18.676 — Stochastic Calculus

Class by Sky Cao

Notes by Sanjana Das

Spring 2025

Lecture notes from the MIT class **18.676** (Stochastic Calculus), taught by Sky Cao. All errors are my own.

Contents

§1 February 3, 2025

§1.1 Syllabus

We'll start by briefly covering the syllabus before we get into the material. What is this course about? Probably we're taking this course because maybe we've heard of the Ito integral and want to learn about that. That'll be the main point of this semester's course; we'll eventually get to stochastic integration and Ito's formula. Eventually we'll get some sort of intuitive understanding of Ito's formula and how you manipulate stochastic integrals — at an intuitive level. But also, hopefully we get some rigorous foundations. There's severla choices you can make with a course like this — you can cover things in less generality but get to the integral much quicker. Sky will take the approach of the textbook, which does things in quite great generality. So it'll take us a while to get to stochastic integration — maybe halfway through the semester. The point is the best time to learn technical details is when you're a student. So we're going to do things in kind of more generality than you might want. Sky thinks if we eventually have to use these things, it's easier to relearn things you've learned before; so it's maybe advantageous to cover the stochastic integral in the generality given by Le Gall. So that's why it'll take us a while to get to integration.

On the back of the syllabus is an outline of topics. It's basically just the chapters of the textbook. We'll follow it closely. But some things Sky will skip because he thinks they're not worth covering in a first course.

After we get to stochastic integration, we'll see how much time is left in the course; the last couple of topics (Markov processes, Brownian motion and PDEs), Sky hopes to cover at least somewhat, but he's not sure how much. But everything up to stochastic integration, we should have time for, so it's a matter of how much more time we have after that.

For prerequisites, we're expected to know 18.675 material (measure-theoretic probability). He highly recommends we've mastered the material in that course before taking this one, since it's foundational to this course — he won't stop us if we haven't, but you'll have to put in a lot of effort (as you wouldn't start learning how to drive in the streets of Manhattan). You also want some knowledge of analysis (L^p spaces, Hilbert spaces, and that type of thing), but that can be filled in by going on Wikipedia (to get certain basic facts).

For logistics, Sky is still deciding on his office hours. The TA and grader, he assumes he'll have at least one of those, but they're assigned in the first or second week of class, so we'll see who it is. There's just homeworks and a final exam, 50% each. The final is usually announced by the third week of class, so we don't know yet when and where; but we should make sure we can make it. For homeworks, we'll have Gradescope. We should already be added on it, though he'll sync the Canvas roster with Gradescope. All homework submissions are on there. Ideally you LaTeX your solutions; if you handwrite, you should make sure it's very neat (ideally on an iPad).

So that's a brief runthrough of the syllabus.

For the first part of the course (which will take 3–4 lectures), we'll talk about Brownian motion. This is one topic that's worth reviewing — the actual construction — because it's the most fundamental example of a continuous-time martingale, and the most basic example with which we'll define stochastic integrals.

So that's what the first couple of lectures will be about.

Sky has posted notes on Canvas, and will be following them. In general he'll try to post handwritten notes for each lecture or module. So we don't have to take our own notes if we don't want to. There's also the textbook, which he's following quite closely; it's also good to read that.

So let's get started.

§1.2 White noise and Brownian motion

The way the textbook constructs Brownian motion is maybe not as standard as other probability textbooks (like Durett). The way Le Gall does it is via this thing called white noise. Sky thinks that's quite interesting,

and maybe later on we'll see other uses for white noise; so he thought it'd be nice to see this alternative construction.

Before we define white noise, let's let H be a real separable Hilbert space. Basically, at least for the purposes of Brownian motion, you can think of H as $L^2(\mathbb{R}_+, dx)$ (where $\mathbb{R}_+$ is the set of nonnegative real numbers, and dx is the Lebesgue measure) — the space of L^2 functions on nonnegative reals is a Hilbert space, where the inner product is $\langle f, g\rangle = \int fg$. So you can think of this as your example of H, though the construction works more generally.

Definition 1.1. A white noise (WN) is a linear isometry $W: H \to L^2(\Omega, \mathcal{F}, \mathbb{P})$ (i.e., a function from H to some fixed probability space) such that for each $h \in H$, we have $W(h) \sim \mathcal{N}(0, \langle h, h\rangle)$.

So we have some fixed probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and for each h, $W(h)$ is some random variable in that space. And it's supposed to be a Gaussian whose variance is the square-norm of h.

Why do we call this white noise? As nonrigorous heuristic discussion, in our example of H, you can think of it as giving you an independent Gaussian for each point in your interval — formally, $W = (W_x)_{x\in[0,\infty)}$ is a collection of IID normals $\mathcal{N}(0,1)$. And here, you formally think of $W(h)$ as this integral $\int dx\, h(x)W(x)$. This is not rigorous because for this integral to be well-defined, W has to be a measurable function. But there's no way W can be measurable — the values at neighboring points in space are completely unrelated to each other. Measurable functions have to be somewhat nice, but this is completely not nice at all. So it's not measurable. But you should really think of white noise as this. Why? It's a linear combination of i.i.d. Gaussians (it's an integral). And a linear combination of i.i.d. centered Gaussians is also a centered Gaussian. So formally, this integral is a centered Gaussian whose variance is the second moment of this thing, i.e., $\mathbb{E}[(\int dx\, h(x)W(x))^2]$. And how do you compute second moments like this? As a heuristic, you can use bilinearity to say this is

$$\int dx\, dy\, h(x)h(y)\mathbb{E}[W_xW_y]$$

(we kind of write the square of an integral as two integrals over dx and dy, and use linearity of expectation). The fact that these are IID enforces that $x = y$, at least at a formal level. And then this becomes $\int dx\, h(x)^2$. And in the case where our Hilbert space is this thing, this is precisely $\langle h, h\rangle$.

So all this is some heuristic discussion for why to think of this as a white noise — heuristically you think of it as coming from this collection of IID standard Gaussians. The way you actually make rigorous sense of this is e.g. by this Hilbert space isometry.

One comment on what the isometry means: In particular, one basic fact we have is that

$$\mathbb{E}[W(h)W(g)] = \langle h, g\rangle$$

for all $h, g \in H$. This for instance you could say is the definition of what an isometry means (between two Hilbert spaces). Another way to derive this is to compute something like $\mathbb{E}[W(h+g)^2]$. On one hand, this is $\langle h+g, h+g\rangle$ by assumption. On the other hand, you can use linearity — we have $W(h+g) = W(h)+W(g)$ — and then expand out to get

$$\mathbb{E}[(W(h) + W(g))^2] = \langle h, h\rangle + 2\mathbb{E}[W(h)W(g)] + \langle g, g\rangle$$

(we expand out and compute each term — for the two terms which are not cross-terms, we use the given definition). But we could also have expanded out

$$\langle h + g, h + g\rangle = \langle h, h\rangle + 2\langle h, g\rangle + \langle g, g\rangle.$$

Then we get this identity. The point is that the two definitions of an isometry (general h and g, or just the case $h = g$) are the same.

§1.3 Construction of white noise

The first thing we need to show about white noise is that it actually exists — we made this definition, but how do you construct a white noise?

> **Claim 1.2 —** White noise exists.

Proof. First, you want to take a basis of your Hilbert space — because H is real and separable, it has a countable orthonormal basis (e_n). What does that mean? It means you can write any element $h \in H$ as a linear combination

$$h = \sum_n \langle h, e_n \rangle e_n$$

(the terms in this sum are projections of h onto each dimension). Then we're going to take an IID sequence $(Z_n)_n$ of standard normal Gaussians $\mathcal{N}(0,1)$, indexed by the same thing as our basis, on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. (There exists *some* probability space where we have a sequence of IID standard Gaussians, e.g. using Kolgomorov extension.)

Then what we do is we define

$$W(h) = \sum_n Z_n \langle h, e_n \rangle$$

as the same linear combination of our standard Gaussians.

This is in principle an infinite series, so why does it converge? We want this to be a random variable in $L^2(\Omega, \mathcal{F}, \mathbb{P})$, so we want to talk about convergence.

One sledgehammer is that by L^2 martingale theory, the right-hand side converges in $L^2(\Omega, \mathcal{F}, \mathbb{P})$. What does that mean? We can define the sequence of partial sums

$$S_N = \sum_{n=1}^{N} Z_n \langle h, e_n \rangle.$$

For each N, this is a finite sum — it's an explicit linear combination of standard Gaussians — and you can verify that S_N is a martingale bounded in L^2. And in 18.675 we saw the L^2 martingale convergence theorem, that a martingale bounded in L^2 converges in L^2 and almost surely. (You prove this using a maximal inequality for L^2 martingales, or something.)

So this infinite series is defined as the L^2 limit of those finite sums — that's what we mean when we say $W(h) = \sum_n Z_n \langle h, e_n \rangle$.

There's another, more elementary, way to see this converges: You can observe that (S_N) is just Cauchy. Why? We can actually explicitly compute the second moments $\mathbb{E}[(S_N - S_M)^2]$ — suppose that $N \geq M$. Then

$$S_N - S_M = \sum_{n=M+1}^{N} Z_n \langle h, e_n \rangle$$

is a linear combination of standard IID Gaussians. And a second moment of a sum is a sum of second moments, so this is

$$\sum_{n=M+1}^{N} \langle h, e_n \rangle^2.$$

And the point is that this goes to 0 — more specifically,

$$\lim_{k \to \infty} \sup_{N,M \geq k} \sum_{n=M+1}^{N} \langle h, e_n \rangle^2 = 0,$$

just because h was in your Hilbert space to begin with. (One fact is that when you have an orthonormal basis, this type of sum has to be a Cauchy sequence — that's exactly what it means that $\sum_n \langle h, e_n\rangle e_n$ converges, that it converges in your Hilbert space; and the reason it converges is that you can verify the partial sums give you an infinite series in your Hilbert space). So all this is to say that instead of using martingale theory, you can directly verify this sequence is Cauchy in $L^2(\Omega, \mathcal{F}, \mathbb{P})$; and then by completeness, you just get that it converges.

So that's what this means.

So that's our candidate for white noise; now we have to check its properties. So we need to say why it's linear, and why it satisfies our distributional assumption.

Linearity is just because the coefficients $\langle h, e_n\rangle$ are linear in h; so we won't check that.

Now we'll check the distributional assumption. This is because for all N, we have $S_N \sim \mathcal{N}(0, \sum_{n=1}^N \langle h, e_n\rangle^2)$ (since it's a sum of Gaussians). And these variances converge to $\langle h, h\rangle$, so $S_N \to \mathcal{N}(0, \sum_n \langle h, e_n\rangle^2)$ (in distribution). (For a sequence of centered Gaussians, convergence in distribution is equivalent to convergence of the variance; you can see this by characteristic functions.) And this variance is precisely $\langle h, h\rangle$.

So S_N converges in distribution to $\langle h, h\rangle$. It also converges in L^2 to $W(h)$, by definition. And then using the fact that the limits have to be the same, this means $W(h) \sim \mathcal{N}(0, \langle h, h\rangle)$ (you can only converge in distribution to one thing, and convergence in L^2 implies convergence in probability and therefore distribution). So $W(h)$ has the distribution we want, which completes the proof. □

§1.4 Brownian motion

Now that we know white noise exists, we can get onto constructing Brownian motion.

Let's make some definitions first.

> **Definition 1.3** (Stochastic process)**.** Fix a measurable space $(E, \mathcal{E})$, and let T be some index set. A stochastic process (indexed by T, and with values in E) is a collection of random variables $(X_t)_{t\in T}$ (all on the same probability space) with values in E.

By default, usually we take $(E, \mathcal{E})$ to be $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ (where $\mathcal{B}(\mathbb{R})$ is the Borel σ-algebra); so if we say 'stochastic process' without specifying E, it's just a real-valued process.

So a stochastic process is just a collection of random variables.

> **Definition 1.4** (Gaussian process)**.** A stochastic process $(X_t)_{t\in T}$ is a centered Gaussian process if any finite linear combination of the X_t's is a centered Gaussian.

As a check of understanding, if T is finite, then this precisely says the law of your process is a centered multivariate Gaussian. Here we're letting T be arbitrary (it could be uncountable). But this still says that any *finite collection* $(X_{t_1}, \ldots, X_{t_n})$ is a centered multivariate Gaussian.

(Some of this is from Chapter 1 of Le Gall.)

§1.4.1 Pre-Brownian motion

Now with all these definitions, we can construct a Gaussian process which basically should be Brownian motion, except that we have a problem of continuity.

Definition 1.5. Let $W: L^2(\mathbb{R}_+, dx) \to L^2(\Omega, \mathcal{F}, \mathbb{P})$ be a white noise. Then the process

$$B_t = W(\mathbf{1}_{[0,t]})$$

is called a pre-Brownian motion.

So we're taking a white noise on our example Hilbert space from earlier; and applying it to a *particular* function.

Remark 1.6. More generally, we'll say any process with the same finite-dimensional distributions as this is called a pre-Brownian motion. Le Gall actually shows that for any such thing, you can construct a white noise on some space such that your white noise arises in this way (every pre-Brownian motion comes from some white noise), but Sky doesn't think we ever need that — it's the law of pre-Brownian motion that's relevant for us, not necessarily how it arises.

Now let's talk about some properties of pre-Brownian motion. First, why the name? As Le Gall emphasizes, this is not very standard terminology; but the reason Le Gall uses this name is it doesn't have continuous sample paths yet, so we're going to have to do something. (The actual construction of Brownian motion is probably the first major result in this course. But before we get to this, let's talk about some properties.)

Proposition 1.7

Pre-Brownian motion is a Gaussian process with covariances $\mathbb{E}[B_t B_s] = t \wedge s$.

(The notation $t \wedge s$ means $\min\{t, s\}$.)

Proof. The fact that it's a Gaussian process follows because W is linear and a white noise — if we take any finite linear combination of pre-Brownian motion, we can write it as $W(h)$ for some Hilbert space element h, and by definition this is Gaussian.

(Even if the pre-Brownian motion doesn't arise from a white noise, it's always equal in distribution to something arising from white noise; so to prove a statement about the law, we can always assume this is true, i.e., that our pre-BM arises from a WN in this way. Then for any finite linear combination, you can write it as $W(h)$ for some Hilbert space element h.)

To finish, we just need to compute the covariance. Going back to the definition of white noise, we have

$$\mathbb{E}[B_t B_s] = \mathbb{E}[W(\mathbf{1}_{[0,t]}) W(\mathbf{1}_{[0,s]})].$$

And recall we saw that $\mathbb{E}[W(h)W(g)] = \langle h, g \rangle$. In this case, the L^2 inner product is $\int fg$, so that means this is

$$\int_0^\infty dx\, \mathbf{1}_{[0,t]}(x) \mathbf{1}_{[0,s]}(x).$$

But this is just integrating 1 from 0 to $t \wedge s$; so this is just $t \wedge s$. □

Now let's cover another proposition about pre-BM — the fact that it has independent increments.

Proposition 1.8

If (B_t) is a pre-Brownian motion, then $B_0 = 0$ (almost surely); $B_t - B_s \sim \mathcal{N}(0, t-s)$ for all $s \le t$; and for any finite sequence $0 = t_0 < t_1 < \cdots < t_n$, the collection

$$B_{t_1} - B_{t_0},\ B_{t_2} - B_{t_1},\ \ldots,\ B_{t_n} - B_{t_{n-1}}$$

are mutually independent. Moreover, this characterizes the law of pre-Brownian motion.

Naturally when your index set is the real line (or a subset), we think of t as time. (So sometimes we'll refer to these as times.)

Proof. First, for the characterization part, this basically follows from combining these three things — then any finite-dimensional distribution $(B_{t_0}, B_{t_1}, \ldots, B_{t_n})$ is determined; and then there's some abstract measure theory results that the finite-dimensional distributions characterize the full law (even if your index set is uncountable). So that's a couple of sentences on why the last thing is true; let's just focus on the first three.

For the first part, by assumption $B_0 \sim \mathcal{N}(0,0)$.

For the second, regarding $B_t - B_s$, there's a couple of things you can do. One thing is we can again assume B comes from some white noise, so then $B_t - B_s = W(\mathbf{1}_{(s,t]})$ (using the linearity of W and the fact that B_t and B_s are both W's of some indicator functions). And by definition, this is precisely $\mathcal{N}(0, t-s)$ (because the norm-squared of $\mathbf{1}_{(s,t]}$ is $t-s$).

Finally for independence, one fact about Gaussian processes is that 0 covariance implies independence. (This isn't true for general random variables, but for Gaussian processes it is.) So we just need to check that any two different terms of our collection have zero covariance. This basically reduces to a fact about orthogonality in your Hilbert space — that

$$\langle \mathbf{1}_{(t_{i-1},t_i]}, \mathbf{1}_{(t_{j-1},t_j]} \rangle = 0.$$

(That's because these are disjoint intervals, and the inner product in $L^2(\mathbb{R}_+)$ corresponds to integrating over the intersection of the intervals.)

And that means any of these two increments has covariance 0 with each other; and that implies independence for multivariate Gaussians. □

Now let's see some more properties, regarding symmetries of pre-BM.

Proposition 1.9 (Symmetries of pre-BM)

Let B be a pre-Brownian motion.

(i) $-B$ is a pre-Brownian motion.

(ii) (Scaling symmetry of time and space) For all $\lambda > 0$, we have that $\lambda^{-1} B_{\lambda^2 t}$ is also a pre-Brownian motion.

(iii) (Markov property) For all $s \geq 0$, the 'restarted process' $B_t^{(s)} = B_{t+s} - B_s$ is also a pre-Brownian motion. Moreover, it is independent of $\sigma(B_r \mid r \leq s)$ (i.e., it's independent of everything that happened before).

These will also be properties of Brownian motion, once we construct it.

As a couple of words, to verify you're a pre-BM, you just have to verify you have the same law; and the law is completely characterized by the previous three properties. So you just have to check those three properties are satisfied. The $-B$ symmetry is most easily seen because $\mathcal{N}(0,1)$ is symmetric; so that's why (i) is true.

For (ii), you check each of these properties. The main one is why you have the right variance — i.e., $B_t - B_s \sim \mathcal{N}(0, t-s)$. This scaling symmetry is quite important, and we'll see an application later on. Intuitively, what it's saying is, let's take λ to be very tiny, and suppose you want to compute some probabiltiy of Brownian motion on the unit timescale $[0,1]$. So you're observing Brownian motion; how large does it get on a unit timescale? You can kind of relate this probability to the probability it goes above a much smaller number λ on a much smaller timescale. So instead of a unit timescale, you have λ^2 times a unit timescale; and instead of seeing if it goes above 1, you see whether it goes above λ. So you can relate properties of Brownian motion on different spatial and time-scales. We'll see multiple examples. For instance, suppose

you ask whether Brownian motion goes above 1, ever — what's the probability your Brownian motion gets larger than 1 at any point in time? Using this scaling symmetry, that's the same as asking whether it gets above any small number λ on your infinite timescale. (You want to know whether $\lambda^{-1}B_{\lambda^2 t}$ goes above 1. And if your timescale is infinite, λ^2 times your infinite timescale is still infinite.) And from that, you'll see that with probability 1, it gets above 1. (We'll cover this after we've actually constructed Brownian motion.)

Finally for the Markov property, the fact that $B_{t+s} - B_s$ is pre-Brownian motion is an explicit computation. The fact that it's independent of everything before time s basically reduces to the third fact, about the independence of increments — to be independent of $\sigma(B_r \mid r \leq s)$ it suffices to be independent of any finite collection of the B_r's, and for a finite collection you're basically saying exactly this type of thing.

§1.5 Some foreshadowing

Now let's do some more foreshadowing. Heuristically, if you have a white noise W on $L^2(\mathbb{R}_+)$, you can write

$$W(f) = \int_0^\infty f_t dB_t$$

(that's how we're eventually going to think of this thing). This is just some notation. But formally, why would you expect this to be true? It's basically inspired by the fact that if $f = \mathbf{1}_{(s,t]}$, then as we saw, we have $W(f) = B_t - B_s$. And that's in line with your intuition of how integrals are supposed to behave — that this should be $\int_0^\infty \mathbf{1}_{(s,t]}(u)\, dB_u$ (because at least on indicator functions of intervals, you should get the difference of endpoints — that's a property that your integrals should satisfy). The whole problem is actually constructing an integral with respect to dB; that's what stochastic integration is about.

Why can't you just do this directly? Eventually for us B is going to be Brownian motion. But the problem is that B is not differentiable anywhere. In fact, it's quite far from differentiable — it's only $(1/2)$-Hölder, in that if you take a small increment, the typical size of $B_t - B_s$ is something like $(t-s)^{1/2}$. This is because $B_t - B_s \sim \mathcal{N}(0, t-s)$, which means its standard deviation is $\sqrt{t-s}$, and the standard deviation is typically how large you are. But if you're differentiable, this had better be $(t-s)^1$. So the fact we have $1/2$ instead of 1 means this can be quite wild.

So you can't just use standard calculus to define this integral. Ultimately, the reason why you can is that Brownian motion has nice independence. And you end up defining it by some abstract Hilbert space isometry arguments, in the spirit of white noise. In fact, we'll define it with more general integrands than $\mathbf{1}_{(s,t]}$; but when you take integrands like this, we will recover the above identity.

So we'll eventually be able to make sense of integrals of the form $\int f_t\, dB_t$; and if you take f of this form, you'll get this thing.

That's what we're working towards in the first six weeks. But before we get to that, we'll need some foundational material.

§1.6 Continuity of sample paths

We'll now take a stochastic process taking values in a metric space (rather than just a general measurable space), since to talk about continuity, we need some topology.

§1.6.1 Definitions

First, what is a sample path?

Definition 1.10. Let (E, d) be a metric space, and let $(X_t)_{t\in T}$ be an E-valued stochastic process. The sample paths of X are the mappings $T \to E$ defined by $t \mapsto X_t(\omega)$ (for each fixed $\omega \in \Omega$).

Here ω is lying in your abstract probability space Ω; so for each element of your probability space, we get some function.

Remark 1.11. When we have a metric or topological space, the σ-algebra we put on it is always the Borel one (the smallest σ-algebra making all your open balls measurable).

Next we'll define what it means to be a *modification*.

Definition 1.12. Let X and $\widetilde{X}$ be E-valued stochastic processes on the same probability space. We say that $\widetilde{X}$ is a modification of X if

$$\mathbb{P}[X_t = \widetilde{X}_t] = 1 \quad \text{for all } t \in T.$$

So basically, to be a modification, for any fixed t, you're almost surely equal.

Definition 1.13. We say $\widetilde{X}$ is indistinguishable from X if there exists a *negligible set* $\mathcal{N} \subseteq \Omega$ (meaning that there exists $F \in \mathcal{F}$ such that $\mathcal{N} \subseteq F$ and $\mathbb{P}[F] = 0$), and for all $\omega \notin \mathcal{N}$, we have $X_t(\omega) = \widetilde{X}_t(\omega)$ for all t.

Indistinguishability is a stronger condition than just being a modification.

What is a negligible set? It's not necessarily measurable — we have a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and we don't necessarily have $\mathcal{N} \in \mathcal{F}$. But it shoudl be contained in some event of probability 0.

And you're indistinguishable if you're the same outside of this negligible set.

Student Question. *Why define negligible sets instead of using F?*

Answer. You probably don't need it; you should be able to just use F.

First, why is this stronger? Well, if for a fixed t you have $X_t \neq \widetilde{X}_t$, then that ω is contained in your negligible set $\mathcal{N}$. The point is to go from 'modification' to 'indistinguishable,' you sort of have to take an intersection over all events corresponding to T. If T is countable, that's fine, and being a modification is equivalent to being indistinguishable. The problem arises if T is uncountable (σ-algebras are defined by countable operations, so uncountable intersections aren't necessarily in your σ-algebra).

Remark 1.14. For our practical purposes, they'll be basically the same; but in principle, the implication does not go downwards, so you always have to be careful. But for us, e.g. if you additionally assume that your sample paths are continuous or right-continuous, then being a modification actually implies you're indistinguishable. So we're going to use that; but otherwise there are counterexamples in full generality.

Student Question. *Can we not just define indistinguishability by taking the definition of a modification and putting 'for all t' inside the bracket?*

Answer. Kind of, but not exactly — heuristically, you think of this definition as

$$\mathbb{P}[X_t = \widetilde{X}_t \text{ for all } t \in T] = 1.$$

But the problem is that when you take probabilities, you assume your event is measurable. And if T is uncountable, then this event $\{X_t = \widetilde{X}_t \text{ for all } t\}$ is not necessarily in $\mathcal{F}$. (But there are some conditions

— e.g., having continuous sample paths — under which it is.)

Remark 1.15. If your σ-algebra is complete, then a subset of a measure-0 set is also measurable (with measure 0). This is not true in general (e.g., for $\mathcal{B}(\mathbb{R})$); but for this reason, we're usually going to work with complete σ-algebras.

(You can in general complete a σ-algebra by taking $\sigma(\mathcal{F}, \mathcal{N})$ where $\mathcal{N}$ is the set of negligible sets. And one can check that this is actually complete. So eventually we'll be using these completed σ-algebras, and then these problems will go away.)

§1.6.2 Overview of Kolgomorov's continuity criterion

Finally, we'll say a bit about the next result, Kolgomorov's continuity criterion, and probably get to it next time. It was covered in 18.675, but it's a foundational result, and the proof technique is good to understand and remember; it's quite important, and shows up in other contexts as well when dealing with stochastic processes. This is going to allow us to construct a continuous modification of a pre-BM.

The Kolgomorov's continuity theorem — we're going to use it to construct continuous modifications, in particular of pre-BM. That'll allow us to construct Brownian motion as a stochastic process with continuous sample paths. Then after that, we'll talk about some sample path properties of Brownian motion.

Brownian motion is continuous, but has some quite pathological properties. At one point people thought every continuous function has to be (piecewise) differentiable; but that's not true, and Brownian motion is a counterexample. Also, BM starts at 0, but we'll see that on any tiny interval, it goes above and below 0 infinitely many times. So it's a very wild continuous function. This kind of relates to the fact that it's 'randomly chosen' — you typically expect 'random continuous functions' to not have the greatest regularity properties.

The first homework will be due next Thursday so it'll be posted soon, based on today's and Wednesday's material (and maybe some foundational 18.675 material as review).

§2 February 5, 2025

§2.1 Kolgomorov's continuity criterion

Last time, the main point was to construct pre-Brownian motion, a stochastic process with the right finite-dimensional distributions. Now we want to construct Brownian motion. We'll do this by proving this continuity theorem, which says that under certain conditions you can get a continuous modification.

Theorem 2.1 (Kolgomorov continuity theorem)

Let $(X_t)_{t\in I}$ be a stochastic process indexed by a bounded interval of $\mathbb{R}$, taking values in a complete metric space (E, d). Assume there exist $q, \varepsilon, C > 0$ such that for all $s, t \in I$, we have

$$\mathbb{E}[d(X_s, X_t)^q] \leq C\,|t-s|^{1+\varepsilon}\,.$$

Then there exists a modification $\widetilde{X}$ of X which is α-Hölder for all $\alpha \in (0, \varepsilon/q)$, i.e.,

$$\sup_{s,t\in I} \frac{d(X_s, X_t)}{|s-t|^\alpha} < \infty.$$

What does this theorem say? You assume you have a stochastic process indexed by some interval, taking values in some complete metric space (for pre-Brownian motion it's just $\mathbb{R}$). The assumption is that in expectation, your stochastic process doesn't vary too much — more precisely, you have some parameters such that the qth moment of $d(X_s, X_t)$ is bounded by some power of $|t-s|$; it turns out that the precise power should be $1+\varepsilon$. (The fact that you need a power greater than 1 is special to the fact that you're assuming your index set I is a subset of $\mathbb{R}$, which is 1-dimensional. If you wanted to do something for higher-dimensional index sets, you'd probably need d instead of 1.) But intuitively, you're just saying in expectation your process shouldn't vary too much. We really care about what happens when s and t are very close (since we want a continuous modification).

And under this assumption you get something more than just continuous — you get an α-Hölder modification, which is stronger. (You have some quantitative estimate on $d(X_s, X_t)$.)

So that's the statement of this theorem. It's quite an important theorem, so it's good to remember at least generally how this proof goes. You may not remember every single detail, but the technique is quite important. Basically any time you want to bound the supremum of a stochastic process, you do something similar to this.

The technique when you bound suprema is you want to do this kind of dyadic decomposition, or *multiscale analysis*. For simplicitly, let's assume $I = [0, 1]$. (It's not going to change the proof much, except for introducing extra notation, so we'll just assume this.)

What's multiscale analysis? For each $n \geq 1$, let D_n be the set of dyadic numbers given by 2^{-n}, i.e.,

$$D_n = \{j2^{-n} \mid 0 \leq j \leq 2^n\}.$$

Basically we're taking a sequence of finer and finer meshes; multiscale refers to the fact that as n increases, you're going down to finer and finer scales, where each time you increase n, you go down to a smaller scale by a constant factor.

So that's D_n. And for notation for later, let $D = \bigcup_n D_n$. Also, fix $0 < \alpha < \varepsilon/q$. First we're going to prove there exists an α-Hölder modification for this fixed α; then we'll show later there exists one for all α.

What does our assumption give us? Let's just start small; then the assumption gives us that for $s, t \in D_n$ which are one dyadic unit apart, meaning that $|s-t| = 2^{-n}$, you're going to basically have control on how

much X_s and X_t vary, or how far they are from each other. This moment assumption basically tells you, just by Markov's inequality, that

$$\mathbb{P}[d(X_s, X_t) > \lambda] \leq \lambda^{-q} \cdot 2^{-n(1+\varepsilon)}.$$

Here 2^{-n} is $|t-s|$ by assumption, and all we did was take both sides to the qth power and apply Markov.

So that's the basic estimate that the assumption gives us. Now we'll do something you might think is quite crude, which is just to union-bound — this implies

$$\mathbb{P}[\max_{\substack{|s-t|=2^{-n} \\ s,t\in D_n}} d(X_s, X_t) > \lambda] \leq \lambda^{-q} 2^{-n\varepsilon}$$

(there's basically just 2^n such s and t). You can see here that the reason we needed a power greater than 1 in the estimate $|t-s|^{1+\varepsilon}$ was that we were using this union bound, and we just want to have some leftover decay.

So we have this estimate. And it's true for any general $\lambda > 0$. Now what we're going to do is, this implies that for any sequence (λ_n) (which we're going to pick soon), we have

$$\sum_{n\geq 1} \mathbb{P}\left[\max_{|s-t|=2^{-n}, s,t\in D_n} d(X_s, X_t) > \lambda_n\right] \leq \sum_n \lambda_n^q 2^{-n\varepsilon}.$$

Basically for each fixed n, I apply the above estimate with my choice of λ_n (which we haven't chosen yet). The point of all this is we're going to see where the choice of λ_n comes from. We basically want this to be summable, so we need to choose λ_n so that this is summable; and your choice of λ_n is basically going to inform why we restrict $\alpha \in (0, \varepsilon/q)$.

We want the right-hand side to be finite. And the reason we have our restriction on α is that if we take

$$\lambda_n = (2^{-n})^\alpha,$$

then this is actually summable — we get $\sum_n 2^{n(q\alpha-\varepsilon)}$, and by assumption this power of 2 is negative, so this is less than ∞.

What have we just done? Well, we showed that the sum of all these probabilities is finite. So by Borel–Cantelli I, this tells you that these events almost surely don't happen infinitely often — i.e., almost surely, for all n large, you're going to have that

$$\max_{|s-t|=2^{-n}, s,t\in D_n} d(X_s, X_t) \leq \lambda_n = 2^{-n\alpha}.$$

So we see this assumption on the moments gives us this control on the small-scale behavior of the process, at least on these dyadic points, just by Borel–Cantelli.

Now we're almost there. At least for points which satisfy the condition $|s-t| = 2^{-n}$ and $s, t \in D_n$, you have the Hölder estimate. Now we want to go from this to saying you have some Hölder modification. That's going to mostly be an analytic lemma. Basically, what this lemma says is that if you have a function defined on $D = \bigcup_n D_n$ such that you have the above estimate, then f is actually α-Hölder on D.

Lemma 2.2 (Analytic lemma)

Let $f: D \to (E, d)$ be such that for all large n, we have

$$\max_{\substack{|s-t|=2^{-n} \\ s,t\in D_n}} d(f(s), f(t)) \leq 2^{-n\alpha}.$$

Then f is α-Hölder on D.

This doesn't involve probability — it's just a fact about functions that if you satisfy this condition, then you're actually α-Hölder on all of D. Here s and t don't necessarily satisfy this condition — they're just general dyadic numbers — and you want to use this condition to still get good control.

We'll prove this at the end, but we'll just apply it for now. By the analytic lemma, you get that X is α-Hölder on D. And then, basically the rest is analytic facts (i.e., real analysis). Since D is dense in $[0,1]$ (recall that we fixed $I = [0,1]$ for simplicity) and (E,d) is complete, you get that X has a unique α-Hölder extension to the entire interval $[0,1]$. (It's kind of natural how you would extend X — for a given general value of t, you'd take a sequence in your dense subset D converging to t, and you just need to show the corresponding X-values converge; that's guaranteed because X is α-Hölder. And then you verify the extension itself is α-Hölder, but that's just some density argument.)

So we'll call this extension $\widetilde{X}$. This is our candidate modification of X. It's α-Hölder, because we just said it is. So what we need to show, by the definition of a modification, is that

$$\mathbb{P}[X_t = \widetilde{X}_t] = 1 \quad \text{for al } t.$$

First, by construction it's true for all $t \in D$, because when you restrict $\widetilde{X}$ to D, it's X itself (that's implicit in the word 'extension'). And then the moment assumption in the statement of the theorem implies that in fact, $X_{t_n} \xrightarrow{p} X_t$ (denoting convergence in probability) for any sequence $t_n \to t$ — this is because you get that the qth moment of distances converges to 0, and that's enough to imply convergence in probability.

Then you're basically done — $\widetilde{X}$ is α-Hölder and therefore continuous, so

$$\widetilde{X}_t = \lim_{n\to\infty} \widetilde{X}_{t_n}.$$

And if you took $t_n \in D$ (which you can always do, because D is dense), then by construction this is equal to $\lim_n X_{t_n} = X_t$. (The first is an almost sure limit, and the last is a limit in probability.) And almost sure convergence implies convergence in probability, and limits for convergence in probability are unique; so you get $\widetilde{X}_t = X_t$ almost surely (for any fixed t).

So we've verified that $\widetilde{X}$ is an α-Hölder modification of X.

Student Question. *How does the moment assumption imply convergence in probability?*

Answer. The assumption says $\mathbb{E}[d(X_{t_n}, X_t)^q] \to 0$. And then you can finish by Markov — you have

$$\mathbb{P}[d(X_{t_n}, X_t) > \lambda] \le \lambda^{-q}\mathbb{E}[d(X_{t_n}, X_t)^q] \to 0.$$

(Here you only need q to be any positive power, and you're fine.)

We're almost done proving the theorem (assuming the analytic lemma). But now we just need to comment on why we can find a modification that works for *all* α simultaneously. This is just a stitching-together argument — take some $\alpha_n \nearrow \varepsilon/q$ (i.e., an increasing sequence always strictly less than ε/q, but converging to ε/q). Then for each α_n, we have an α_n-Hölder modification of X, which we'll denote by $\widetilde{X}^n$.

Claim 2.3 — We have $\mathbb{P}[\widetilde{X}^n_t = \widetilde{X}^m_t \text{ for all } t] = 1$ for all n and m.

In other words, we're saying that these processes are all indistinguishable. Note that this event is actually measurable — the $\widetilde{X}$'s are continuous, so this equality can be written on a countable dense subset of t's, and then it's a countable intersection.

Proof. We can write the left-hand side as

$$\mathbb{P}[\widetilde{X}^n_t = \widetilde{X}^m_t \text{ for all } t \in D]$$

(because $\widetilde{X}^n$ and $\widetilde{X}^m$ are continuous). But now you can use countable additivity to say this is 1 (the countable intersection of probability-1 events is probability-1). □

So now we can just define our modification $\widetilde{X}$ as

$$\widetilde{X} = \widetilde{X}^1 \mathbf{1}_E,$$

where E is the event

$$E = \{\widetilde{X}^n = \widetilde{X}^m \text{ for all } t \text{ and } n, m\}.$$

This event E is probability-1 — before we fixed some m and n, but you can take a countable intersection and get a probability-1 intersection even if you quantify over all. And if $\widetilde{X}^1$ is a modification of X, then this thing is also a modification of X (since we just modified $\widetilde{X}^1$ up to indistinguishability).

And why is this α-Hölder? Because on this event, $\widetilde{X} = \widetilde{X}^n$ for any n, which is α_n-Hölder; and if I take n large enough then α_n gets arbitrarily close to ε/q.

Student Question. *How do you get uniqueness of the extension (in the unique α-Hölder extension)?*

Answer. Any continuous extension is going to have to be unique — if you're equal on a dense subspace, then it has to be equal everywhere.

To recap, the main thing was kind of the beginning, where we use the moment assumption to conclude that if you choose the λ_n in the right way, you can apply Borel–Cantelli to get the statement about $\max d(X_s, X_t) \leq 2^{-n\alpha}$. The rest is just some real analysis and stitching things together; it's maybe necessary but not the main point. The main point is how you use the moment assumption, and that's by going down to these dyadic scales.

§2.1.1 Proof of the analytic lemma

Now let's go to the analytic lemma. You have to be slightly careful when doing this. First, one preliminary reduction is that the assumption implies there's some C such that for *all* n, you have

$$\max_{|s-t|=2^{-n}, s,t \in D_n} d(f(s), f(t)) \leq C 2^{-n\alpha}$$

(because it's true for all large n, and then you can make C large enough to handle the finitely many small remaining values of n). That will be somewhat convenient — that this thing holds for all n in the argument.

Now let's fix some $s < t$, with $s, t \in D$. Our goal is to show that basically

$$d(f(s), f(t)) \lesssim |s-t|^\alpha$$

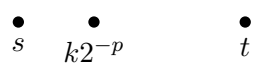
(where $\lesssim$ means 'up to constants,' where the constant doesn't depend on s and t). This is basically just restating the Hölder condition. (We don't care about the constant.)

We'll make a drawing.

We have our points s and t. The first thing you want to do is go on the dyadic scale on par with the distance between s and t — in other words, we let p be the smallest integer such that $2^{-p} \leq t - s$. So we go down to this dyadic scale.

Once we've gone down to this dyadic scale, let's let k be the smallest integer such that $k2^{-p} \geq s$.

So what have we just done? We take the spacing of our mesh fine enough that it separates s and t, and then we find the first point in this mesh which is in our interval.



First we can make one observation. We drew this point to be between s and t; it turns out this has to be the case, i.e., that $k2^{-p} \leq t$. Why? By the assumption that k is the first such integer, we have

$$(k-1)2^{-p} < s,$$

which implies $k < s + 2^{-p}$ (just moving 2^{-p} onto the right-hand side). And we assumed that $2^{-p} \leq t - s$, so this is less than t. So we *can* draw the above picture.

Now we've gone down to this scale that separates the point. And what you want to do is write s and t as a combination of dyadic points, starting from $k \cdot 2^{-p}$ (where our points get exponentially closer to s or to t).

In other words, we can write

$$s = k2^{-p} - a_1 2^{-p-1} - \cdots - a_n 2^{-p-n}$$

and similarly

$$t = k2^{-p} + b_1 2^{-p-1} + \cdots + b_m 2^{-p-m},$$

where $a_i, b_i \in \{0, 1\}$. Why do we start at $-p-1$? Well, I can't start at 2^{-p}, because if I subtracted one whole power then I'd go strictly less than s. So these points I drew which are getting increasingly closer to t and s are the partial sums of these expressions.

Let's make some notation for these partial sums — we'll let s_r be the first $r+1$ terms of this sum, so

$$s_r = k2^{-p} - a_1 2^{-p-1} - \cdots - a_r 2^{-p-r}.$$

Similarly, let's let

$$t_\ell = k2^{-p} + b_1 2^{-p-1} + \cdots + b_\ell 2^{-p-\ell}.$$

What's the point of all this? Here's where we're going to use our assumption in the analytic lemma. We eventually want to cmpare the values of f at s and t. So you find this common point $k2^{-p}$, and compare the values of f between every two consecutive points. You have to be kind of clever in choosing this, but this argument turns out to work.

Also, to initialize, we have $s_0 = t_0 = k2^{-p}$.

Now let's estimate $d(f(s), f(t))$. We have $s = s_n$ and $t = t_m$, so we can write this as

$$d(f(s), f(t)) = d(f(s_n), f(t_m)) \leq \sum_{i=1}^{n} d(f(s_i), f(s_{i-1})) + \sum_{j=1}^{m} d(f(b_j), f(b_{j-1}))$$

(where the first sum compares the distances between successive s_r-values, and the second does likewise for t_ℓ's). The fact that $s_0 = t_0$ means that you don't need an extra $d(f(s_0), f(t_0))$ term here — this common point is just equal, so this distance is just 0, and we don't need to write it.

Now that we've written it in this form, there's nothing left to do besides to apply our assumption, so let's do that. By the assumption, we can bound our distance by the sum of two series. The distance scale between successive points is $|s_i - s_{i-1}| = 2^{-p-i}$ (or maybe we're off by a factor of 2, but it doesn't matter). And so by using the assumption, we get that this whole first sum is estimated by

$$\sum_{i=1}^{\infty} C \cdot 2^{-(p+i)\alpha}$$

(here n in our given condition is actually $p+i$, from the way we set things up; there might be some constant here with ± 1 in the exponent, but that doesn't matter for this part). Similarly, the second sum we can estimate by basically the same thing,

$$\sum_{j=1}^{\infty} C 2^{-(p+j)\alpha}.$$

(The sums are finite, but we might as well make them infinite.)

And this is the whole point of going to dyadic scales — you get these geometric series, which are going to be dominated by their first terms. And the first terms are basically $2^{-p\alpha}$ (it might be $+1$, but we don't care about factors of 2). So we get something like

$$2^{p\alpha} \cdot \frac{1}{1-2^{-\alpha}} \lesssim 2^{-p\alpha}.$$

(The extra thing is a constant we don't care about.)

And the way we set things up, we went down to the distance scale of $t-s$ (i.e., $2^{-p} \le t-s$), so this is bounded by $|t-s|^\alpha$. And that's precisely what we wanted to show. So that concludes the proof.

So this is an analytic lemma, which we include for completeness; but you do have to be slightly clever when doing this.

§2.2 Brownian motion

Now we'll use this to modify pre-Brownian motion. We'll first restrict to just $[0,1]$, because Kolgomorov's continuity theorem requires a bounded interval.

Corollary 2.4

Let B be a pre-Brownian motion on $[0,1]$. Then B has a modification which is α-Hölder for all $\alpha < 1/2$.

So it's just barely below $1/2$ in terms of Hölder regularity.

Why should you expect $1/2$ to be the limit? It probably can't be any larger because you recall that $\mathbb{E}[|B_t - B_s|^2] = t-s$ (it's a normal with variance $t-s$). And we kind of mentioned last time that this means $|B_t - B_s|$ is sort of of order $(t-s)^{1/2}$. So even in expectation you can't hope for anything with higher regularity than $1/2$ (higher regularity is stronger, because you think of $t-s$ as small). So $1/2$ is the natural limit, and you can go just below it. (The fact that you can't get something exactly equal to $1/2$ is because if you want to take a sup over all s and t, you'll have to pay a tiny factor in regularity.)

Proof. We kind of almost said this, but the idea is you want to take higher moments. We know that

$$B_t - B_s = (t-s)^{1/2}\mathcal{N}(0,1)$$

in distribution, which means that for any $2p$th moment, we have

$$\mathbb{E}[|B_t - B_s|^{2p}] = (t-s)^p \mathbb{E}[Z^{2p}]$$

(we're trying to verify the assumtpion in the statement of Kolgomorov), where $Z \sim \mathcal{N}(0,1)$. We don't really care what $\mathbb{E}[Z^{2p}]$ is; it's just some finite constant. So we get

$$\mathbb{E}[|B_t - B_s|^{2p}] \le C_p (t-s)^p.$$

Written a bit more evocatively, we can say $p = 1 + (p-1)$. Now applying Kolgomorov with $q = 2p$ and $\varepsilon = p-1$, you obtain a modification which is α-Hölder for all $\alpha < (p-1)/2p = 1/2 - 1/2p$. And then you send $p \to \infty$. Again, you maybe want to do this stitching argument to say why you can get a modification that works for *all* α simultaneously, but it's basically the same argument from earlier. □

So that's Brownian motion on $[0,1]$. For Brownian motion on $[0,\infty)$, you could apply the corollary on each bounded interval $[n, n+1]$; then for each interval like this, you get a modification of pre-Brownian motion which is continuous and α-Hölder for all $\alpha < 1/2$. And then you can kind of 'stitch' the modifications together again to obtain a continuous modification of pre-Brownian motion on this infinite interval $[0,\infty)$ which is *locally* α-Hölder for all $\alpha < 1/2$. What 'locally' means is that if you take I to be any bounded subinterval of $[0,\infty)$, then you get

$$\sup_{s,t\in I} \frac{d(X_s, X_t)}{|s-t|^\alpha} < \infty.$$

Student Question. *How do you stitch together the modifications?*

Answer. Let's say I have a pre-Brownian motion B, and $\widetilde{B}^1$ is a modification on $[0,1]$, $\widetilde{B}^2$ is a modification on $[1,2]$, and so on (I can get these by applying the corollary). And now I want to define

$$\widetilde{B}^{12} = \widetilde{B}^1 \mathbf{1}_{[0,1]} + \widetilde{B}^2 \mathbf{1}_{[1,2]}.$$

Why do I get a continuous function (almost surely)? For that, I have $\widetilde{B}^1(1) = B(1) = \widetilde{B}^2(1)$ almost surely, because both of these are modifications of our pre-Brownian motion B. So to be precise, we should really define $\widetilde{B}^{12}$ as in the above formula on the event $\widetilde{B}^1(1) = \widetilde{B}^2(1)$; and off this event, we just set it to 0 (we don't really care about a probability-0 event).

So I can stitch together a Brownian motion on $[0,1]$ and $[1,2]$. And I can iterate this. You're only doing a countable number of things, so this will result with a continuous modification of pre-Brownian motion. (There's no issues with uncountable intersections because you just do this a countable number of times.)

That leads to defining Brownian motion.

Definition 2.5. A process (B_t) is a Brownian motion if:

- B is a pre-Brownian motion (i.e., it has the right finite-dimensional distributions).
- B has continuous sample paths.

We showed above that Brownian motion is guaranteed to exist (at least on some probability space).

§2.3 Brownian motion as a function-valued random variable

Now that we know Brownian motion exists, we'll talk about some general considerations relating to Brownian motion.

Notation 2.6. Let $C(\mathbb{R}_+, \mathbb{R})$ be the space of continuous functions $\mathbb{R}_+ \to \mathbb{R}$ (where $\mathbb{R}_+ = [0,\infty)$).

We kind of want to describe the law of Brownian motion as a random variable taking values in this space (it does that because by definition it has continuous sample paths). So you can think of Brownian motion not just as a stochastic process, but as a random variable taking values in this space.

But we want to endow this space with a measurable structure, i.e., a σ-algebra. So what should that σ-algebra be?

Notation 2.7. We let $\mathcal{C}$ be the smallest σ-algebra such that all the coordinate functions $w \mapsto w(t)$ (for any fixed t, this defines a function $C(\mathbb{R}_+, \mathbb{R}) \to \mathbb{R}$, by taking the value of our function at time t) are measurable.

This is some σ-algebra. It's an exercise on the first homework that you can alternatively describe this as a Borel σ-algebra, with respect to the natural metric on this space (the one giving uniform convergence on compact sets). We'll maybe never use this equivalence in the course — the fact that this σ-algebra is equal to the Borel one — but if you want to learn more about Brownian motion, it's quite important. As an side, one fact about Brownian motion is that if you take rescaled random walks — you take discrete-time random walks at time 0 through n, but you rescale them to be a continuous function on $[0,1]$, doing linear interpolation — these rescaled random walks can also be thought of as random variables in our space (of continuous functions on $[0,1]$), which we'll call $(S_t^n)_{t\in[0,1]}$.

And one fact, called Donsker's invariance principle, is that in a sense $S^n \xrightarrow{d} B$. What does this mean? Last semester we covered convergence in distribution for *real*-valued random variables. Here these random variables are not real-valued but function-valued; but it turns out you can make the theory of convergence in distribution work verbatim. The key property is that $C(\mathbb{R}_+,\mathbb{R})$ is a complete separable metric space (this is what you need for the theorems about convergence in distribution for $\mathbb{R}$-valued randomv ariables to carry over).

Generally for such things, you want to take the Borel σ-algebra; that's the natural one to put. So that's why you might care about the Borel σ-algebra.

But in this class we're not spending as much time on Brownian motion as other classes might, so this probably won't come up too much.

Claim 2.8 — Given a Brownian motion B, the mapping $(\Omega,\mathcal{F})\to(\mathcal{C}(\mathbb{R}_+,\mathbb{R}),\mathcal{C})$ given by $\omega\mapsto(B_t(\omega))_t$ is measurable.

When we constructed Brownian motion, we thought of it as a stochastic process. But now we're saying you can also think about it as a measurable map between these spaces.

Why is it measurable? It's basically because $\omega\mapsto B_t(\omega)$ is measurable for all t, as a map $(\Omega,\mathcal{F})\to(\mathbb{R},\mathcal{B})$. So we have that $\omega\mapsto(B_t(\omega))_t$ is measurable. This is because to check this map is measurable, by the definition of the σ-algebra, we just need to check each coordinate function is a random variable. And that's true by the definition of a stochastic process — every random variable in our stochastic process is a random variable.

So $\mathcal{C}$ is the natural σ-algebra to place on our space such that you get a measurable map.

Now we can define the *law* of Brownian motion.

Definition 2.9. The Wiener measure, denoted by W, is the law of Brownian motion B as a random variable taking values in $(C(\mathbb{R}_+,\mathbb{R}),\mathcal{C})$.

(Le Gall uses the letter W; this unfortunately clashes with Sky's notation for white noise.)

In other words, for every $A\in\mathcal{C}$, we ddefine

$$W(A)=\mathbb{P}[B\in A].$$

(That's usually what the law of a random variable means.) You could also say it's the pushforward of $\mathbb{P}$ under B.

§2.3.1 The canonical construction

One foundational observation is that W does not depend on the Brownian motion you chose to define W — it doesn't depend on the underlying probability space $(\Omega,\mathcal{F},\mathbb{P})$. Why not? Well, you can consider a collection of sets which generate $\mathcal{C}$; the most natural collection are the *cylindrical sets*

$$A=\{w\in C(\mathbb{R}_+,\mathbb{R})\mid(w(t_1),\dots,w(t_n))\in A_1\times\cdots\times A_n\}$$

(where A_i's are some fixed Borel subsets of $\mathbb{R}$, and the t_i's are fixed times). It's a general measure theory fact that cylindrical sets of this form generate $\mathcal{C}$; this basically follows from the definition of $\mathcal{C}$ itself. Definitely if one of these coordinate functions $w(t)$ is measurable, then any event A when $n = 1$ must be measurable (by assumption), and then you can take a finite intersection of such events to get events of this form. So the σ-algebra generated by collections of events of this form is exactly $\mathcal{C}$ itself.

And if we evaluate W on such a cylindrical event, by definition this just depends on the finite-dimensional distributions of your Brownian motion — i.e.,

$$W(A) = \mathbb{P}[B_{t_1} \in A_1, \ldots, B_{t_n} \in A_n].$$

And by definition, the finite-dimensional distributions of Brownian motion don't depend on your underlying probability space; they're just some explicit multivariate Gaussian. So that's why this is true — why this measure W doesn't depend on the underlying probability space.

That leads to something called the *canonical construction* of Brownian motion, which we may use later on (but for now it's just an isolated definition).

Definition 2.10. The canonical construction of Brownian motion is given by choosing $(\Omega, \mathcal{F}, \mathbb{P}) = (\mathcal{C}(\mathbb{R}_+, \mathbb{R}), \mathcal{C}, W)$, and

$$B_t(\omega) = \omega(t)$$

(this is a Brownian motion).

So I could choose my probability space to be precisely this (the continuous functions equipped with measure W).

This is always kind of confusing. But ω is an element of Ω, which we are choosing to be the space of continuous functions; and we're trying to define a random variable on this space. We define this random variable to just be the coordinate function itself, which is measurable by the definition of $\mathcal{C}$. So you can certainly define a collection of random variables in this way. And the claim is that this collection of random variables is precisely a Brownian motion. Why is that true? Brownian motion is supposed to have two properties — be a pre-BM and have continuous sample paths. The second is true by construction — the sample paths are these functions $\omega(t)$, which are continuous (because ω takes values in the space of continuous functions). And it's pre-BM because you can unwind the cylindrical sets thing to get that the finite-dimensional distributions are correct.

We'll probably see this later on; sometimes it's nice to assume your Brownian motion is constructed in this way. But we probably won't see that until we get to stochastic integration.

§2.4 Sample path properties

In the remaining six minutes, we'll talk about some sample path properties of Brownian motion. To set this up, we'll fix some Brownian motion and consider the associated continuous-time filtration $\mathcal{F}_t = \sigma(B_s \mid s \leq t)$. (We will talk about continuous-time filtrations and martingales later.)

We'll define $\mathcal{F}_{0+} = \bigcap_{t>0} \mathcal{F}_t$.

Intuitively, how do you think about these? Intuitively $\mathcal{F}_t$ is all the information you've seen up to time t in your Brownian motion; and $\mathcal{F}_{0+}$ is all the information you see infinitesimally after time 0.

Theorem 2.11 (Blumenthal's 0–1 law)

The σ-algebra $\mathcal{F}_{0+}$ is trivial, i.e., $\mathbb{P}[A] \in \{0, 1\}$ for all $A \in \mathcal{F}_{0+}$.

Proof. If you recall the proof of Kolmogorov's 0–1 law for the tail σ-algebra for a sequence of IID random variables, all you want to show is that $\mathcal{F}_{0+}$ is independent of itself — because once you have this, you get that for any event $A \in \mathcal{F}_{0+}$, we have

$$\mathbb{P}[A] = \mathbb{P}[A \cap A] = \mathbb{P}[A]^2,$$

which means $\mathbb{P}[A] \in \{0, 1\}$.

How do we show it's independent of itself? Let's fix $A \in \mathcal{F}_{0+}$, and fix some sequence of points $0 < t_1 < \cdots < t_n$ and a bounded measurable function $g \colon \mathbb{R}^n \to \mathbb{R}$. Then we can compute

$$\mathbb{E}[\mathbf{1}_A g(B_{t_1}, \ldots, B_{t_n})].$$

To show independence, we eventually want to show that this is the product of expectations; and this is kind of what we're leading towards.

First, what you can do is by continuity of sample paths, you can write this as

$$\lim_{\varepsilon \to 0} \mathbb{E}[\mathbf{1}_A g(B_{t_1} - B_\varepsilon, \ldots, B_{t_n} - B_\varepsilon)]$$

(here we're just using the fact that $B_\varepsilon \to 0$ as $\varepsilon \to 0$, and the fact that g is bounded and measurable so you can use bounded convergence to swap the limits). But htis itself is going to be equal to

$$\lim_{\varepsilon \to 0} \mathbb{P}[A]\mathbb{E}[g(B_{t_1} - B_\varepsilon, \ldots, B_{t_n} - B_\varepsilon)].$$

Why? We have $A \in \mathcal{F}_\varepsilon$ (because $A \in \mathcal{F}_{0+}$, which is contained in $\mathcal{F}_\varepsilon$); and by the Markov property, we know that $(B_t - B_\varepsilon)_{t \geq \varepsilon}$ is independent of $\mathcal{F}_\varepsilon$ (we saw a Markov property of pre-BM, so it's also true for Brownian motions). And then you can back out the limit and get that this is equal to

$$\mathbb{P}[A]\mathbb{E}[g(B_{t_1}, \ldots, B_{t_n})].$$

After you get this, it's not so hard to show independence, but maybe we'll see that next time. □

§3 February 10, 2025

Today we'll prove some sample path properties of Brownian motion. And time permitting, we'll talk about the strong Markov property.

§3.1 Sample path properties

Here the theme is that despite being continuous, Brownian motion can be quite wild. We'll prove some properties that are kind of hard to exhibit if you were to try to directly construct a function — we're challenged to try constructing a function by hand that has the properties we'll see. This is in a sense reminiscent of the probabilistic method, where you want to show the existence of a combinatorial object with some properties; and you do so by picking a random object and showing that with positive probability, it has those properties. This is somewhat reminiscent — if you tried constructing directly a function satisfying these properties, it might be hard. But instead you can just kind of pick a random function. And it's not even just with positive probability, but *almost sure*, that Brownian motion will have these properties.

The first thing which is not a sample path property, but will be useful when proving them:

Definition 3.1. Let $\mathcal{F}_t = \sigma(B_s \mid s \leq t)$, and $\mathcal{F}_{0+} = \bigcap_{t>0} \mathcal{F}_t$.

Intuitively, $\mathcal{F}_{0+}$ consists of all the information you see infinitesimally after time 0.

Theorem 3.2 (Blumenthal's 0–1 law)

The σ-algebra $\mathcal{F}_{0+}$ is trivial (i.e., every event has probability 0 or 1).

Proof. The proof technique (as with every 0–1 law) is to show that $\mathcal{F}_{0+}$ is independent of itself. Fix $A \in \mathcal{F}_{0+}$ and times $0 < t_1 < \cdots < t_n$, and some bounded continuous function $g : \mathbb{R}^n \to \mathbb{R}$.

Then we can consider $\mathbb{E}[\mathbf{1}_A g(B_{t_1}, \ldots, B_{t_n})]$. Why are we looking at something like this? We want to show this is the product of expectations. Once we show this, we're done — functions of all finite collections $(B_{t_1}, \ldots, B_{t_n})$ are going to generate $\sigma(B_t \mid t > 0)$, which is equal to $\sigma(B_t \mid t \geq 0)$. (This is true because by definition, your Brownian motion is 0 at $t = 0$, so B_0 is just a constant — there's no information you gain or lose by adding B_0.) So if we can show this factors as the product, then we'll get that this σ-algebra is independent of $\mathcal{F}_{0+}$, which is something even stronger than what we needed.

This is a kind of perturbation trick which we'll see quite often — what you want to do is kind of apply the Markov property of Brownian motion, because that's the one independence property we know of so far. To do so, wee can write this as

$$\lim_{\varepsilon \to 0} \mathbb{E}[\mathbf{1}_A g(B_{t_1} - B_\varepsilon, \ldots, B_{t_n} - B_\varepsilon)]$$

(because Brownian motion is continuous and $B_0 = 0$, so the pointwise limit holds, and you can pull the limit out using bounded convergence). And then we can use the fact that $A \in \mathcal{F}_{0+} \subseteq \mathcal{F}_\varepsilon$ ($\mathcal{F}_{0+}$ stores information infinitesimally after 0, and you know that information if you know the process up to time ε). And we also know this restarted process $(B_{t+\varepsilon} - B_\varepsilon)_{t \geq 0}$ is independent of $\mathcal{F}_\varepsilon$ (we stated this for pre-BM, but it also holds for BM). So we can write this as

$$\lim_{\varepsilon \to 0} \mathbb{P}[A]\mathbb{E}[g(B_{t_1} - B_\varepsilon, \ldots, B_{t_n} - B_\varepsilon)].$$

And now we can back out the limit and get that this is equal to $\mathbb{P}[A]\mathbb{E}[g(B_{t_1}, \ldots, B_{t_n})]$, which is what we wanted.

And now we can finish as indicated before — this means $\mathcal{F}_{0+}$ is independent of $\sigma(B_t \mid t > 0) = \sigma(B_t \mid t \geq 0)$. And $\mathcal{F}_{0+} \subseteq \sigma(B_t \mid t \geq 0)$, so we're done. □

Student Question. *What do we mean when we say these expectations generate the σ-algebra?*

Answer. To check two σ-algebras are independent, it suffices to check independence on a generating collection. For instance, vectors of the form $(B_{t_1}, \ldots, B_{t_n})$ for fixed $t_1 < \cdots < t_n$ generate $\sigma(B_t \mid t > 0)$. And then you use the approximation fact: in principle, to say two σ-algebras $\mathcal{G}$ and $\mathcal{F}$ are independent, I have to say $\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B]$ for all $A \in \mathcal{G}$ and $B \in \mathcal{F}$. But rather than checking this for every event, it actually suffices to check that $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ for all bounded random variables which are $\mathcal{G}$-measurable and $\mathcal{F}$-measurable (respectively). And you don't actually have to consider *every* bounded measurable function; it's enough to consider functions of finitely many of the B_t's. (This is some sort of monotone class argument — often when checking independence, you want to check on the most convenient collection you can, and continuous is perfectly fine for that.)

Now let's look at some consequences.

Proposition 3.3

(i) We have that almost surely, for all $\varepsilon > 0$,

$$\sup_{0 \le s \le \varepsilon} B_s > 0 \quad \text{and} \quad \inf_{0 \le s \le \varepsilon} B_s < 0.$$

(ii) For any $a \in \mathbb{R}$, let $T_a = \inf\{t \ge 0 \mid B_t = a\}$. Then almost surely, $T_a < \infty$ for all $a \in \mathbb{R}$.

In other words, for (i), Brownian motion goes above and below 0 on any ε-interval.

First, note that we stated this for all ε, so you might think this is an uncountable intersection; but it suffices to just consider a sequence $\varepsilon_n \swarrow 0$, so there's no issues with uncountability.

Again (ii) looks like an uncountable intersection, but because BM has continuous sample paths, if you go above a then you must have also gone above $\lfloor a \rfloor$; so we could've just stated this for all integers instead. So it's again a countable intersection.

Note that (ii) has the consequence

$$\limsup_{t \to \infty} B_t = \infty,$$

and similarly $\liminf_{t \to \infty} B_t = -\infty$ (this isn't about probability; it's just that if you have a continuous function with the property (ii), then you have to oscillate infinitely often between $+\infty$ and $-\infty$).

It's maybe not hard to construct a function satisfying (ii), but it's quite hard to construct a function satisfying (i) — on any interval, you go above and below 0 infinitely many times. So there's never a first time you go above or below 0; this is quite hard to construct.

Proof of (i). Consider some sequence $\varepsilon_n \searrow 0$, and consider the event

$$A = \bigcap_{n \ge 1} \left\{ \sup_{0 \le s \le \varepsilon_n} B_s > 0 \right\} = \bigcap_{n \ge 1} E_n$$

(this is a countable intersection, where the nth event E_n is the event that your Brownian motion goes above 0 on $[0, \varepsilon_n]$). First, we can see that $E_n \supseteq E_{n+1}$ — we took ε_n to be decreasing, so in E_{n+1} we're taking a sup over a smaller interval, which means

$$\sup_{0 \le s \le \varepsilon_{n+1}} B_s \le \sup_{0 \le s \le \varepsilon_n} B_s.$$

First, note that $A \in \mathcal{F}_{0+}$. Intuitively, this is because it only depends on your BM infinitesimally after 0. To make this precise, these events are increasing, so instead of starting this intersection at 1, you could have started at any N. And by considering that, you get that $A \in \mathcal{F}_{\varepsilon_N}$ for all N, and by definigion of $\mathcal{F}_{0+}$, this means it has to be in $\mathcal{F}_{0+}$.

And because we've got a decreasing intersection, by continuity of probability you have $\mathbb{P}[A] = \lim_n \mathbb{P}[E_n]$.

What are we doing? We want to show $\mathbb{P}[A] = 1$. And you want to apply Blumenthal's 0–1 law, so we just need to show the right-hand side is greater than 0. What's one soft way of showing it's greater than 0? You can just lower-bound it by

$$\lim_n \mathbb{P}[E_n] \ge \limsup_n \mathbb{P}[B_{\varepsilon_n} > 0]$$

(we've got a sup, so if $B_{\varepsilon_n} > 0$, then the sup over the interval is also greater than 0). But B_{ε_n} is a normal distribution centered at 0, so you can explicitly compute this, and it's actually just $\frac{1}{2}$. So Blumenthal implies $\mathbb{P}[A] = 1$.

And then to get the statement about inf, you can just replace B with $-B$ (this is the reflection symmetry of Brownian motion — you could have applied this result to $-B$, and if you unpack what that says, you get the inf part of the claim). □

Student Question. *Why is $A \in \mathcal{F}_{0+}$?*

Answer. Because the events E_n are decreasing, you have $\bigcap_{n\geq 1} E_n = \bigcap_{n\geq N} E_n$ for all N. And this is in $\mathcal{F}_{\varepsilon_N}$, since you're only looking at your BM on the time interval $[0, \varepsilon_N]$. So this implies $Ain \bigcap_{N\geq 1} \mathcal{F}_{\varepsilon_N} = \mathcal{F}_{0+}$.

Student Question. *Why is $\lim_n \mathbb{P}[E_n] \geq \limsup_n \mathbb{P}[B_{\varepsilon_n} > 0]$?*

Answer. We have $\mathbb{P}[E_n] \geq \mathbb{P}[B_{\varepsilon_n} > 0]$. The left-hand side has a limit; the right-hand side may or may not, but that's not an issue. Then this is just a fact about real numbers, that if $x_n \geq y_n$ and $\lim x_n$ exists, then $\lim x_n \geq \limsup y_n$. (Here in fact the limit does exist — it's just $\frac{1}{2}$ — but you don't actually need that.)

Proof of (ii). Let's first be a bit less ambitious and just try to show that BM goes above 1 almost surely — i.e., that

$$\mathbb{P}\left[\sup_{s\geq 0} B_s > 1\right] = 1.$$

There's a scaling argument that will show B_s goes above any finite number almost surely, but let's first try to show this.

To show this, we again use scaling symmetry. Fix $\lambda > 0$. Then we have this space-time scaling — we can define a new Brownian motion $B_t^\lambda = \lambda^{-1} B_{\lambda^2 t}$. (The way you remember this scaling is that at a fixed time t, I should be $\mathcal{N}(0, t)$.) There's the scaling symmetry that (B_t^λ) is also a Brownian motion. And what's the consequence of that? Again, intuitively you're just saying that if you're asking about a property of BM on some time interval (e.g., whether it goes above a threshold), that's the same as asking if it goes above a smaller threhsold on a smaller time-interval (think of λ as small). If the original BM goes above 1, then this rescaled BM goes over λ.

But now we're just asking that BM goes above 1 on the *infinite* time interval, so rescaling the time interval does nothing. So the question of whether BM goes above 1 is the same as whether it goes above λ — in other words, our original probability is

$$\mathbb{P}\left[\sup_{s\geq 0} B_s^\lambda > 1\right] = \mathbb{P}\left[\sup_{s\geq 0} B_{\lambda^2 s} > \lambda\right] = \mathbb{P}\left[\sup_{s\geq 0} B_s > \lambda\right]$$

(we're on an infinite time interval, so we don't care about the λ^2). So we get that

$$\mathbb{P}\left[\sup_{s\geq 0} B_s > 1\right] = \mathbb{P}\left[\sup_{s\geq 0} B_s > 0\right]$$

(here we're taking a sequence $\lambda_n \to 0$ and taking a limit of the above argument). But the right-hand side is 1 by (i) (in fact, we know on every tiny interval it goes above 0 infinitely often).

So we get the statement for 1. And the proof immediately suggests how you upgrade this to any positive number — again by scaling, we have

$$\mathbb{P}[\sup_{s\geq 0} B_s > 1] = \mathbb{P}[\sup_{s\geq 0} B_s > n]$$

for all n (you can just take λ to be n in this scaling symmetry).

(Technically, the statement is about the times T_a; but the event that T_a is finite is exactly the same as the event $\sup_{s\geq 0} B_s \geq a$, probably using the fact that your Brownian motion is continuous and starts at 0.)

So we've showed that BM goes above every level almost surely. That works for positive a's; for negative a's you use reflection symmetry (replacing B by $-B$). □

Student Question. *Could you also use some chaining argument with Bayes's rule, e.g. showing that*

$$\mathbb{P}[\sup B_s \geq 2 \mid \sup B_s \geq 1]?$$

Answer. Maybe, but this isn't *a priori* clear. You want to follow your BM until it hits 1, and then restart at this random time when it hits 1, and ask whether this restarted BM goes above 1. But this is kind of using the strong Markov property that we'll get to later. This is a *random* time, so you're asking, if we restart BM at a *random* time, is it still a BM? And this is true, but we haven't proved it yet. (Specifically,we'd want to define $B_t^1 = B_{T_1+t} - B_{T_1}$; then $\{\sup B_s \geq 2\} = \{\sup B_s \geq 1\} \cap \{\sup B_s^1 \geq 1\}$.) So we don't have the tools for this yet, but it's a natural strategy that you could also use once you have the strong Markov property.

Student Question. *What does a σ-algebra being trivial mean?*

Answer. The probability of every event is either 0 or 1. (This is more general than every event being $\emptyset$ or Ω — you could have other sets of probability 0.)

Remark 3.4. Some comments about scaling symmetry: The scaling $\lambda^{-1}B_{\lambda^2 t}$ is quite special. If you've taken PDE, this is parabolic scaling; so you might think Brownian motion is to the heat equation (which has the same parabolic scaling symmetry). And in fact, this is true. Solutions to the heat equation model how heat spreads. If you think about it at the particle level, the heat equation is about the distribution of particles, but if you look at individual particles they're supposed to be following Brownian motion.

As a corollary of (i) of Proposition 3.3:

Corollary 3.5

Almost surely, the sample path $t \mapsto B_t$ is not monotone on any nonempty interval.

In other words, if you draw the graph of your BM, there's no interval where it's monotone. Again, it's not so easy to try to construct a function with this property.

Proof. The proof is kind of immediate from (i) and the Markov property — by the Markov property and (i), we obtain that almost surely, for all $q \in \mathbb{Q}$ with $q > 0$ (i.e., rational times) and all $\varepsilon > 0$, we have

$$\sup_{q \leq s \leq q+\varepsilon} B_s > B_q \quad \text{and} \quad \inf_{q \leq s \leq q+\varepsilon} B_s < B_q.$$

(Explicitly, you bring B_q to the other side, so the first statement essentially says that $\sup_{q \leq s \leq q+\varepsilon}(B_s - B_q) > 0$, and you can write this as a statement about restarted Brownian motion at time q.) And we took a countable collection of times, so we're just taking a countable intersection of almost sure events, which is still almost sure. And this implies the statement we want, by an analysis fact. □

The next property is less about sample paths, but we'll use it to prove something about sample paths.

Proposition 3.6

Fix $t > 0$, and let $0 = t_0^n < t_1^n < \cdots < t_{p_n}^n = t$ be a sequence of subdivisions of $[0, t]$ with mesh size tending to 0 (as $n \to \infty$). Then we have

$$\sum_{i=1}^{p_n} (B_{t_i^n} - B_{t_{i-1}^n})^2 \xrightarrow{L^2} t.$$

So we're fixing t and taking a sequence of subdivisions of the interval $[0,t]$. The way you visualize it is you take $[0,t]$ and put down a bunch of points, where the first point is 0 and the last is t; and in the middle you have a bunch of points. The *mesh size* is the maximum distance between any two consecutive points; so the subdivision gets increasingly finer.

Why would you expect this? If you took the expectation of the left-hand side, the expectation of the increment squared is precisely $t_i^n - t_{i-1}^n$. And then you can telescope; you're summing the lengths of all intervals, so you get t. So the *expectation* of the left-hand side is t; and L^2 convergence basically says that the *variance* of the left-hand side converges to 0.

Proof. The first thing you'd try is to literally write out what this means — you basically write out

$$\mathbb{E}\left[\left(\sum_{i=1}^{p_n}(B_{t_i^n} - B_{t_{i-1}^n})^2 - t\right)^2\right]$$

(this is the variance, and we want to show this goes to 0). And the point is this is just a sum of squares of normal random variables, so you can just compute. First, we can write this as

$$\mathbb{E}\left[\left(\sum_{i=1}^{p_n}(B_{t_i^n} - B_{t_{i-1}^n})^2\right)^2\right] - t^2$$

(using the fact that $\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$, and here we just computed that the first moment is t). When you expand this out, first you get a fourth power from when there's no cross-terms — i.e., something of the form $(B_{t_i^n} - B_{t_{i-1}^n})^4$. And then you also get cross-terms. So you end up with

$$\sum_{i=1}^{p_n}(B_{t_i^n} - B_{t_{i-1}^n})^4 + \sum_{1\le i\ne j\le p_n}\mathbb{E}[(B_{t_i^n} - B_{t_{i-1}^n})(B_{t_j^n} - B_{t_{j-1}^n})^2] - t^2.$$

And you shouldn't be afraid of any of these things, because they're just normals, which means we have exact formulas for these things. If we call the first sum I_1, then we have

$$I_1 \lesssim \sum_{i=1}^{n}(t_i^n - t_{i-1}^n)^2$$

(we're basically taking the fourth moment of a Gaussian whose variance is $t_i^n - t_{i-1}^n$ — if Z is a standard normal, then $\sigma^{1/2}Z$ is a normal with variance σ, and $\mathbb{E}[(\sigma^{1/2}Z)^4] = \mathbb{E}[Z^4]\sigma^2 = 3\sigma^2$; so the actual constant above is 3, but that's not relevant). Now we want to bound this, so we can use the fact that our mesh size is tending to 0; so we can bound this by

$$\max_i |t_i^n - t_{i-1}^n| \sum_{i=1}^{p_n}(t_i^n - t_{i-1}^n)$$

(basically, L^∞ and L^1 controls L^2). But then the sum is precisely t by telescoping, and the max tends to 0. So that handles I_1, and so I_1 is fine.

Now let's consider I_2 (the sum over $i \ne j$, together with the $-t^2$). For this, you first use the fact that because these are all disjoint time intervals, we have independence of increments; so we can actually write this expectation of products as a product of expectations. And then we're taking the product of variances of normals. So

$$I_2 = \sum_{1\le i\ne j\le p_n}(t_i^n - t_{i-1}^n)(t_j^n - t_{j-1}^n) - t^2.$$

And we want to show this goes to 0 (so we need a way to cancel out this t^2).

But the first sum is all cross-terms, so we could have written it as

$$\left(\sum_{i=1}^{p_n}(t_i^n - t_{i-1}^n)\right)^2 - \sum_{i=1}^{p_n}(t_i^n - t_{i-1}^n)^2 - t^2$$

(you can imagine expanding the first thing out, and then the second thing subtracts out the non-cross terms).

But the first sum telescopes, so it gives t^2, which cancels out the $-t^2$. And we just bounded $\sum_{i=1}^{p_n}(t_i^n - t_{i-1}^n)^2$ (we showed it goes to 0). So $I_2 \to 0$, as desired. □

Student Question. *Does this also work for pre-BM?*

Answer. Yes, since it only looks at finite-dimensional distributions.

The summary is that whenever you have Gaussian random variables, you can kind of just do the first thing you'd think to do — to compute everything — and you can do that.

Student Question. *Why is there not a 2 in front of the sum for I_2?*

Answer. We're not assuming i and j are ordered in the sum.

Now we can use this to prove the following property of Brownian motion. First, we need to define a concept.

Definition 3.7. We say a function $f\colon [a,b] \to \mathbb{R}$ has *infinite variation* if the sup of the quantiites

$$\sum_{i=1}^{p} |f(t_i) - f(t_{i-1})|$$

over all subdivisions of $[a,b]$ is infinite.

This is kind of some discrete approximation of the variation of your function; the fact that this is infinite is the source of the name 'infinite variation.'

Remark 3.8. What are functions where the sup is finite? Well, continuously differentiable functions — if f' were continuous, then you could bound this by $\|f'\|_{L^\infty}\sum_{i=1}^{p}|t_i - t_{i-1}| = (b-a)\|f'\|_{L^\infty}$. So continuously differentiable functions *don't* have infinite variation.

Proposition 3.9

Almost surely, Brownian motion has infinite variation on any nonempty interval.

In particular, Brownian motion cannot be continuously differentiable on any nonempty interval. Again, it's hard to construct a function like this (which is continuous, but not continuously differentiable on any nonempty interval).

Proof. It suffices to prove this on some interval $[0,t]$ — by the Markov property, you can then extend this to any interval (not necessarily starting at 0) — this only looks at the increments, so you can look at the restarted Brownian motion. And the reason you can fix an interval is because we can take a countable intersection of intervals with rational endpoints; any interval has a smaller interval with rational endpoints contained in it, and if you have infinite variation on a smaller interval, you also have infinite variation on the larger one. So you can just reduce to this case.

And the fact that the quadratic variation (which we computed earlier) is positive is basically incompatible with having finite variation — if you had finite variation, then this thing would have to be 0. That's basically the same as our earlier argument, more or less — if you have finite variation, then on any interval, your variation is basically the length of that interval. Then when you square, you're getting the square of the length of that interval; and we just showed that goes to 0.

So that's some heuristic reason for why if the function has finite variation, then our quadratic sum from earlier has to be 0. But now we have to actually prove the contrapositive of that. So how do we do that?

On the left-hand side, let's look at our quadratic variation

$$\sum_{i=1}^{p}(B_{t_i} - B_{t_{i-1}})^2.$$

First, we can upper-bound this by

$$\sup_i |B_{t_i} - B_{t_{i-1}}| \cdot \sum_{i=1}^{p} |B_{t_i} - B_{t_{i-1}}|.$$

(We have two factors on the left; we bound one of them by the sup, and then we sum over the rest.)

Now by the L^2 convergence that we showed, we already showed that the left-hand side converges in L^2 to $t > 0$. That implies it also converges in *probability* to t. And if you have a sequence converging in probability, you can find a subsequence converging almost surely (by some Borel–Cantelli trick from 18.675).

So you can obtain a sequence of subdivisions (t_i^n) such that

$$\sum_{i=1}^{p_n}(B_{t_i^n} - B_{t_{i-1}^n})^2 \to t.$$

(We started with any sequence with mesh size going to 0, got convergence in probability, and then used Borel–Cantelli to get an almost surely converging subsequence.)

And now you're basically done — first, we have

$$\sup_i |B_{t_i^n} - B_{t_{i-1}^n}| \to 0$$

surely (in fact, not just almost surely), because we're taking our subdivisions to have mesh size going to 0, and BM is uniformly continuous on any compact interval. (A continuous function on $[0, t]$ is uniformly continuous, and the mesh size is going to 0, so this thing must go to 0.)

And that has to imply

$$\sum_{i=1}^{p_n} |B_{t_i^n} - B_{t_{i-1}^n}| \to \infty$$

(because the left-hand side goes to something strictly positive, and the first term goes to 0; so the last term has to go to ∞). □

Student Question. *How did we get the existence of a subsequence which converges almost surely?*

Answer. What you kind of want to say is — if we have a sequence $X_n \xrightarrow{p} X$, by definition you're saying $\mathbb{P}[|X_n - X| \geq \varepsilon] \to 0$ for all $\varepsilon > 0$. And the thing to remember is that we can obtain a subsequence (X_{n_k}) such that

$$\mathbb{P}[|X_{n_k} - X| \geq 2^{-k}] \leq 2^{-k}$$

(so we're first picking $\varepsilon = 2^{-k}$, and because the probability goes to 0, we can find some n large enough that it's less than 2^{-k}). And we set this up so that this sequence of probabilities is summable, so Borel–Cantelli I tells us that almost surely $|X_{n_k} - X| \leq 2^{-k}$ for sufficiently large k; and that gives that

our sequence converges.

Remark 3.10. This trick is good to remember — we also used something like this in the proof of Kolgomorov's continuity criterion (the fact that you can upgrade convergence in probability to almost sure, using some exponential decay trick).

Student Question. *Is this theorem still true if we have the additional condition that the mesh size goes to* 0*, instead of taking a* sup *over all subdivisions? In other words, is it true that for* any *sequence of subdivisions with mesh size going to* 0*, the corresponding variation should be infinite?*

Answer. Here we're starting with some sequence with mesh size going to 0; we can find some subsequence which converges almost surely, and along that subsequence we get infinite variation. However, that doesn't imply it for the original sequence. Still, Sky thinks the answer should be yes.

§3.2 The strong Markov property of Brownian motion

Now we can discuss elements of the strong Markov property, so let's move on to that.

What does this basically say? First, the Markov property says that if we fix a (deterministic) time t and start observing your Brownian motion from this deterministic time (so you kind of restart it), that's still a Brownian motion.

In the strong Markov property, now you restart at a *random* time T; and still, the restarted motion is Brownian. That's basically the statement of the SMP.

But of course, T can't just be any arbitrary random time. It has to satisfy some conditions — in particular, it has to be a *stopping time*.

§3.2.1 Stopping times

Definition 3.11. We write $\mathcal{F}_\infty = \sigma(B_t \mid t \geq 0)$.

(This is all the information in your BM at all times.)

Recall that we also write $\mathcal{F}_t = \sigma(B_s \mid s \leq t)$.

Now we'll define stopping times. We probably saw them defined in discrete time in 18.675, but here we'll define them for continuous time (here we'll do it for this specific filtration; then we'll move onto continuous-time martingales later, where we'll see the general definition — it's just this, but for a general filtration).

Definition 3.12. A random variable T with values in $[0, \infty]$ is called a stopping time if for all $t \geq 0$,

$$\{T \leq t\} \in \mathcal{F}_t.$$

(Note that T is allowed to be ∞.) The motivation for this is the same as in discrete-time — you think of yourself as observing the stochastic process as you go, and at some point based on the values of your process, you decide to stop. In particular, you have to decide to stop based on only what you've seen so far, and not based on the stuff you haven't seen. That's basically what this says — whether you stop by time t only depends on what has happened up to time t.

Note that if T is a stopping time, then for the event that you stop *strictly* before time t, we can write

$$\{T < t\} = \bigcup_{\substack{0 \leq q < t \\ q \in \mathbb{Q}}} \{T \leq q\} \in \mathcal{F}_t$$

(since each of these events is in $\mathcal{F}_q \subseteq \mathcal{F}_t$).

Student Question. *Earlier, we proved that BM hits each individual point almost surely. Is it actually true that it hits* all *points almost surely — i.e., that it covers the number line?*

Answer. Yes. For instance, you can take the integers 1, 2, 3, 4, 5, ...; we know BM hits each of these integers at some point. And then you can take the countable intersection of these events. (In particular, the fact that the lim sup is ∞ and the lim inf is $-\infty$ implies that it hits any level.)

Student Question. *Could you then make a modification of BM such that every hitting time would be finite?*

Answer. You probably could, but it's less clear whether you'd ever need that in a proof. We will see examples where you want to modify things so that something happens surely, but maybe for other properties and not exactly this one.

Example 3.13

Any deterministic time $T = t_0$ is a stopping time.

Intuitively, you don't have to look at your process at all to determine whether to stop. To verify this from the definition, we have

$$\{T \leq t\} = \{t_0 \leq t\} = \begin{cases} \emptyset & t < t_0 \\ \Omega & t \geq t_0 \end{cases}$$

(both of which are in $\mathcal{F}_t$).

Example 3.14

The time $T = T_a = \inf\{t \geq 0 \mid B_t = a\}$ (as defined earlier) is a stopping time.

This again uses the characterization of $\{T_a \leq t\}$ in terms of the sup of your BM. Assume $a > 0$ (otherwise you use the inf); then we have

$$\{T_a \leq t\} = \left\{\sup_{s \leq t} B_s \geq a\right\}.$$

And $\sup_{s \leq t} B_s$ is $\mathcal{F}_t$-measurable as a random variable (you can use continuity to replace this with a countable sup — you can restrict to $q \leq t$ with $q \in \mathbb{Q}$ — and each B_q is $\mathcal{F}_q$-measurable and therefore $\mathcal{F}_t$-measurable, and it's a general fact that for any sequence of random variables X_n, $\sup_n X_n$ is a random variable).

These are some first immediate examples of stopping times. Another, which we'll use later on:

Example 3.15

If T is any stopping time and t_0 a deterministic time, then $T + t_0$ is also a stopping time — we have

$$\{T + t_0 \leq t\} = \{T \leq t - t_0\} \in \mathcal{F}_{(t-t_0)\vee 0} \subseteq \mathcal{F}_t.$$

More generally, the sum of two stopping times is also a stopping time (though this requires slightly more argument).

Definition 3.16. Let T be a stopping time. The stopped σ-algebra $\mathcal{F}_T$ is defined as

$$\mathcal{F}_T = \{A \in \mathcal{F}_\infty \mid A \cap \{T \leq t\} \in \mathcal{F}_t \text{ for all } t\}.$$

Intuitively, you think of this as all the information you've seen up to your random time T — this is just saying that on the event that you've stopped by time t, you can determine A by just the information up to time t (in other words, you get all the information you've seen up to the random stopping time).

Exercise 3.17. This set $\mathcal{F}_T$ is a σ-algebra, and T is $\mathcal{F}_T$-measurable.

We've just defined $\mathcal{F}_T$ as a collection of sets, so this needs to be verified. The second statement intuitively makes sense — $\mathcal{F}_T$ encodes all the information we've seen up to a random time, and if we know all the information up to that time, we should know that time. But you have to check this.

There's a bunch of preliminaries we need to set up to talk about the strong Markov property rigorously; this is part of the setup, and next class we'll do the rest and state the strong Markov property.

§4 February 12, 2025

Today we'll state and prove the strong Markov property of Brownian motion and see an application to the reflection principle.

§4.1 Strong Markov property

We've constructed Brownian motion, and defined a filtration and stopping times with respect to it; recall that given a stopping time T, we can define the *stopped σ-algebra*

$$\mathcal{F}_T = \{A \in \mathcal{F}_\infty \mid A \cap \{T \le t\} \in \mathcal{F}_t \text{ for all } t \ge 0\},$$

which is heuristically all the information you've seen up to your stopping time.

Now we want to consider the value of our Brownian motion at our random stopping time — so we let

$$\mathbf{1}_{T<\infty} B_T = \begin{cases} B_{T(\omega)}(\omega) \mid T < \infty \\ 0 \qquad\qquad\qquad T = \infty \end{cases}$$

(this is a function on Ω).

Claim 4.1 — $\mathbf{1}_{T<\infty} B_T$ is $\mathcal{F}_T$-measurable.

So it's not just a random variable, but even measurable with respect to your information up to time t.

This is not obvious — by construction B_t is measurable for any fixed t, but now you're taking your Brownian motion at a random time. We'll use this type of proof again and again. To show something's measurable, you typically want to show it's the limit of more simple indicator functions or linear combinations of indicators; those should be more easily verified to be measurable, and the limit of measurable functions is measurable. So you basically want a discrete approximation.

Proof. Fix $n \ge 1$; then we can approximate B_T by the function

$$X_n = \sum_{j=1}^{\infty} \mathbf{1}\left[\frac{j-1}{n} \le T < \frac{j}{n}\right] B_{(j-1)/n}.$$

So we're basically approximating based on the values of our stopping time — if T is finite, then it has to lie in one of these intervals of length $1/n$, and you take B at the left endpoint. We have to show this itself is a random variable.

The idea is that we have to say the X_n's are $\mathcal{F}_T$-measurable. For that, it's very important we're taking the *left* endpoint. If T is greater than the left endpoint, then I 'know' the value of $B_{(j-1)/n}$, because this time comes *before* my stopping time, rather than after.

First, $X_n \to \mathbf{1}_{T<\infty} B_T$ as $n \to \infty$ pointwise (for all $\omega \in \Omega$). So it suffices to show that $X_n \in \mathcal{F}_T$ (i.e., that X_n is $\mathcal{F}_T$-measurable) for all n (then we can use the fact that a pointwise limit of measurable functions is measurable). And because X_n is a linear combination of functions, it suffices to show that each of those individual functions is measurable (technically it's an infinite sum, but we can again use the pointwise limit fact).

To see this, fix $0 \le s \le t$. Let's consider the function

$$\mathbf{1}(s \le T < t) B_s$$

(here we're just replacing $(j-1)/n$ and j/n with general s and t). We want to say this is $\mathcal{F}_T$-measurable; how do we do this?

Let's take some Borel set $A \in \mathcal{B}$ (we write $\mathcal{B}$ for $\mathcal{B}(\mathbb{R})$); we want to show that $\{\mathbf{1}(s \le T < t) B_s \in A\}$ is in $\mathcal{F}_T$.

What does it mean for this event to be in the stopped σ-algebra $\mathcal{F}_T$? I hav to verify that if I intersect this with the event $\{T \le u\}$, I have to show that this is in $\mathcal{F}_u$ — so I have to show that for all $u \ge 0$,

$$E = \{\mathbf{1}(s \le T < t) B_s \in A\} \cap \{T \le u\} \in \mathcal{F}_u.$$

(This is what it means to be $\mathcal{F}_T$-measurable; here we're just unpacking the definition.)

Let's assume $0 \notin A$; this will be slightly more convenient. Under this condition, we can write E on a case-by-case basis. If $u < s$, then this event is just $E = \emptyset$, because we have a condition $T \le u$ and $T \ge s$. And $\emptyset$ is certainly in $\mathcal{F}_u$.

Now, if $u \le s < t$, this becomes $\{B_s \in A\} \cap \{s \le T \le u\}$. And $\{B_s \in A\}$ is in $\mathcal{F}_s$ (because B_s is $\mathcal{F}_s$-measurable by definition), and $\mathcal{F}_s \subseteq \mathcal{F}_u$. What about the second part? Here we have to use the fact that T is a stopping time, combined with the fact that we can write this as $\{T \le u\} \setminus \{T < s\}$. And last time we briefly remarked that by defiition $\{T \le u\} \in \mathcal{F}_u$, and you can still show that $\{T < s\} \in \mathcal{F}_s \subseteq \mathcal{F}_u$ (by taking an increasing union).

(Where do we use the fact $0 \notin A$? This was in saying $E = \emptyset$ when $u < s$; otherwise we might get Ω instead of $\emptyset$.)

Finally, when $u > t$, we can write this as $\{B_s \in A\} \cap \{s \le T < t\}$. By similar considerations as before, this is $\mathcal{F}_t$-measurable, and therefore $\mathcal{F}_u$-measurable.

We've only considered the case where $0 \notin A$; but the case where $0 \in A$ follows by taking complements. Specifically, you can write

$$\{\mathbf{1}(s \le T < t) B_s \in A\} = \Omega \setminus \{\mathbf{1}(s \le T < t) \in A^{\mathsf{c}}\}.$$

And A^{c} doesn't contain 0, so we know this thing is in $\mathcal{F}_T$, and when you take the complement it's still in $\mathcal{F}_T$. □

Maybe if this weren't a math course you might not care so much about showing measurability — you might just believe it's true — but for a foundational course it's good to cover the details, in case we come across it again. In practice you maybe don't worry too much about these details when doing research, but it's good to see it at least once.

So now we've defined Brownian motion at your stopping time and shown it's a random variable, and actually $\mathcal{F}_T$-measurable.

Theorem 4.2 (Strong Markov property)

Let T be a stopping time, and assume $\mathbb{P}[T < \infty] > 0$. For $t \geq 0$, define

$$B_t^{(T)} = \mathbf{1}(T < \infty)(B_{T+t} - B_T).$$

Then under the probability measure $\mathbb{P}[\bullet \mid T < \infty]$, $(B_t^{(T)})$ is a Brownian motion independent of $\mathcal{F}_T$.

The idea is that we define a restarted process based on your random stopping time — I start observing my Brownian increments after my random stopping time T. If T were deterministic this is the usual thing you do when looking at the Markov property; the reason for 'strong' is that we're looking at random times.

Note that we know $B_t^{(T)}$ is a random variable — $B_T \in \mathcal{F}_T$, and for B_{T+t}, a stopping time plus a deterministic time is still a stopping time. So this actually is a random variable.

We're looking at the conditional probability measure $\mathbb{P}[\bullet \mid T < \infty]$, which is defined as

$$\frac{\mathbb{P}[A \cap \{T < \infty\}]}{\mathbb{P}[T < \infty]}.$$

So this is an extension of the Markov property to these random stopping times.

Proof. For simplicity, let's first assume that $T < \infty$ almost surely, meaning that $\mathbb{P}[T < \infty] = 1$ (so that this conditioned measure is your original measure). In the course of the proof, we'll see that the proof also carries over to the more general case.

The kind of technique is that you want to perform a discrete approximation on your stopping time T, very similar to what we did before, and apply the usual Markov property. So we'll see how we do that.

First, as some setup, let's fix an event $A \in \mathcal{F}_T$; and let's also fix times $0 \leq t_1 < \cdots < t_p$. We're going to write out what it means to be independent — fix $F \colon \mathbb{R}^p \to \mathbb{R}$ which is a bounded continuous function on t real variables. Then the statement of independence just reduces to showing (similarly to what we did in Blumenthal's 0–1 law) that

$$\mathbb{E}[\mathbf{1}_A F(B_{t_1}^{(T)}, \ldots, B_{t_p}^{(T)})] = \mathbb{P}[A]\mathbb{E}[F(B_{t_1}^{(T)}, \ldots, B_{t_p}^{(T)})].$$

This would imply precisely the independence statement we want. And actually if you remove the T's, showing that

$$\mathbb{E}[\mathbf{1}_A F(B_{t_1}^{(T)}, \ldots, B_{t_p}^{(T)})] = \mathbb{P}[A]\mathbb{E}[F(B_{t_1}, \ldots, B_{t_p})],$$

that would imply that this restarted process precisely is a Brownian motion — because then you could take $A = \Omega$, and you'd be saying we have the right finite-dimensional distributions.

So how do we show this? This is where the discrete approximation comes in (plus the usual Markov property). Let's now define the process X_t^n, which you should think of as an approximation to the restarted process B_t^T. Again we're going to take indicators based on which interval my stopping time T lies in. But now I'm going to kind of do things based on the *right* endpoint — so I let

$$X_t^n = \sum_{i=1}^{\infty} \mathbf{1}\left[\frac{i=1}{n} \leq T < \frac{i}{n}\right] \cdot \left(B_{i/n+t} - B_{i/n}\right)$$

(before we needed to use the left endpoint, but here we'll actually need the right endpoint to apply the Markov property).

The first thing is we have $X_t \to B_t^{(T)}$ as $n \to \infty$ (for all t and ω); that's the whole reason why we took this approximation. Then since F is bounded and continuous, you get that

$$\text{LHS} = \lim_{n \to \infty} \mathbb{E}[\mathbf{1}_A F(X_{t_1}^n, \ldots, X_{t_p}^n)]$$

(where LHS refers to the quantity on the left-hand side of what we want to show, and we're just replacing B with X).

Now we can use the fact that

$$F(X^n_{t_1}, \dots, X^n_{t_p}) = \sum_{i=1}^{n} \mathbf{1}\left[\frac{i-1}{n} \le T < \frac{i}{n}\right] F(B^{i/n}_{t_1}, \dots, B^{i/n}_{t_p})$$

(where thee notation $B^{i/n}$ refers to what happens when you restart your motion at i/n, i.e., $B_{i/n+t} - B_{i/n}$).

And now you can input this formula into our limit; and you want to swap the infinite summation and expectation. You can justify this by some sort of dominated convergence — F is bounded, and the sum of indicators is at most 1, because T can only be in one of these intervals. So there's some sort of bounded convergence justification. And then you get that in the end, this is equal to

$$\lim_{n\to\infty} \sum_{i=1}^{\infty} \mathbb{E}\left[\mathbf{1}_A \mathbf{1}\left[\frac{i-1}{n} \le T < \frac{i}{n}\right] F(B^{i/n}_{t_1}, \dots, B^{i/n}_{t_p})\right].$$

And here's where we're going to apply the usual Markov property — by teh usual Markov property, this restarted process $B^{i/n}$ is independent of all of your information up to time i/n, i.e., of $\mathcal{F}_{i/n}$. And to get this independence, we just want to show that

$$\mathbf{1}_A \mathbf{1}\left[\frac{i-1}{n} \le T < \frac{i}{n}\right] \in \mathcal{F}_{i/n}.$$

Why is this true? First, we can write this as the indicator of an intersection, namely

$$A \cap \left\{\frac{i-1}{n} \le T < \frac{i}{n}\right\} = A \cap \left\{T < \frac{i}{n}\right\} \setminus \left\{T < \frac{i-1}{n}\right\}.$$

And the second term is in $\mathcal{F}_{(i-1)/n} \subseteq \mathcal{F}_{i/n}$ (essentially by definition — technically we have a $<$ but not $\le$, but we can again take an increasing union, so that's fine). And the first term is also in $\mathcal{F}_{i/n}$ — again you can use the definition of $\mathcal{F}_T$ and take your endpoint to be a sequence of rationals going up to i/n (essentially, we deal with the $<$ in the same way as for the second term). (We could probably avoid this issue by swapping which endpoint of the discrete approximation interval is strict.)

So this is $\mathcal{F}_{i/n}$-measurable and the remainder is independent of that, so now we can split the expectation of the product; so this becomes

$$\lim_{n\to\infty} \sum_{i=1}^{\infty} \mathbb{P}\left[A \cap \left\{\frac{i-1}{n} \le T < \frac{i}{n}\right\}\right] \mathbb{E}[F(B_{t_1}, \dots, B_{t_n})]$$

(where we dropped the i/n superscripts because by the usual Markov property, $B^{i/n}$ is distributed like a usual Brownian motion). And for the first piece, when we sum over i, what we get is really just $\mathbb{P}[A \cap \{T < \infty\}]$; so this becomes

$$\lim_{n\to\infty} \mathbb{P}[A \cap \{T < \infty\}] \mathbb{E}[F(B_{t_1}, \dots, B_{t_p})]$$

(since our indicators form a partition of $[0, \infty)$). But now this is actually independent of n, so we can remove this limit.

And so then we're basically done, at least in the case $T < \infty$ almost surely — this is just equal to $\mathbb{P}[A]$.

More generally, we didn't really use the assumption $T < \infty$ anywhere except on this last step. So we obtain more generally that

$$\text{LHS} = \mathbb{P}[A \cap \{T < \infty\}] \mathbb{E}[F(B_{t_1}, \dots, B_{t_p})].$$

And this should suffice. Because under the conditioned measure, the expectation has to be divided by $\mathbb{P}[T < \infty]$; so we'd have to divide both sides by that probability. And then

$$\frac{\text{LHS}}{\mathbb{P}[T<\infty]} = \frac{\mathbb{P}[A \cap \{T<\infty\}]}{\mathbb{P}[T<\infty]}\mathbb{E}[F(B_{t_1},\ldots)] = \mathbb{P}[A \mid T<\infty]\mathbb{E}[\cdots],$$

and if you unwrap what the statement is saying, it's basically just this identity. □

Student Question. *Is the event $\{T<\infty\}$ $\mathcal{F}_T$-measurable?*

Answer. Yes. Intuitively it has to be because you should know, up to your random time, whether it's finite or not. By definition you have to check $\{T<\infty\} \cap \{T \le t\} \in \mathcal{F}_T$ for all t; but this is just the event $\{T \le t\}$, and that's the definition of a stopping time.

More generally, we stated last time that T itself is $\mathcal{F}_T$-measurable, so any event like this is also $\mathcal{F}_T$-measurable. And in the proof of this, you should be able to do something similar.

§4.2 The reflection principle

Now let's discuss one application of the strong Markov property, the *reflection principle*.

Theorem 4.3 (Reflection principle)

Let $S_t = \sup_{0\le s\le t} B_s$. Then for $a \ge 0$ and $b \le a$, we have

$$\mathbb{P}[S_t \ge a,\ B_t \le b] = \mathbb{P}[B_t \ge 2a - b].$$

Corollary 4.4

The distribution of S_t is the same as that of $|B_t|$.

(It's not obvious how this follows from the previous statement, but we'll show it.)

The picture is you start looking at your Brownian motion, and you draw a line at level a. If your Brownian motion ever gets to level a, and it ends at some level b (which could be positive or negative, as long as it's at most a), what you want to do is *reflect* everything past the first time you hit a. If you have any path starting from a that gets you down to b, then when you reflect it, you'll end up at $a + (a - b) = 2a - b$. So that's the source of this identity.

You also see why b has to be at most a here — you have to have $S_t \ge B_t$, so if $b > a$, then you wouldn't have to consider that.

Remark 4.5. That's why this is called the reflection principle — you're reflecting your path at a time. Something like this is also true of discrete random walks — you can come up with some bijection. And there's also a general result that random walks converge to Brownian motion, so this is the continuum version of the discrete reflection principle. But of course you can't really prove a bijection for continuous walks; that's not the way to prove this.

Now let's actually prove this.

Proof. Let T_a be the first time the Brownian motion gets to a, i.e., $T_a = \inf\{t \mid B_t = a\}$. Note that $\{S_t \ge a\} = \{T_a \le t\}$ (these are two ways of saying you've reached level a by time t).

Now we can write

$$\mathbb{P}[S_t \geq a, B_t \leq b] = \mathbb{P}[T_a \leq t,\, B^{(T_a)}_{t-T_a} \leq b - a]$$

(note that T_a is a stopping time). Why? The proof for why the second two are the same is because

$$B^{(T_a)}_{t-T_a} = B_t - B_{T_a} = B_t - a$$

by definition. And now you add a to both sides and get back what you want. (But really you should think of it in terms of the picture.)

And now you use the strong Markov property. Now we want to condition on the stopped σ-algebra $\mathcal{F}_{T_a}$; and we can say this is

$$\mathbb{E}[\mathbf{1}(T_a \leq t)\mathbb{P}[B^{(T_a)}_{t-T_a} \leq b - a \mid \mathcal{F}_{T_a}]]$$

(we can pull the indicator out because it is $\mathcal{F}_{T_a}$-measurable). And now we use the reflection symmetry of Brownian motion and the strong Markov property — because $B^{(T_a)}_{t-T_a}$ is a Brownian motion, this is equal to

$$\mathbb{E}[\mathbf{1}(T_a \leq t)\mathbb{P}[-B^{(T_a)}_{t-T_a} \leq b - a \mid \mathcal{F}_{T_a}]].$$

And now we just want to back out of the conditioning. So now this is further equal to

$$\mathbb{P}[T_a \leq t,\, B^{(T_a)}_{t-T_a} \geq a - b]$$

(moving the negative to the other side). But now if you kind of again rewrite what this stopped Brownian motion is (that it's $B_t - B_{T_a} = B_t - a$), this becomes

$$\mathbb{P}[T_a \leq t,\, B_t \geq 2a - b].$$

But then if $B_t \geq 2a - b$, well, $2a - b \geq a$ by assumption, so this extra event $T_a \leq t$ is unnecessary (you know you already hit a higher level than a, so you definitely hit level a); this means this is equal to $\mathbb{P}[B_t \geq 2a - b]$. □

The key step is the reflection step, where we replace $B^{(T_a)}_{t-T_a}$ with its negative. (Le Gall says this slightly differently; when you condition on T_a you can sort of treat T_a as a 'constant,' since T_a is already known.)

Proof of Corollary. Now, why is the second property true? Let's look at the event $S_t \geq a$. We can split

$$\mathbb{P}[S_t \geq a] = \mathbb{P}[S_t \geq a,\, B_t \leq a] + \mathbb{P}[S_t \geq a,\, B_t > a].$$

The second term is just $\mathbb{P}[B_t > a]$, since if you go above a then certainly $S_t \geq a$. For the first, we can apply the reflection principle. So we get that this is

$$\mathbb{P}[B_t \geq a] + \mathbb{P}[B_t > a]$$

(since $2a - a = a$). But we can change $>$ to $\geq$ because B_t has a continuous distribution; so in summary, you get

$$\mathbb{P}[S_t \geq a] = 2\mathbb{P}[B_t \geq a] = \mathbb{P}[|B_t| \geq a]$$

(by the symmetry of normal distributions). And then you're done. □

§4.3 Generalizations of Brownian motion

That almost concludes the beginning module on Brownian motion. We'll just make two brief definitions, which we'll use later.

Definition 4.6. If Z is a (real-valued) random variable, we say (X_t) is a Brownian motion started from Z if $X_t = Z + B_t$ where B is a Brownian motion which is independent of Z.

So by default, whenever we say 'Brownian motion' we assume it starts at 0. But you can also start it at more general points. You can of course start it as constants; but you can also make the starting point a random variable, as long as it's independent of the ensuing increments.

We can also define higher-dimensional BMs.

Definition 4.7. A $\mathbb{R}^d$-valued stochastic process $B_t = (B_t^1, \dots, B_t^d)$ is a *d-dimensional Brownian motion* if its components are IID Brownian motions.

Many of the same properties we showed for Brownian motion also hold here — for example, Blumenthal's 0–1 law, the Markov property, the strong Markov property. If you want, you can go back into the proofs and redo them for this higher-dimensional case; but it's all going to work.

§4.4 Overview — continuous-time stochastic processes

With that, we'll move to more general continuous-time stochastic processes.

The first part of this module is again stuff related to measurability; it's maybe not the most interesting thing in the world, and on a day-to-day basis if you use these things you don't really care about measurability and assume it'll work out. But on a foundational level, it's good to know these concepts. At the end of the module, we'll get on to actual martingale stuff — for instance, the optional stopping theorems, which you can use to prove things about BM. For instance, you can get the distributions of these stopping times T_a using the optional stopping theorem; we'll see that at the end of the module.

§4.5 Filtrations

First, let's set up a basic framework to discuss continuous-time stochastic processes.

Throughout, we fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Definition 4.8 (Filtration). A *filtration* on $(\Omega, \mathcal{F}, \mathbb{P})$ is a collection $(\mathcal{F}_s)_{0 \le s \le \infty}$ of sub-σ-algebras of $\mathcal{F}$, meaning that

$$\mathcal{F}_s \subseteq \mathcal{F}_t \quad \text{for all } 0 \le s \le t \le \infty.$$

This kind of generalizes what we saw for Brownian motion and discrete-time stochastic processes. It's basically an increasing collection of sub-σ-algebras of $\mathcal{F}$. Usually you might assume $\mathcal{F}_\infty$ is the generated σ-algebra from all the finite $\mathcal{F}_t$'s, but in this definition you don't necessarily need to assume that.

Example 4.9

The canonical filtration of Brownian motion is $\mathcal{F}_t = \sigma(B_s \mid s \le t)$ and $\mathcal{F}_\infty = \sigma(B_s \mid s \ge 0)$. More generally, you can replace B by any stochastic process indexed on $[0, \infty)$, and you get a filtration. (That's usually how we get our filtrations.)

Next, we'll introduce a kind of alternative filtration.

Definition 4.10. Given a filtration $(\mathcal{F}_t)$, we define $\mathcal{F}_{t+} = \bigcap_{s>t} \mathcal{F}_s$. (We also define $\mathcal{F}_{\infty+} = \mathcal{F}_\infty$.)

Intuitively, this is all the information known in your process infinitesimally after time t. (We already saw $\mathcal{F}_{0+}$ in the statement of Blumenthal's 0–1 law.)

This is also a filtration; you have $\mathcal{F}_t \subseteq \mathcal{F}_{t+}$ for any fixed t (because $\mathcal{F}_t \subseteq \mathcal{F}_s$ for any $s > t$).

Definition 4.11. We say that a filtration $(\mathcal{F}_t)$ is right-continuous if $\mathcal{F}_t = \mathcal{F}_{t+}$ for all t.

So this is some sort of regularity property for your filtration, which is sometimes convenient to have.

Student Question. *When would this not be true?*

Answer. In the canonical filtration for Brownian motion — $\mathcal{F}_0$ is the trivial σ-algebra (because B_0 is a constant). But $\mathcal{F}_{0+}$ for instance contains the event

$$\{\sup_{0 \le s \le \varepsilon} B_s > 0 \text{ for all } \varepsilon > 0\}.$$

This is in $\mathcal{F}_{0+}$ but not $\mathcal{F}_0$.

But we did show that $\mathcal{F}_{0+}$ is trivial, so it's not *too* far off.

Also, by construction $\mathcal{F}_{t+}$ itself is right-continuous — you can write

$$\bigcap_{s>t} \mathcal{F}_{s+} = \bigcap_{s>t} \bigcap_{r>s} \mathcal{F}_r = \bigcap_{r>t} \mathcal{F}_r = \mathcal{F}_{t+}.$$

§4.6 Completions

Now let $(\mathcal{F}_t)$ be a filtration. Let $\mathcal{N}$ be the collection of $(\mathcal{F}_\infty, \mathbb{P})$-null sets, i.e.,

$$\mathcal{N} = \{E \subseteq \Omega \mid \text{exists } F \in \mathcal{F}_\infty \text{ with } \mathbb{P}[F] = 0 \text{ and } E \subseteq F\}.$$

So basically this means you're contained in an almost not sure event. (We also call these *negligible* sets.)

Definition 4.12. A filtration $(\mathcal{F}_t)$ is complete if $\mathcal{N} \subseteq \mathcal{F}_0$.

This notion of completeness is convenient to have in some settings that we'll see, probably later on. Because it's convenient, you can always guarantee yourself that you're working with a complete filtration: If $(\mathcal{F}_t)$ is *not* complete, then we can take the *completed filtration*

$$\mathcal{F}'_t = \sigma(\mathcal{F}_t, \mathcal{N}).$$

This itself will again be a filtration. And we implicitly also extend $\mathbb{P}$ to $(\Omega, \mathcal{F}'_\infty, \mathbb{P}')$. This is what Problem 6 on the homework is about — you can extend your probability measure onto a completed one, and on this completed probability space you can consider this completed filtration. Le Gall maybe doesn't make this super explicit, that you have to replace your original probability measure by the completed one; this is maybe not a big deal, but it is something you should do.

As a comment, if you do this for Brownian motion, you can check that the strong Markov property still holds with respect to the completed filtration — the restarted process (B^t_s) will be independent of $(\mathcal{F}^B_t)'$. (The usual Markov property is when we don't take the 'prime' indicating completion. You actually have to prove something because in principle this is a bigger filtration — you added these null sets. But the point is you only basically added probability 0 or 1 sets. And independence is a statement about 'the expectation of a product is the product of expectations'; that isn't affected by introducing probability 0 or 1 sets.)

And we have to check this, but this should also imply the completed filtration will contain $\mathcal{F}^B_{t+}$. So you should have this stronger statement of the Markov property — that $(B^t_s)_{s\geq 0}$ is actually independent of $(\mathcal{F}^B_{t+})$, everything you know infinitesimally after time t.

(Sky thinks the completion of $\mathcal{F}_t$ should be the completion of $\mathcal{F}_{t+}$, but this would have to be proven.)

§4.7 Stochastic processes at random times

What we're going to work towards next is that we want to consider stochastic processes ata *random* time T — so we want to look at X_T (T will always be a stopping time; otherwise the theory doesn't work). This is something you do all the time in discrete-time martingale theory — if you want to talk about the optional stopping theorem, you have to talk about your process at a random time.

In discrete time it's trivial that this is measurable — you could write $X_T = \sum_n X_n \mathbf{1}[T = n]$, and each individual thing is a random variable, so this whole thing is too (and you can also show it's in $\mathcal{F}_T$ by the same proof).

But continuous time is different; you can't do exactly this thing. But as we saw with Brownian motion, you basically can by discrete approximation.

So the approach to talking about measurability of X_T is morally reducing to discrete-time statements by discrete approximation; but you always have to do something extra in continuous time.

So we want to build up some results saying why X_T would first be measurable at all, and then why it's $\mathcal{F}_T$-measurable.

§4.8 Measurable, adapted, and progressive stochastic processes

Definition 4.13. A stochastic process (X_t) taking values in a measurable space $(E, \mathcal{E})$ is measurable if the mapping $(\Omega \times \mathbb{R}_+, \mathcal{F} \otimes \mathcal{B}(\mathbb{R}_+)) \to (E, \mathcal{E})$ given by $(\omega, t) \mapsto X_t(\omega)$ is measurable.

(You can take $(E, \mathcal{E})$ to be $\mathbb{R}$ and its Borel sets, but we can say something quite general.)

By definition, a stochastic process is a collection of random variables. But we're now imposing a more stringent condition that if you put this collection together into a map from our product space $\Omega \times \mathbb{R}_+$ to our measurable space, this map should be measurable (with respect to the product σ-algebra on the left and $\mathcal{E}$ on the right).

In this definition, we kind of assumed X itself is a stochastic process, so each X_t is measurable. But this condition is stronger than that. In particular, if you satisfy this condition, you're a stochastic process. Why? We can fix some t_0; and we want to say that the map $\omega \mapsto X_{t_0}(\omega)$ is measurable (as a map $(\Omega, \mathcal{F}) \to (E, \mathcal{E})$) — that's the definition of being a random variable. But you can exhibit this as a composition of two measurable maps

$$(\Omega, \mathcal{F}) \to (\Omega \times \mathbb{R}_+, \mathcal{F} \otimes \mathcal{B}(\mathbb{R}_+)) \to (E, \mathcal{E}),$$

given by $\omega \mapsto (\omega, t_0)$ and $(\omega, t) \mapsto X_t(\omega)$. If you compose these two maps, you get precisely the one we want. We're assuming the second is measurable. And the first is also measurable — by definition of the product σ-algebra, it suffices to just verify measurability on product sets $\{(\omega, t_0) \in F \times B\}$ (where B is a Borel set). But if you unwind the definition, this is either F if $t_0 \in B$, or $\emptyset$ if $t_0 \notin B$; and each of these is in $\mathcal{F}$.

So all that is to say that being measurable is stronger than being a stochastic process, and that's how you verify this statement.

Now in the remaining discussion, implicitly we're always fixing a filtration $(\mathcal{F}_t)$; all the ensuing definitions are with respect to this filtration that we've fixed.

Definition 4.14. A stochastic process (X_t) taking values in $(E, \mathcal{E})$ is called *adapted* if for all t, $X_t \in \mathcal{F}_t$ (i.e., X_t is $\mathcal{F}_t$-measurable).

For instance, Brownian motion was adapted to its canonical filtration.

Definition 4.15. A stochastic process (X_t) is called *progressive* if for all t, the map

$$(\Omega \times [0,t], \mathcal{F}_t \otimes \mathcal{B}([0,t])) \to (E, \mathcal{E})$$

given by $(\omega, t) \mapsto X_t(\omega)$ is measurable.

The only difference from before (with the definition of measurability) is that instead of $\mathcal{F}$, you took $\mathcal{F}_t$; so this is a more stringent condition.

Le Gall doesn't make it clear when he introduces this, but this is precisely the concept you need to eventually verify the measurability of X_T. We'll see later that if you assume T is a stopping time and X is progressive, then X_T is going to be a random variable measurable with respect to $\mathcal{F}_T$. So this is precisely the concept you need; and the proof of measurability is basically composing two maps together. So that's why we have this definition — because we're working towards talking about X_T as a random variable.

(But we'll get to this later, following the order in Le Gall.)

Remark 4.16. A progressive process is both adapted and measurable. It's definitely measurable, because $\mathcal{F}_t$ is a smaller σ-algebra than $\mathcal{F}$; so if it's measurable with respect to this smaller σ-algebra, then it's also measurable with respect to the original one.

For why it's adapted, you use the same argument as above, but now replacing $\mathcal{F}$ with $\mathcal{F}_t$. We want to say that $\omega \mapsto X_{t_0}(\omega)$ is $\mathcal{F}_{t_0}$-measurable. And this is the composition of the two maps $(\omega, t) \mapsto X_t(\omega)$ and $\omega \mapsto (\omega, t_0)$. The first is measurable because of progressiveness. And the second is basically the same proof — if we consider $\{(\omega, t_0) \in F \times B\}$, we either get F or $\emptyset$.

Often we start with an adapted process. And we said the concept of progressiveness is important. So how do we verify that an adapted process is progressive? That's the purpose of the next proposition. This essentially says that 'adapted' plus 'right-continuous' implies 'progressive.'

Proposition 4.17

Let (X_t) be a stochastic process valued in a metric space (E, d) (with the Borel σ-algebra). Suppose that X is adapted and has right-continuous sample paths. Then X is progressive.

You probably need a metric space here because we're talking about continuity.

Remark 4.18. You could also assume X has left-continuous sample paths; either would work.

Most of our processes in this course are actually going to be continuous; we're just making the slightly more general statement because the same proof works. But when we talk about local martingales, we're generally going to assume continuity, because that already suffices to discuss a lot of examples.

So that's the statement, and it's going to allow us to verify progressivity. The proof at a very high level is that you want to use dicrete approximation again; that's what you do over and over in this chapter.

§5 February 18, 2025

§5.1 Progressivity

Student Question. *What does it intuitively mean for a process to be progressive?*

Answer. Sky doesn't know if there's an intuitive notion; but it's the exact notion you need to verify measurability of these stochastic processes at random times. So that's what the main motivation for this definition should be.

At the end of last class and today, we'll be developing some basic language for continuous-time stochastic processes. Last time we defined the notion of a progressive process:

Definition 5.1. A stochastic process (X_t) valued in $(E, \mathcal{E})$ is *progressive* if for all t, the map

$$(\Omega \times [0,t], \mathcal{F}_t \times \mathcal{B}([0,t])) \to (E, \mathcal{E})$$

defined by $(\omega, t) \mapsto X_t(\omega)$ is measurable.

Last time we commented that any progressive process is adapted. The proof is kind of similar to (as we'll see) why when you have a progressive process and evaluate it at a stopping time, that thing is a random variable.

But first, one fundamental proposition is that it's not too hard to verify a process is progressive — it's not too restrictive an assumption. As long as your process is right-continuous and adapted, it's going to be progressive. (Left-continuity would also suffice. All the processes in this course are probably going to be even continuous; but maybe if you'll have to work with continuous-time Poisson processes, which are right-continuous but not continuous, you might want this.)

Proposition 5.2

Let (X_t) be a stochastic process valued in a metric space $(E < d)$. If X is adapted and has right-continuous sample paths, then X is progressive.

Proof. The general approach is to do a discrete approximation — for the discrete approximation it should be relatively easy to verify this property of progressiveness.

So let's fix a time t_0; we're going to verify this property (in the definition of progressiveness) for t_0. We'll define our discrete approximation X^n, which will be a process on $[0, t_0]$, as

$$X_t^n = \sum_{i=1}^{n} \mathbf{1}\left[\frac{i-1}{n} t_0 \le t \le \frac{i}{n} t_0\right] X_{it_0/n} + \mathbf{1}[t = t_0] X_{t_0}.$$

We're basically splitting based on which bucket our time t is in; and then we take the right endpoint of that bucket. Why we choose the right endpoint instead of the left endpoint is informed by our assumption of right-continuity — by right-continuity, for all $t \in [0, t_0]$ and all ω, we have $X_t^n(\omega) \to X_t(\omega)$. This is really because we're taking the right endpoint — so you have a sequence of times converging to t from the *right*.

Then because the limit of measurable functions is measurable, it suffices to show that each of these discrete approximations is measurable, in the sense that when we think of it as a function $(\omega, t) \mapsto X_t^n(\omega)$, this is measurable with respect to $\mathcal{F}_t \times \mathcal{B}([0,t])$.

What does it mean to be measurable? If we fix $A \in \mathcal{B}(E)$ (we only needed E to be a metric space to talk about continuity; for the rest of this E could have been a general measurable space) and then look at

$\{(\omega, t) \mid X_t^n(\omega) \in A\}$, we can basically write this as a union of product sets, directly from the definition of our discrete approximation — it's

$$\bigcup_{i=1}^{n} \{X_{it_0/n} \in A\} \times \left[\frac{i-1}{n}t_0, \frac{i}{n}t_0\right) \bigcup \{X_{t_0} \in A\} \times \{t_0\}$$

(we're just splitting based on which bucket your time t lies in). So you can just explicitly write out what this event is. But then you're done — because by assumption $\{X_{it_0/n} \in A\} \in \mathcal{F}_{t_0}$ and $\{X_{t_0} \in A\} \in \mathcal{F}_{t_0}$ (that's where we use the assumption that X is adapted). And these intervals are definitely measurable subsets of $[0, t_0]$. So our σ-algebra $\mathcal{F}_t \times \mathcal{B}([0, t_0])$ includes all these sets, and we're done. □

Here's a definition, though this doesn't really get used until later on.

Definition 5.3 (Progressive σ-algebra)**.** Let $\mathcal{P}$ be the collection of sets $A \in \mathcal{F} \otimes \mathcal{B}(\mathbb{R}_+)$ such that the process $X_t(\omega) = \mathbf{1}[(\omega, t) \in A]$ is a progressive process. We call $\mathcal{P}$ the *progressive σ-algebra*.

So we're looking at all sets A such that you get a progressive process when you get an indicator. Here are some facts about $\mathcal{P}$ (which we may use later).

Exercise 5.4. This collection $\mathcal{P}$ is a σ-algebra.

(This is either on the current homework or will be on the next one.)

Exercise 5.5. A set $A \subseteq \Omega \times \mathbb{R}_+$ is in $\mathcal{P}$ if and only if for all $t \geq 0$, we have

$$A \cap (\Omega \times [0, t]) \in \mathcal{F}_t \otimes \mathcal{B}([0, t]).$$

This is very similar to the definition of progressive processes; maybe that similarity is why you might think to check this condition.

Exercise 5.6. A process X is progressive if and only if the map $(\Omega \times \mathbb{R}_+, \mathcal{P}) \to (E, \mathcal{E})$ defined by $(\omega, t) \mapsto X_t(\omega)$ is measurable.

This is basically saying that the progressive σ-algebra is the smallest σ-algebra for which all progressive processes are measurable; that's one way to think about it.

§5.2 Stopping times

Now let's define stopping times in slightly more generality than for Brownian motion (it's basically the same, you just have a general filtration).

Definition 5.7 (Stopping times)**.** A random variable $T : \Omega \to [0, \infty]$ is a *stopping time* with respect to $(\mathcal{F}_t)$ if $\{T \leq t\} \in \mathcal{F}_t$ for all $t \geq 0$.

Definition 5.8. We define the *stopped σ-algebra*

$$\mathcal{F}_T = \{A \in \mathcal{F}_\infty \mid A \cap \{T \leq t\} \in \mathcal{F}_t \text{ for all } t \geq 0\}.$$

Note that T is allowed to be infinite. (This is the exact same definition as in Brownian motion.)

Exercise 5.9. The collection $\mathcal{F}_T$ is indeed a σ-algebra.

(This is on the homework.)

Some fundamental facts:

Fact 5.10 — If T is a stopping time, then $\{T < t\}$ is also $\mathcal{F}_t$-measurable.

Proof. We can write $\{T < t\} = \bigcup_{q \in \mathbb{Q}, q<t} \{T \leq q\} \in \mathcal{F}_t$. □

Fact 5.11 — The event $\{T = \infty\}$ is $\mathcal{F}_\infty$-measurable.

Proof. We can write $\{T = \infty\} = (\bigcup_{n \in \mathbb{N}} \{T \leq n\})^c$. □

Since $\mathcal{F}_t \subseteq \mathcal{F}_{t+}$ (recall that this is all your information infinitesimally after time t), a stopping time with respect to $\mathcal{F}_t$ is also one with respect to $\mathcal{F}_{t+}$.

The next proposition characterizes when you're a stopping time with respect to this larger filtration $(\mathcal{F}_{t+})$. (It's easier to be a stopping time with respect to this filtration, because it's larger.)

Proposition 5.12

Let $\mathcal{G}_t = \mathcal{F}_{t+}$.

(i) A random variable T is a stopping time with respect to $(\mathcal{G}_t)$ if and only if $\{T < t\} \in \mathcal{F}_t$ for all $t > 0$. This is equivalent to the statement that $T \wedge t \in \mathcal{F}_t$ for all $t > 0$.

(ii) Let T be a stopping time with respect to $(\mathcal{G}_t)$. Then the stopped σ-algebra $\mathcal{G}_T$ can be written as

$$\mathcal{G}_T = \{A \in \mathcal{F}_\infty \mid A \cap \{T < t\} \in \mathcal{F}_t \text{ for all } t > 0\}.$$

(Here $\mathcal{G}_T$ is defined in the same way as $\mathcal{F}_T$ from before, with $\mathcal{G}$ in place of $\mathcal{F}$. It's natural based on (i) that this stopped σ-algebra based on $\mathcal{G}$ should have this alternative interpretation.)

Notation 5.13. We write $\mathcal{F}_{T+} = \mathcal{G}_T$.

(This is just notation — *a priori* $\mathcal{F}_{T+}$ doesn't have meaning — but it's natural based on these results.)

Proof of (i). Let's first assume T is a stopping time with respect to $\mathcal{G}$; and we want to show the event $\{T < t\}$ is $\mathcal{F}_t$-measurable. This is kind of the same trick as what we did before — we can write

$$\{T < t\} = \bigcup_{\substack{q \in \mathbb{Q} \\ q<t}} \{T \leq q\}.$$

And for each q, by definition $\{T \leq q\} \in \mathcal{G}_q \subseteq \mathcal{F}_t$ (since $q < t$, so all your information infinitesimally after time q is contained in all your information up to time t). So this whole thing is in $\mathcal{F}_t$, which proves one direction.

Conversely, suppose we have this property that $\{T < t\} \in \mathcal{F}_t$ for all t, and we want to verify that T is a stopping time with respect to $(\mathcal{G}_t)$. Then we have to consider $\{T \leq t\}$; and we can write this as

$$\{T \leq t\} = \bigcap_n \left\{T < t + \frac{1}{n}\right\}.$$

We want to say the right-hand side is in $\mathcal{G}_t$. But because this is a decreasing intersection, you can start the intersection at any finite positive integer n_0 (insteaad of at 1); so this is going to be in $\mathcal{F}_{t+1/n_0}$ for all n_0. So that means it's also in $\bigcap_{n_0} \mathcal{F}_{t+1/n_0}$. And by definition that's $\mathcal{F}_{t+} = \mathcal{G}_t$.

(Intuitively you're just saying you have an intersection of decreasing events, so you don't have to look at the first finitely many.)

That proves the first part of (i); now we want to say this is all equivalent to saying that $T \wedge t$ is $\mathcal{F}_t$-measurable as a random variable.

First, if $T \wedge t \in \mathcal{F}_t$, then by definition, for all s we have

$$\{T \wedge t \leq s\} \in \mathcal{F}_t$$

(that's a consequence of measurability). Now if we take $s < t$, then

$$\{T \leq s\} = \{T \wedge t \leq s\}$$

(because there's no way $t \leq s$). So taking a sequence $s_n \nearrow t$, we get that

$$\{T < t\} = \bigcup_n \{T \leq s_n\} \in \mathcal{F}_t.$$

(It's all similar ideas to what we've seen before; here you're taking an increasing union.)

Conversely, now suppose that T is a stopping time with respect to $(\mathcal{G}_t)$; and we want to verify $T \wedge t$ is $\mathcal{F}_t$-measurable. So we want to take events of the form $\{T \wedge t \leq s\}$ and show that they belong to $\mathcal{F}_t$ (if we can verify this for all s, then that shows $T \wedge t$ is $\mathcal{F}_t$-measurable).

First, if $s < t$, then this is just $\{T \leq s\}$, which we saw is in $\mathcal{F}_t$ (we assumed T is a stopping time with respect to $(\mathcal{G}_t)$, so $\{T \leq s\} \in \mathcal{G}_s \subseteq \mathcal{F}_t$); and if $s \geq t$ then it's just Ω (since we're taking a min), which is definitely in $\mathcal{F}_t$. □

Student Question. *What's the intuition behind this proposition?*

Answer. What it means to be a stopping time means you can decide when to stop based on the information you've seen. Now if T is a stopping time with respect to $\mathcal{G}$, that sort of means you can decide whether to stop using the information infinitesimally afterwards. This is a bit weaker because you have extra information, but only infinitesimally after t.

For these precise statements, Sky isn't sure if there's intuition; it's kind of just playing with set theory. These are all facts that we may need later on, so it's good to at least have seen it once before.

That proves (i); we're going to skip the proof of (ii), because it's basically doing something like

$$\{T < t\} = \bigcup_{\substack{q \in \mathbb{Q} \\ q < t}} \{T \leq q\}$$

again (you just have extra intersections with A everywhere).

Now we'll present a long proposition with some more foundational properties of stopping times. It has parts (a) through (j); we are only going to prove (j), and parts (a)–(i) are on the homework. This is proven in the book, so you can look at the book if you want, but if you actually want to understand things or check you understand things, it's good to try to prove them on your own first — it's a good exercise to know that you can manipulate all these definitions.

Proposition 5.14

(a) For a stopping time T, $\mathcal{F}_T \subseteq \mathcal{F}_{T+}$. Furthermore, if $(\mathcal{F}_t)$ is right-continuous (i.e., $\mathcal{F}_t = \mathcal{G}_t$ for all t), then $\mathcal{F}_T = \mathcal{F}_{T+}$.

(b) If $T = t$, then $\mathcal{F}_T = \mathcal{F}_t$ and $\mathcal{F}_{T+} = \mathcal{F}_{t+}$.

(c) T is $\mathcal{F}_T$-measurable.

(d) Let $A \in \mathcal{F}_\infty$, and set

$$T^A(\omega) = \begin{cases} T(\omega) & \omega \in A \\ \infty & \omega \notin A. \end{cases}$$

Then $A \in \mathcal{F}_T$ if and only if T^A is a stopping time.

(e) Let S and T be stopping times with $S \le T$. Then $\mathcal{F}_S \subseteq \mathcal{F}_T$ and $\mathcal{F}_{S+} \subseteq \mathcal{F}_{T+}$.

(f) $S \vee T$ and $S \wedge T$ are also stopping times. Furthermore,

$$\mathcal{F}_{S\wedge T} = \mathcal{F}_S \cap \mathcal{F}_T.$$

Furthermore, $\{S \le T\} \in \mathcal{F}_{S\wedge T}$ and $\{S = T\} \in \mathcal{F}_{S\wedge T}$.

(g) If (S_n) is a monotone increasing sequence of stopping times (written $S_n \uparrow$), then $\lim_n S_n$ is also a stopping time.

(h) If (S_n) is a monotone decreasing sequence of stopping times (written $S_n \downarrow$), then $\lim_n S_n$ is a stopping time with respect to $(\mathcal{F}_{t+})$. If we call $\lim_n S_n = S$, we also have

$$\mathcal{F}_{S+} = \bigcap_n \mathcal{F}_{S_n+}.$$

(i) If $S_n \downarrow$ and S_n is eventually constant (i.e., for all ω, the sequence $S_n(\omega)$ eventually stays at some fixed value), then $S = \lim_n S_n$ is a stopping time (with respect to the original filtration), and $\mathcal{F}_S = \bigcap_n \mathcal{F}_{S_n}$.

(j) Let T be a stopping time. A function $\omega \mapsto Y(\omega)$ defined on $\{T < \infty\}$ taking values in $(E, \mathcal{E})$ is $\mathcal{F}_T$-measurable if and only if for all $t \ge 0$, the restriction of $\omega \mapsto Y(\omega)$ to $\{T \le t\}$ is $\mathcal{F}_t$-measurable.

For (a), intuitively $\mathcal{F}_T$ is all the information up to your random time, and $\mathcal{F}_{T+}$ is all the information up to and infinitesimally after your random time; so it makes sense $\mathcal{F}_T \subseteq \mathcal{F}_{T+}$. To prove it you have to play with set theory. The second part is also intuitive — right continuity means this is true for any *deterministic* time, and this says it's also true for your random time.

For (b), this is also natural if you think of $\mathcal{F}_T$ as all the information up to your random time (which is no longer random).

For (d), we're assuming T is a stopping time, and we modify the stopping time by setting it to T on A and ∞ off.

For (e), this is again quite natural — all the information up to your first random time has to be contained in all the information up to your second, larger, random time.

For (f), $\mathcal{F}_{S\wedge T}$ is all the information up to the *first* (smaller) one of your random times. Why should $\{S \le T\}$ be contained in this? Imagine you're observing your process and deciding when to stop. If I see that I've stopped S and I haven't stopped T, that should be known based on all the information up to time S, which in this case is $S \wedge T$ (since I'm assuming S stopped before T); so if you think about something like that, you kind of guess that these should be true.

For (h) vs. (i), (i) says that if you have the additional property of being eventually constant, then you don't

have to look infinitesimally ahead to know when to stop. This is intuitive because if you know eventually $S_n = S$, then all the information up to time S_n should be all the information up to time S (at least, when you take this intersection).

Property (j) is the one thing we'll actually prove in class, because we'll eventually come back to why when you have a progressive process and take a random time, that gives you a random variable; and to prove that, we need property (j).

For (j), what we really mean is that Y is a function on $\{T < \infty\} \subseteq \Omega$. We're saying that to verify Y is measurable with respect to $\mathcal{F}_T$, you can kind of reduce to verifying measurability for fixed times. To be precise, when we say the restriction of a map is itself measurable, what you're saying is that we have a map Y defined on the set $\{T < \infty\}$ with the σ-algebra $\{A \cap \{T < \infty\} \mid A \in \mathcal{F}_T\}$; and we're saying that this map to $(E, \mathcal{E})$ is measurable. Then the restriction of Y to $\{T \leq t\}$ is a map

$$Y_t : (\{T \leq t\}, \{A \cap \{T \leq t\} \mid A \in \mathcal{F}_t\}) \to (E, \mathcal{E})$$

(basically we just replaced $\{T < \infty\}$ with the smaller event $\{T \leq t\}$). And we're saying the first thing is measurable if and only if the second is measurable for all t.

Proof. First let's suppose Y is $\mathcal{F}_T$-measurable. Now we want to verify that Y_t is measurable for any t. So let's take $B \in \mathcal{E}$; we want to look at an event of the form $\{Y_t \in B\}$, and we want to show this is contained in the relevant σ-algebra.

But what is this event? We can think of it as $\{Y \in B\} \cap \{T \leq t\}$ (because all we did was restrict to a smaller space, and we can write this in terms of the original map and then add in the restriction). And we're assuming Y is measurable, so we can write $\{Y \in B\} = A \cap \{T < \infty\}$ for some $A \in \mathcal{F}_T$. And because $\{T \leq t\} \subseteq \{T < \infty\}$, this is just $A \cap \{T \leq t\}$. And if $A \in \mathcal{F}_T$, then by definition this is in $\mathcal{F}_t$; so that verifies what we wanted (e.g., we can write it as $(A \cap \{T \leq t\}) \cap \{T \leq t\}$; the first event A' is in $\mathcal{F}_t$ by the fact that $A \in \mathcal{F}_T$).

That proves one direction. Conversely, now we're going to assume Y_t is $\mathcal{F}_t$-measurable for all t; and we want to show that $\{Y \in B\}$ is in the relevant σ-algebra $\{A \cap \{T < \infty\} \mid A \in \mathcal{F}_T\}$. By the definition of what $\mathcal{F}_T$ is, we need to intersect this with $\{T \leq t\}$ and show that this is in $\mathcal{F}_t$. But this is just

$$\{Y \in B\} \cap \{T \leq t\} = \{Y_t \in B\} \in \mathcal{F}_t.$$

So this implies that $\{Y \in B\} \in \mathcal{F}_T$. And $\{Y \in B\} = \{Y \in B\} \cap \{T < \infty\}$, because Y itself is just a function on $\{T < \infty\}$. □

Remark 5.15. There's probably some redundancy in how we defined things — we shouldn't have to add in these intersections that are basically the whole space — but it doesn't really matter. The reason we had them is that we have functions that are defined on just a subset of Ω rather than all of Ω, so you have to talk about a σ-algebra on this subset; and the natural one is just that you take any subset in the original σ-algebra and intersect it with this subset.

The abstract thing is that if we have measurable spaces $(E_1, \mathcal{E}_1) \to (E_2, \mathcal{E}_2)$ and a measurable function f, now if I take some $F_1 \in \mathcal{E}_1$, I want to define a new function $f|_{F_1}$ on F_1; and if I want to talk about its measurability properties, I have to define a new σ-algebra; and the natural one is $\{F_1 \cap F \mid F \in \mathcal{E}_1\}$. (Here F_1 was the event $\{T \leq t\}$.)

But what happens is that if I take an event $A \cap \{T < \infty\}$ and then I'm intersecting it with $F_1 = \{T \leq t\}$, then $\{T < \infty\}$ is redundant and I can write this as $A \cap \{T \leq t\} \in \mathcal{F}_t$. So maybe that's why this extra thing was redundant.

§5.3 Progressive processes at stopping times

Now let's get back to what we said at the beginning about progressive processes — (j) is kind of exactly the thing you need, and we'll see why.

Theorem 5.16

Let (X_t) be a progressive process valued in some measurable space $(E, \mathcal{E})$. Let T be a stopping time. Then the map $\omega \mapsto X_{T(\omega)}(\omega)$ is $\mathcal{F}_T$-measurable on the event $\{T < \infty\}$.

(We fix a filtration beforehand, which gives a notion of progressive processes and stopping times.)

In hindsight, the reason we did all this effort in (j) to talk about measurability on $\{T < \infty\}$ is that certainly stopping times can be ∞, but for a general stochastic process there's no notion of X_∞ (for example, B_∞ is undefined, because you'll oscillate infinitely often). So in general you can only talk about X_T when T is finite. (We'll see that with martingales satisfying certain properties you *do* have an almost sure limit as $T \to \infty$, and in that case you can define this even without restricting to the case where T is finite.)

Proof. We'll just apply (j); by (j), it suffices to verify that for all $t \geq 0$, the map $\omega \mapsto X_{T(\omega)}(\omega)$ is $\mathcal{F}_t$-measurable on the event $\{T \leq t\}$ (that's what (j) is telling us — that we can just reduce to looking at these deterministic times).

One reduction is that on the event $\{T \leq t\}$, we can write

$$X_{T(\omega)}(\omega) = X_{T(\omega)\wedge t}(\omega).$$

This means it suffices to show that $X_{T(\omega)\wedge t}(\omega)$ is $\mathcal{F}_t$-measurable on $\{T \leq t\}$.

And basically, the reason this is measurable is that we're composing two maps: the first is the map

$$(\{T \leq t\}, \mathcal{F}_t) \to (\Omega \times [0,t], \mathcal{F}_t \times \mathcal{B}([0,t]))$$

defined by $\omega \mapsto (\omega, T(\omega) \wedge t)$. And the second is just the thing where you evaluate your process on this pair — it's the map

$$(\Omega \times [0,t], \mathcal{F}_t \times \mathcal{B}([0,t])) \to (E, \mathcal{E})$$

defined by $(\omega, s) \mapsto X_s(\omega)$.

One checks that our map is a composition of these two. And by assumption the second one is measurable, because the process is progressive (and this is precisely the definition of a progressive process). So now it suffices to check that just the first one is measurable (then the composition of measurable maps is measurable, so this would allow us to conclude what we want).

To check this, we take a generating event in the σ-algebra $\mathcal{F}_t \times \mathcal{B}([0,t])$ — so we can take a product event $A \times B$ where $A \in \mathcal{F}_t$ and $B \in \mathcal{B}([0,t])$. And now we want to look at

$$\{\omega \mid (\omega, T(\omega) \wedge t) \in A \times B\} \cap \{T \leq t\}.$$

(This is the pre-image of $A \times B$ under our map, and we want to verify this is in $\mathcal{F}_t$.) Certainly $\{T \leq t\} \in \mathcal{F}_t$, so that's fine. Meanwhile, for the first term, we can write this as

$$(A \cap \{T \wedge t \in B\}) \cap \{T \leq t\}.$$

And $A \in \mathcal{F}_t$ and $\{T \leq t\} \in \mathcal{F}_t$. Meanwhile, for the second term, there's several things you could do; for example, the random variable $T \wedge t$ is $\mathcal{F}_t$-measurable (T is a stopping time with respect to the filtration, so it's a stopping time with respect to $(\mathcal{G}_t)$; and we had a proposition saying that being a stopping time with respect to this larger filtration is equivalent to saying $T \wedge t$ is $\mathcal{F}_t$-measurable). So all these sets are in $\mathcal{F}_t$ and we're done.

(We could also use (f) — t itself is a stopping time, so $T \wedge t \in \mathcal{F}_{T\wedge t} \subseteq \mathcal{F}_t$.) □

§5.4 Some preliminaries

Now we know what progressive processes are, and we can talk rigorously about this random time — this is what you really want to do in martingale theory (talk about your stopped processes), and now we have the foundations to do that (we know these stopped processes define random variables).

Before we get to martingale theory, there are one or two preliminary processes that will be helpful.

Proposition 5.17

Let T be a stopping time, and let $S : \Omega \to [0, \infty]$ be $\mathcal{F}_T$-measurable such that $S \geq T$. Then S is also a stopping time.

In particular, the reason this is useful is that if T is a stopping time, then we can define a sequence of discrete approximations

$$T_n = \sum_{k=1}^{\infty} \mathbf{1}[(k-1)2^{-n} < T \leq k2^{-n}]k2^{-n} + \infty \mathbf{1}[T = \infty].$$

Then this means T_n is a stopping time and $T_n \searrow T$. (By definition $T_n \geq T$ because we took the right endpoint of each interval. And to check T_n is $\mathcal{F}_T$-measurable, you just need to check each of these events in the indicator functions is $\mathcal{F}_T$-measurable, which you can do just by going back to the definitions.)

Proof. We want to show S is a stopping time, so we need to show $\{S \leq t\} \in \mathcal{F}_t$. First, we know $\{S \leq t\} \in \mathcal{F}_T$ because S is $\mathcal{F}_T$-measurable. And we can write

$$\{S \leq t\} = \{S \leq t\} \cap \{T \leq t\}$$

(because $S \geq T$, so the intersection doesn't do anything). And the reason you introduce this is because by definition of what it means to be in $\mathcal{F}_T$, this is in $\mathcal{F}_t$. So S is a stopping time. □

There's just one final proposition before we get into martingale theory. It basically gives you examples of stopping times; and it's the most frequent examples we'll use.

Proposition 5.18

Let (X_t) be a progressive process valued in a metric space (E, d).

(i) If X has right-continuous sample paths and O is an open set of (E, d), then

$$T_O = \inf\{t \mid X_t \in O\}$$

is a stopping time with respect to $(\mathcal{F}_{t+})$.

(ii) If X has continuous sample paths and $F \subseteq E$ is closed, then

$$T_F = \inf\{t \mid X_t \in F\}$$

is a stopping time (with respect to $(\mathcal{F}_t)$).

For (i), we're looking at the first time your process enters this open set O; and this isn't necessarily a stopping time with respect to the original filtration, but it is with respect to the enlarged one. And the second part says that if you have *continuous* sample paths and you look at a *closed* set, then the first time you enter is a stopping time with respect to the original filtration.

We'll prove (i), and leave (ii) for next time.

Proof of (i). In order to be a stopping time with respect to $\mathcal{F}_{t+}$, we can just look at events of the form $\{T_0 < t\}$ and show it's in $\mathcal{F}_t$. By right-continuity, we can write

$$\{T_0 < t\} = \bigcup_{s \in D} \{X_s \in O\}$$

where $D \subseteq [0,t]$ is a countable dense subset of $[0,t]$. (This is using right continuity and the fact that O is open — so if there's some time when your process is in O, then going slightly beyond that time it must still stay in O.) And each of these things is $\mathcal{F}_s$-measurable because X is progressive (so in particular adapted), so you get this result; that verifies T_O is a stopping time with respect to this enlarged filtration. □

Next time we'll get into continuous-time martingales. It might be good to review the discrete-time theory, because we'll use lots of that as input and not reprove them (e.g., convergence theorems for martingales).

§6 February 19, 2025

Yesterday we finished off stating this proposition:

Proposition 6.1

(ii) Let X be progressive, valued in a metric space (E,d), and continuous. If $F \subseteq E$ is closed, then

$$T_F = \inf\{t \geq 0 \mid X_t \in F\}$$

is a stopping time with respect to $(\mathcal{F}_t)$.

(We didn't rewrite (i), which we proved yesterday; we'll prove (ii) today.) This is probably going to be the most relevant case for us — if you think about the first time Brownian motion hits a point, for example, that's an example of this stopping time. But the statement is that if you have a progressive proccess taking values in some metric space with *continuous* sample paths, then the first time you hit a closed set is a stopping time (with respect to your original filtration).

Proof. First, note that the function on your metric space E defined by

$$x \mapsto d(F,x) = \inf_{y \in F} d(y,x)$$

(basically, the distance from your point to your closed set) is continuous. We'll prove this later, but first let's assume this is true.

Then th process given by $(d(F,X_t))_t$ (taking the distance from your process X to your closed set F) has continuous sample paths; and it's also adapted (just because, well, you're taking this continuous function and applying it to X_t; X_t is adapted because you had a progressive process). And you can express the event $\{T_F \leq t\}$ as

$$\{T_F \leq t\} = \left\{ \inf_{s \in D} d(F, X_s) = 0 \right\}$$

where D is some countable dense subset of $[0,t]$. Here we're using continuity of the distance process to replace the inf over *all* times with just a countable set; and we want a countable set so that we can say this event is a countable intersection, so it's going to be measurable. And so this is in $\mathcal{F}_t$ — any individual one of these events is $\mathcal{F}_s$-measurable because the process is adapted, and $s \leq t$, so this whole thing is in $\mathcal{F}_t$.

So that verifies you have a stopping time, assuming this distance function is continuous; now let's show that. It's actually going to be Lipschitz continuous. Basically the picture is that you have some closed set

F and points x and z; and somehow if we look at the individual distances between x and F and z and F, the difference between them can't be greater than the distance between x and z, basically by the triangle inequality.

So let $x, z \in E$, and let $y_0 \in F$ be such that $d(F, z) = d(y_0, z)$. Here's where we use the fact that F is closed; that means this infimum is actually attained by some point y_0. Then we have

$$d(F, x) - d(F, z) \leq d(y_0, x) - d(y_0, z)$$

(because $d(F, z) = d(y_0, z)$ by assumption, and $d(F, z)$ is an infimum, so we can bound it by the distance to some point). And by the triangle inequality, this is at most $d(x, z)$. (It's more clear if you move the minus to the other side.)

And this is symmetric in x and z, so if we swap their roles we get that

$$|d(F, x) - d(F, z)| \leq d(x, z).$$

So it's actually Lipschitz continuous, not just continuous. And that shows this proposition. □

Student Question. *Where did we use that X was progressive?*

Answer. It seems nowhere. Maybe the fact you're assuming X is continuous means that if you just assume adaptedness, then it's automatically progressive. And we have to assume continuity; so we could have just said it's continuous and adapted; but that automatically means it's progressive.

That concludes some of this foundational material about stopping times and σ-algebras. Now we can move on to continuous-time martingales.

§6.1 Continuous-time martingales

Let's start by defining a martingale. First, in the background we're always fixing some filtration (as usual); all definitions are with respect to that filtration.

Definition 6.2. An adapted $\mathbb{R}$-valued process (X_t) such that $X_t \in L^1$ for all $t \geq 0$ is called:

- A *martingale* if for all $0 \leq s < t$, we have $\mathbb{E}[X_t \mid \mathcal{F}_s] = X_s$.
- A *supermartingale* if for all $0 \leq s < t$, we have $\mathbb{E}[X_t \mid \mathcal{F}_s] \leq X_s$.
- A *submartingale* if for all $0 \leq s < t$, we have $\mathbb{E}[X_t \mid \mathcal{F}_s] \geq X_s$.

This is all exactly the same as in discrete time, except that now your times are allowed to be continuous. Martingales model a process that tends not to increase or decrease, at least in expectation.

One thing to note is that if you have a supermartingale, then by negating it you get a submartingale. So it's often the case that if you prove something for supermartingales, it also extends to submartingales (as long as you keep track of the $-$).

§6.2 Examples of martingales

Often, the main thing you want to do is find martingales, because they satisfy very nice properties and also let you do explicit computations. Somehow the fact that it's a process that doesn't have a trend of increase or decrease kind of tells you something is 'conserved,' and this leads to exact computations of things, as it turns out.

But anyways, let's get into some examples. How do you spot martingales, basically? Here's a very general setup that allows you to find martingales.

Example 6.3

Let Z be adapted and with independent increments, i.e., $Z_t - Z_s$ is independent of $\mathcal{F}_s$ for all $0 \le s < t$. Then:

(i) If we additionally have $Z_t \in L^1$ for all t, then $\widetilde{Z}_t = Z_t - \mathbb{E}[Z_t]$ is a martingale.

(ii) If $Z_t \in L^2$ for all t, then $Y_t = \widetilde{Z}_t^2 - \mathbb{E}[\widetilde{Z}_t^2]$ is a martingale.

(iii) If for some $\theta \in \mathbb{R}$, for all $0 \le s < t$ you have $\mathbb{E}[\exp(\theta Z_t)], \mathbb{E}[\exp(\theta(Z_t - Z_s))] < \infty$, then

$$X_t = \frac{\exp(\theta Z_t)}{\mathbb{E}[\exp(\theta Z_t)]}$$

is a martingale.

So for (i), if your Z_t's are integrable, then we just need to subtract out the mean to get a martingale. For (ii), if the Z_t's are actually in L^2, then we get a 'quadratic' martingale — where you subtract out the mean and also the variance. And as you can see, there's a more general martingale that gives these examples — you can imagine taking cubes and fourth powers and so on, and in principle if you assume enough integrability, you can get corresponding martingales, though the things you subtract become more and more complicated. But it turns out you can encode them all in some 'exponential' martingale, as in (iii). Here you have to *divide* by the mean, instead of subtracting.

Why are (i) and (ii) encoded in (iii)? Basically if (iii) is true for θ in some interval around 0, you basically want to differentiate X_t in θ. If you differentiate once and set $\theta = 0$, the martingale you get should be the one in (i); and if you differentiate more and more times, you should get higher-order martingales. (But we won't go into this.)

Proof. For (i), let's just try to compute the conditional expectation $\mathbb{E}[Z_t \mid \mathcal{F}_s]$ (the thing we want to look at). First we can write

$$\mathbb{E}[Z_t \mid \mathcal{F}_s] = \mathbb{E}[Z_t - Z_s \mid \mathcal{F}_s] + Z_s$$

(by assumption Z_s is $\mathcal{F}_s$-measurable, so you don't need a conditional expectation on it). And now using the independent increments property, we have $\mathbb{E}[Z_t - Z_s \mid \mathcal{F}_s] = \mathbb{E}[Z_t - Z_s]$. And now we can rearrange things to write

$$\mathbb{E}[Z_t \mid \mathcal{F}_s] = \mathbb{E}[Z_t - Z_s] + Z_s = \widetilde{Z}_s + \mathbb{E}[Z_t].$$

And if you move $\mathbb{E}[Z_t]$ to the other side, you get $\widetilde{Z}_t$; that verifies the martingale property for this linear martingale.

That's what you do in general — if you have to guess what you need to subtract from $\widetilde{Z}_t$ to get a martingale, you kind of do this computation and see how to group things in an appropriate way. We'll also see this for the quadratic martingale.

For the quadratic martingale (ii), let's try to compute

$$\mathbb{E}[\widetilde{Z}_t^2 \mid \mathcal{F}_s] = \mathbb{E}[(\widetilde{Z}_t - \widetilde{Z}_s)^2 \mid \mathcal{F}_s] + 2\mathbb{E}[\widetilde{Z}_s(\widetilde{Z}_t - \widetilde{Z}_s) \mid \mathcal{F}_s] + \widetilde{Z}_s^2$$

(again we write $\widetilde{Z}_t = (\widetilde{Z}_t - \widetilde{Z}_s) + \widetilde{Z}_s$; then we have a square, so we get three terms; and we don't need to take a conditional expectation on the last term because $\widetilde{Z}_s$ is adapted).

The first thing to note is that the cross-term $\mathbb{E}[\widetilde{Z}_s(\widetilde{Z}_t - \widetilde{Z}_s) \mid \mathcal{F}_s]$ is just 0 — you can pull out $\widetilde{Z}_s$ because it's adapted, and then use the martingale property to say that the expected increment is 0.

And for the last term, you can remove the conditioning by independence, so we get

$$\mathbb{E}[(\widetilde{Z}_t - \widetilde{Z}_s)^2] + \widetilde{Z}_s^2.$$

Now the first thing you can try with this is to expand out the square again, giving

$$\mathbb{E}[\widetilde{Z}_t^2] - 2\mathbb{E}[\widetilde{Z}_t\widetilde{Z}_s] + \mathbb{E}[\widetilde{Z}_s^2] + \widetilde{Z}_s^2.$$

Again we get this cross-term. But the thing to realize — which is always good to know — is that you can always kind of look at the increments. What this means is that we can write this as

$$\mathbb{E}[\mathbb{E}[\widetilde{Z}_s\widetilde{Z}_t \mid \mathcal{F}_s]] = \mathbb{E}[\widetilde{Z}_s\mathbb{E}[\widetilde{Z}_t \mid \mathcal{F}_s]].$$

And now using the martingale property, this is just $\mathbb{E}[\widetilde{Z}_s^2]$.

And now we can group these two; so in the end this whole thing is

$$\mathbb{E}[\widetilde{Z}_t^2] - \mathbb{E}[\widetilde{Z}_s^2] + \widetilde{Z}_s^2.$$

And now it's in the right form — if you move $\mathbb{E}[\widetilde{Z}_t^2]$ to the left-hand side you verify the martingale property for Y_t. This type of thing where you realize you can do this conditional expectation thing — we're going to be using it over and over again.

For this exponential martingale (iii), let's again look at $\mathbb{E}[\exp(\theta Z_t) \mid \mathcal{F}_s]$. Again you want to use the independent increment property, so you can write this as

$$\exp(\theta Z_s)\mathbb{E}[\exp(\theta(Z_t - Z_s)) \mid \mathcal{F}_s].$$

And again by the independent increment property, we can remove the conditional expectation, so this is

$$\exp(\theta Z_s)\mathbb{E}[\exp(\theta(Z_t - Z_s))].$$

Now what do we want to say? It's kind of a trick — what you want to do is separate this and write it as a product of expectations, or at least this was Sky's first instinct. But that's not possible because Z_t and Z_s aren't independent. But what you *do* know is that if we take the expectation of both sides, we get

$$\mathbb{E}[\exp(\theta Z_t)] = \mathbb{E}[\exp(\theta Z_s)]\mathbb{E}[\exp(\theta(Z_t - Z_s))].$$

That should give us an identity for $\mathbb{E}[\exp(\theta(Z_t - Z_s))]$; we can now replace it with this ratio and move it to the left-hand side. And that concludes — now we get

$$\mathbb{E}[\exp(\theta Z_t) \mid \mathcal{F}_s] = \exp(\theta Z_s) \cdot \frac{\mathbb{E}[\exp(\theta Z_t)]}{\mathbb{E}[\exp(\theta Z_s)]}.$$

And we can move $\mathbb{E}[\exp(\theta Z_t)]$ to the left-hand side, and that verifies the thing we claimed was a martingale is indeed a martingale. □

Remark 6.4. One thing we didn't explicitly check was that these things are all integrable; but that just follows from the assumptions. In (i), if Z is integrable so is $\widetilde{Z}$. In (ii), if $Z \in L^2$ then $\widetilde{Z}^2$ is integrable. And in (iii), you're basically just assuming that everything up here is integrable. In Le Gall, he may have stated it without the second assumption that the exponential moments of the *increments* are all finite. Sky isn't sure whether it's true without that — because in this proof you need to assume $\mathbb{E}[\exp(\theta(Z_t - Z_s))]$ is integrable. But there may be an argument showing that if all $\mathbb{E}[\exp(\theta Z_t)]$'s are finite then these are finite as well.

§6.2.1 Martingales from Brownian motion

What process Z are we going to use the most often? Well, it's Brownian motion. So if we look at what we get when Z is Brownian motion, we get a bunch of martingales associated with Brownian motion; so let's see that.

But before we do that, we need to slightly expand our definition of Brownian motion. When you talk about martingales, you're always fixing a filtration beforehand. So we want to talk about Brownian motion with respect to that filtration.

Definition 6.5. We say a process (B_t) is a *$(\mathcal{F}_t)$-Brownian motion* if B is a Brownian motion, it is adapted, and it has independent increments with respect to $(\mathcal{F}_t)$.

Recall being a Brownian motion means you have continuous sample paths and the right finite-dimensional distributions (you're a Gaussian process with the right covariance function).

Naturally you'd want it to be adapted with respect to your filtration. Of course, if you just take your filtration to be the filtration generated by the Brownian motion, all this is true. But the point is sometimes you can't — sometimes you're given a filtration and can't choose it. And you might still want this independent increment property with respect to that larger filtration.

You can also define the same thing for $\mathbb{R}^d$-valued Brownian motion.

Then applying the earlier example to Brownian motion:

Example 6.6

If (B_t) is a $(\mathcal{F}_t)$-Brownian motion, then the following are all martingales: B_t itself, $B_t^2 - t$, and $\exp(\theta B_t - \theta^2 t/2)$ (for any θ).

Here we're using the fact that B_t has mean 0 and variance t. And for the third, we're using the fact that the normal distribution has good tails, so you always have exponential integrability (no matter what θ is). You can get the extra correction $-\theta^2 t/2$ because this thing had better have mean 1 — it's a martingale, so it has to have the same mean as at time 0, where it's just 1. So whatever you subtract here has to make it so that the whole thing has mean 1. And $\exp(\theta B_t)$ is the moment generating function of Brownian motion; so you just have to remember that that's $\exp(\theta^2 t/2)$, and you just have to cancel that to get something with mean 1.

For what we were saying about differentiating in θ, if we differentiate this exponential in θ and set $\theta = 0$, this should just be B_t — when the derivative hits the θB_t term you get B_t, and for the second term you still have a θ, so setting $\theta = 0$ makes that term go away.

And if you differentiate twice in θ and set $\theta = 0$, now you're going to get $B_t^2 - t$.

So this is what we mean when we say if you differentiate in θ, you generate all these martingales with higher and higher degree polynomials. These polynomials are quite special and are called *Hermite polynomials*; but we probably won't get into that.

Student Question. *Can you give some intuition to why all the derivatives are martingales?*

Answer. The point is that you should think of this as a family of martingales indexed by θ now — very abstractly, if you have

$$\mathbb{E}[M_t(\theta) \mid \mathcal{F}_s] = M_s(\theta)$$

for all θ (so you get a martingale for each θ), then you can differentiate both sides by θ (as many times as you want) and get

$$\partial_\theta^k \mathbb{E}[M_t(\theta) \mid \mathcal{F}_s] = \partial_\theta^k M_s(\theta).$$

And if you formally believe you can take the θ inside the expectation, you get

$$\mathbb{E}[\partial_\theta^k M_t(\theta) \mid \mathcal{F}_s] = \partial_\theta^k M_s(\theta),$$

which says this kth derivative is also a martingale. You can often formally justify taking the derivative inside; but anyways once you have a guess for what the martingale should be, you can try to verify directly that the polynomial you get is a martingale. But that's the intuition — you just differentiate both sides with respect to your parameter.

So these are the basic examples of martingales. There's another example in the notes, which Le Gall talks about, using white noise; but we'll probably just skip it.

§6.3 General facts about martingales

Now let's discuss some general facts about martingales. They're all extensions of discrete time theorems or propositions about martingales; so let's see these.

§6.3.1 Submartingales

Proposition 6.7

Let (X_t) be an adapted process, and let $f : \mathbb{R} \to \mathbb{R}_+$ be a convex function such that $\mathbb{E}[f(X_t)]$ is finite for all t.

(i) If (X_t) is a martingale, then $(f(X_t))$ is a submartingale.

(ii) if (X_t) is a submartingale and f is nondecreasing, then $(f(X_t))$ is a submartingale.

So we have a positive convex function such that $f(X_t)$ is integrable for all t.

Proof. For both cases, you have to know about conditional Jensen's inequality, or the fact that Jensen's inequality still holds for conditional expectations; so you get that

$$\mathbb{E}[f(X_t) \mid \mathcal{F}_s] \geq f(\mathbb{E}[X_t \mid \mathcal{F}_s])$$

(if you have an expectation of a convex function, you can take it outside). And now we want to conclude that this is at least $f(X_s)$.

In the martingale case, it's actualaly exactly equal, since $\mathbb{E}[X_t \mid \mathcal{F}_s] = X_s$. In the second case, if you have a submartingale, then $\mathbb{E}[X_t \mid \mathcal{F}_s] \geq X_s$, and you have to use the fact that f is nondecreasing. But in either case you get this inequality

$$\mathbb{E}[f(X_t) \mid \mathcal{F}_s] \geq f(\mathbb{E}[X_t \mid \mathcal{F}_s]) \geq f(X_s).$$

(So the main thing is conditional Jensen's; but that requires some trick to prove, which we won't go over.) □

Some useful things to know as a consequence of this proposition:

Example 6.8

If you have a martingale (X_t), then if you take pth powers, $|X_t|^p$ is a submartingale for all $p \geq 1$ (assuming that $|X_t|^p$ is integrable).

(This is because the function $x \mapsto |x|^p$ is convex; this is why we needed $p \geq 1$.)

As another consequence:

Notation 6.9. For any random variable X, we define its *positive part* as $X^+ = \max\{X, 0\}$.

The function $x \mapsto x^+$ is a convex function, and it's nondecreasing. So we get the following:

Example 6.10

If (X_t) is a submartingale, then (X_t^+) is also a submartingale.

Here you don't have to assume additional integrability — you're already assuming $\mathbb{E}[|X_t|]$ is finite (that's what it means for X_t to be integrable); but this is $\mathbb{E}[X_t^+ + X_t^-]$, where $X^- = \max\{-X, 0\}$. (This is still a convex function, though it's actually decreasing.) So if you assume this is finite, then $\mathbb{E}[X_t^+]$ is also finite.

Now let's get to another fact about submartingales. The point of all these facts is basically that martingales or submartingales are very special processes. Here's one way they're special.

Proposition 6.11

Let (X_t) be a submartingale. Then for all $t \geq 0$, we have

$$\sup_{0 \leq s \leq t} \mathbb{E}[|X_s|] < \infty.$$

In other words, your submartingale is L^1-bounded on any finite interval. This is one consequence of the submartingale property — it's maybe the first example of why submartingales are special, that you're actually L^1-bounded on any finite interval.

Proof. First of all, since we know (X_t^+) is a submartingale by what we just said, if you use the submartingale property you obtain that

$$\mathbb{E}[X_s^+] \leq \mathbb{E}[X_t^+] \quad \text{for all } 0 \leq s \leq t.$$

And so at least if you replaced the absolute value with just the positive part, then you'd be done (this thing is just bounded by the endpoint $\mathbb{E}[X_t^+]$). But of course you also have to deal with the negative part (because $|X_s| = X_s^+ + X_s^-$); how do we do that?

For that, we basically want to say the expectations cannot get too negative. First, we have

$$\mathbb{E}[X_s^+ - X_s^-] = \mathbb{E}[X_s] \geq \mathbb{E}[X_0]$$

(using the submartingale property) — so we basically said the expectation (not the L^1-norm) cannot get too negative. But now you can turn this into an estimate on the negative part by moving it over — we get

$$\mathbb{E}[X_s^-] \leq \mathbb{E}[X_s^+] - \mathbb{E}[X_0].$$

And we already had a bound for $\mathbb{E}[X_s^+]$. So we get that

$$\sup_{0 \leq s \leq t} \mathbb{E}[|X_s|] \leq 2\mathbb{E}[X_t^+] - \mathbb{E}[X_0].$$

And that finishes. □

§6.3.2 Uncorrelated increments

Now let's talk about L^2 martingales briefly, and the fact that they have nice 'uncorrelated increments.' (When we say (M_t) is a L^2 martingale we mean $M_t \in L^2$ for every fixed t; this doesn't necessarily mean that (M_t) is bounded in L^2.)

Proposition 6.12

Let (M_t) be an L^2 martingale. Let $0 \leq s < t$, and let $s = t_0 < t_1 < \cdots < t_p = t$ be a subdivision of $[s,t]$. Then we have

$$\mathbb{E}\left[\sum_{i=1}^{p}(M_{t_i} - M_{t_{i-1}})^2 \mid \mathcal{F}_s\right] = \mathbb{E}[M_t^2 - M_s^2 \mid \mathcal{F}_s] = \mathbb{E}[(M_t - M_s)^2 \mid \mathcal{F}_s].$$

What is this saying? Well, the first and third things being equal basically mean that even conditionally, the increments of any L^2 martingale are uncorrelated — you can write $M_t - M_s$ as the sum of hte increments $M_{t_i} - M_{t_{i-1}}$. And now we're saying when you look at the second moment, you just need to look at the diagonal terms. Usually when you take the second moment of a sum, you get diagonal terms and cross-terms; and this basically says the cross-terms disappear because you're uncorrelated.

This is quite a special property and somehow encodes a bunch of cancellation. For example, typically if you want to bound the square of a sum, you're not worried about the diagonal terms — if your sum has n terms, you'll only have n diagonal terms in the square, but $O(n^2)$ cross-terms. So if you want to bound the square of a sum, you usually worry about the cross-terms, because they're larger in number. But here you're saying all the cross-terms are 0, which encodes the cancellations. On some level, this is why stochastic integration works — because you have this kind of cancellation or uncorrelated increments — as we'll see later.

Proof. Let's compute one of the squared increments

$$\mathbb{E}[(M_{t_i} - M_{t_{i-1}})^2 \mid \mathcal{F}_s].$$

The first thing you want to do is use the tower property — we want to condition on the left endpoint, so we write this as

$$\mathbb{E}[\mathbb{E}[(M_{t_i} - M_{t_{i-1}})^2 \mid \mathcal{F}_{t_{i-1}}] \mid \mathcal{F}_s]$$

(we can do this because $s \leq t_{i-1}$ by definition). Now there's nothing to do but to expand this square, and we get

$$\mathbb{E}[\mathbb{E}[M_{t_i}^2 \mid \mathcal{F}_{t_{i-1}}] - 2\mathbb{E}[M_{t_i}M_{t_{i-1}} \mid \mathcal{F}_{t_{i-1}}] + M_{t_{i-1}}^2 \mid \mathcal{F}_s].$$

And this type of calculation is very similar to what we just did with finding this L^2 martingale — you conditioned in $\mathcal{F}_{t_{i-1}}$, so you can pull out $M_{t_{i-1}}$ from the cross-term, and then use the martingale property on M_{t_i}. And then this middle term just becomes $M_{t_{i-1}}^2$, and you can group it with the last term; so this becomes

$$\mathbb{E}[\mathbb{E}[M_{t_i}^2 \mid \mathcal{F}_{t_{i-1}}] - M_{t_{i-1}}^2 \mid \mathcal{F}_s].$$

And now basically you just use tower again and remove the interior conditioning, so in the end you get

$$\mathbb{E}[M_{t_i}^2 - M_{t_{i-1}}^2 \mid \mathcal{F}_s].$$

Again, the key point was that

$$\mathbb{E}[M_{t_i}M_{t_{i-1}} \mid \mathcal{F}_{t_{i-1}}] = M_{t_{i-1}}^2,$$

and so we have a coefficient of $-2 + 1 = -1$.

And now when we sum over i, we get something telescoping, and everything cancels except the endpoints; that gives you the first identity.

And going from here to the second expression is the exact same computation — we expand out the square, and group the cross-term with the left endpoint. That proves this. □

This type of thing is going to be used all the time soon — even starting from the next chapter on semi-martingales, but definitely when we talk about stochastic integration.

§6.3.3 Maximal inequalities

Another thing to recall from discrete-time martingale theory that we're going to use:

Proposition 6.13 (Doob's maximal inequality)
If (X_n) is a supermartingale, then for all $\lambda > 0$ and times $k \geq 0$, we have

$$\lambda \mathbb{P}\left[\sup_{0 \leq n \leq k} |X_n| > \lambda\right] \leq \mathbb{E}[|X_0|] + 2\mathbb{E}[|X_k|].$$

(In this proposition we're in discrete time, which is why we index by n instead of t.)

This says you can bound the tail probability of your maximum by the expectations of the endpoints. To compare this with what you'd get from Markov's inequality, if you applied Markov's inequality at just the endpoint, what you'd get is

$$\lambda \mathbb{P}[|X_k| > \lambda] \leq \mathbb{E}[|X_k|].$$

If you tried to do something similar here, you'd have to bound the expectation of the sup. But you can't generally control that; what you *do* have control on is the sup of expectations.

So somehow this is something very special about supermartingales — the fact that you have this Markov-like inequality for the *sup* of your process is quite strong.

Proposition 6.14 (L^p maximal inequality)
If (X_n) is a martingale, then for all $k \geq 0$ and $p > 1$, we have

$$\mathbb{E}\left[\sup_{0 \leq n \leq k} |X_n|^p\right] \leq \left(\frac{p}{p-1}\right)^p \mathbb{E}[|X_k|^p].$$

Again this is quite a strong inequality — if your martingale was in L^p for some $p \geq 1$, then this lets you actually control the expectation of the sup, based on just the expectation of the endpoint. So if your martingale is in L^p, then you can always control the left-hand side. Often if you're stuck on some estimates, the missing thing is remembering this.

These are the discrete-time versions. We're taking them as givens, so we assume these are true; and we're going to use them to prove the corresponding statements in continuous time.

The standard approach to all these things is you just want to reduce the continuous-time result to the discrete-time one; so you're going to take an increasing mesh of times and reduce to this discrete-time result. But to be able to do that, you need to assume some sort of continuity.

Proposition 6.15

Let (X_t) be right-continuous.

(i) If (X_t) is a supermartingale, then

$$\lambda \mathbb{P}\left[\sup_{0 \le s \le t} |X_s| > \lambda\right] \le \mathbb{E}[|X_0|] + 2\mathbb{E}[|X_t|].$$

(ii) If (X_t) is a martingale, then

$$\mathbb{E}\left[\sup_{0 \le s \le t} |X_s|^p\right] \le \left(\frac{p}{p-1}\right)^p \mathbb{E}[|X_t|^p].$$

(Right-continuous means that it has right-continuous sample paths. Left-continuous probably also works.)

We'll again call these Doob's maximal inequality and the L^p maximal inequality; they're the same as the discrete-time ones, except we take a sup over continuous time.

Proof. For (i), let's take a sequence D_m of increasing *finite* subsets of $[0,t]$ such that $\bigcup_m D_m$ is dense in $[0,t]$, and such that your two endpoints 0 and t are contained in all the D_m's. (You can always ensure this; for example, you can take $D_m = \{it/m \mid 0 \le i \le m\}$.)

Then for each m, you can apply the discrete-time version. In discrete-time you usually index your property by n. Here our times are some arbitrary points, but that's fine — it's more just a notational difference. But if you do apply that proposition for discrete time, you get

$$\lambda \mathbb{P}\left[\sup_{s \in D_m} |X_s| > \lambda\right] \le \mathbb{E}[|X_0|] + 2\mathbb{E}[|X_t|].$$

And then you can take $m \to \infty$. And you note that

$$\sup_{s \in D_m} |X_s| \to \sup_{s \in [0,t]} |X_s|,$$

because you took the union of these D_m's to be dense. (Here we're also using the fact that X is right-continuous to get this convergence. The precise convergence result you might apply is e.g. Fatou's lemma; that would show the result.)

And that means as you take $m \to \infty$, this probability on the left converges to $\mathbb{P}[\sup_{s \in [0,t]} |X_s| > \lambda]$.

The proof of (ii) is basically the same — you take the same sequence, use the discrete-time result, and then take the limit, which you can justify by Fatou's lemma or monotone convergence or something like that. □

§6.4 Martingale convergence theorem

The first main theorem in dicrete-time martingale theory is the martingale convergence theorem — if your martingale is bounded in L^1, it converges almost surely. We basically want to prove the same result in continuous time, but before we do that, we need some setup.

We're going to first prove this convergence theorem for L^1-bounded martingales. Eventually we want to prove optional stopping theorems — they're one thing that makes martingales super powerful, because you usually use those to get explicit calculations. For that, you need the martingale convergence theorem and to talk about uniform integrability; so that's what we're working towards. And after we cover OST, we can discuss some applications to Brownian motion.

To prove the martingale convergence theorem in discrete time, you had to look at up-crossing numbers; so let's define what those are.

Definition 6.16. Let $f: I \to \mathbb{R}_+$ for some $I \subseteq \mathbb{R}$. For fixed $a < b$, we define the *up-crossing number* $M_{ab}^f(I)$ as the maximum integer k such that there exists a sequence $s_1 < t_1 < \cdots < s_k < t_k$ of elements of I such that $f(s_i) \leq a$ for all i, and $f(t_i) \geq b$ for all i.

(Despite the notation, I doesn't have to be an interval; it's just some arbitrary subset.)

Intuitively, this is the maximum number of up-crossings your function does on I — from below a to above b. So the picture to have in mind is that you're graphing your function f, and maybe it takes values at a discrete set of points, or maybe it's continuous in some parts (I could be arbitrary). And then you set two levels a and b (as horizontal lines). And you're basically asking for the number of times your function starts below a and goes above b.

So that's the picture; and this is how we formalize it.

Remark 6.17. As special cases, if no such choice of s_i's and t_i's exist, you set the up-crossing number to 0; if for all k such a choice exists, you set it to ∞ (i.e., you have an infinite number of up-crossings).

The key to the proof of the martingale convergence theorem in discrete time was Doob's up-crossing inequality, which we'll now state.

Proposition 6.18 (Doob's upcrossing inequality)

If (X_n) is a (discrete-time) supermartingale, then for all $n \geq 0$ and all $a < b$, we have

$$\mathbb{E}[M_{ab}^X(\{0, \ldots, n\})] \leq \frac{1}{b-a}\mathbb{E}[(X_n - a)^-].$$

So this says that for all n, the expected number of up-crossings of your process on some $[a, b]$ is bounded by something involving the endpoint of your process. (Here we think of X as the function.)

We're not going to prove this. The proof is something about defining the right martingale or submartingale that basically counts up-crossings, and then applying one of these submartingale expectation inequalities (maybe the maximal inequality), and that eventually leads you to this bound on the number of upcrossings. But we're just going to use this as input.

Theorem 6.19 (Supermartingale convergence)

Let (X_t) be a right-continuous supermartingale. If (X_t) is L^1-bounded (i.e., $\sup_t \mathbb{E}[|X_t|] < \infty$), then there exists a random variable $X_\infty \in L^1$ such that $X_t \to X_\infty$ almost surely (as $t \to \infty$).

So that's the supermartingale convergence theorem.

Proof. Again, you just kind of want to discretize. So let D be a countable dense subset of $\mathbb{R}_+$, and let's fix some time $T > 0$ (this is a fixed time, not a stopping time). Let (D_m) be a sequence of finite subsets of D such that $D_m \nearrow D$ and $T \in D_m$ for all m. (To make this super precise, what you probably want to do is first fix T and then let D be your countable dense subset — otherwise there's no way D can be countable and contain every possible time. So probably you first fix T, and then your countable dense subset D.)

By Doob's upcrossing inequality, for all $a < b$ and all m, we have

$$\mathbb{E}[M_{ab}^X(D_m \cap [0, T])] \leq \frac{1}{b-a}\mathbb{E}[(X_T - a)^-]$$

(this is just the up-crossing inequality applied to this specific set of finite times). Taking $m \to \infty$, we note that because D_m is increasing, this upcrossing number is monotone increasing (if you make your set of

possible times bigger, you can only have more upcrossings). So using monotone convergence, you get that

$$\mathbb{E}[M_{ab}^X(D \cap [0,T])] \le \frac{1}{b-a}\mathbb{E}[(X_T - a)^-].$$

(Basically the claim is that this thing with D_m converges upwards to the thing with D.)

Now what you do is send $T \to \infty$, and use the fact that X is L^1-bounded on the right. (The remark about swapping the orders doesn't really matter — you can actually send $T \to \infty$ through the integers. So then you *can* find some countable dense D which works for all integers T.)

Then we get that

$$\mathbb{E}[M_{ab}^X(D)] < \infty$$

(the right-hand side is uniformly bounded in T because X is L^1-bouned, and the left-hand side converges upwards to this).

And once you get here, the rest is just basically various analysis facts; we're out of time, so we'll end here and finish the proof next time. □

§7 February 24, 2025

§7.1 Supermartingale convergence theorem

Today we'll start by finishing the proof of the supermartingale convergence theorem.

Theorem 7.1 (Supermartingale convergence)
Let (X_t) be a right-continuous supermartingale. If (X_t) is L^1-bounded, then there is $X_\infty \in L^1$ such that

$$\lim_{t\to\infty} X_t = X_\infty \quad \text{almost surely.}$$

So L^1 boundedness implies almost sure convergence. This is one of the special things about the theory of martingales (or supermartingales).

Proof. Last time, we got up to the following statement: that

$$\mathbb{E}[M_{ab}^X(D)] < \infty,$$

where we fixed $a < b$ and a countable dense subset $D \subseteq \mathbb{R}_+$, and $M_{ab}^X(D)$ is the number of times you go from below a to above b when you restrict your times to D (the upcrossing number). This was basically reducing to the discrete-time Doob's upcrossing inequality; always with arguments of this flavor, you take some discrete subdivision and take something to ∞, and we got this.

Starting from here, this implies $M_{ab}^X(D)$ is finite almost surely. Now we can take a countable collection of a and b — we have that almost surely for all rationals $a < b$, the up-crossing number $M_{ab}^X(D)$ is finite.

And this is already enough to imply that X_t converges, at least within D — we get that almost surely,

$$X_\infty = \lim_{\substack{t\to\infty \\ t\in D}} X_t \quad \text{exists.}$$

(*A priori*, though, this limit might be $\pm\infty$.)

Why? It's easier to see the contrapositive. If the limit doesn't exist, that means $\liminf_t X_t < \limsup_t X_t$ — that's basically what it means for you not to have a limit. But then if this is true, you can find a rational

a slightly larger than the lim inf and b slightly smaller than the lim sup (with $a < b$), and for these your upcrossing number will be infinite (because you have to oscillate infinitely often between these two levels). So if the limit does not exist, then your upcrossing number must be infinite (for some a and b); and the contrapositive gives the above statement.

Even more, by Fatou's lemma, we get that

$$\mathbb{E}[|X_\infty|] \le \liminf_{D \ni t \to \infty} \mathbb{E}[|X_t|].$$

And the right-hand side is finite by our L^1 boundedness assumption. So we get that $|X_\infty| < \infty$ almost surely and $X_\infty \in L^1$.

The last thing to say is that by right-continuity, we actually have $\lim_{t\to\infty} X_t = \lim_{D\ni t\to\infty} X_t$ (i.e., the limit over all reals is the same as if you just restrict to your countable dense subset). And that basically finishes the proof. □

The point is that you just use the usual discrete-time ideas, but there's always some density argument behind it.

§7.2 Uniform integrability and closed martingales

So now we have this supermartingale convergence theorem. We'll want to get to optional stopping theorems; before that, we have to recall the definition of uniform integrability.

Definition 7.2. A martingale (X_t) is *closed* if there exists $Z \in L^1$ such that $X_t = \mathbb{E}[Z \mid \mathcal{F}_t]$ for all t.

By the tower property of conditional expectations, any process that satisfies this is a martingale; but this (being closed) is a stronger assumption.

Definition 7.3. We say a collection of random variables $(X_\lambda)_{\lambda\in\Lambda}$ (where Λ is an arbitrary index set) is *uniformly integrable* if:

(i) $\sup_{\lambda\in\Lambda} \mathbb{E}[|X_\lambda|] < \infty$.

(ii) For all $\varepsilon > 0$, there exists $\delta > 0$ such that for all A with $\mathbb{P}[A] \le \delta$, we have $\sup_{\lambda\in\Lambda} \mathbb{E}[|X_\lambda| \mathbf{1}_A] \le \varepsilon$.

Intuitively, the way to think about this is that it says the X_λ have 'uniformly decaying tails.'

This was covered in 18.675, so we won't go into the proofs of the basic properties; but we'll use some of the main theorems that were proved about them.

Exercise 7.4. A single L^1 random variable is uniformly integrable; more generally, any finite collection of L^1 random variables is uniformly integrable.

This is a nice exercise ((i) is clear, but (ii) is a good exercise to check).

Theorem 7.5

Let (X_t) be a right-continuous martingale. Then the following are equivalent:

(i) X is closed.

(ii) X is uniformly integrable.

(iii) X_t converges almost surely and in L^1.

Moreover, if any of these properties hold, then we have $X_t = \mathbb{E}[X_\infty \mid \mathcal{F}_t]$ for all t, where X_∞ is the almost sure limit of X_t.

Saying X is closed just means there's some Z with $X_t = \mathbb{E}[Z \mid \mathcal{F}_t]$; the point of the final statement is that Z is just the most natural variable you'd think of.

Proof. (This proof is kind of cheating because we'll just apply some results and not give the details, since it was covered in 18.675.)

First, why does (i) imply (ii)? More generally, if you start with $Z \in L^1$, then the collection $(\mathbb{E}[Z \mid \mathcal{G}])_{\mathcal{G} \subseteq \mathcal{F}}$ (where $\mathcal{G}$ ranges over *all* sub-σ-algebras) is uniformly integrable. The fact that this is true definitely implies the $(i) \to (ii)$ implication, since you can just take a subcollection of $\mathcal{G}$'s (the ones corresponding to your filtration). This result is again a good exercise if you haven't seen it before (but before you prove this, Sky recommends seeing why a *single* L^1 random variable is uniformly integrable).

For $(ii) \to (iii)$, by the supermartingale convergence theorem we have that $X_t \to X_\infty$ almost surely (for some X_∞), because one condition of uniform integrability is that you're L^1 bounded. And now uiform integrability implies that $X_t \to X_\infty$ in L^1. (This is the whole point of introducing uniform integrability — you want to say when almost sure convergence implies convergence in L^1, and this is exactly the condition you need. This result is sometimes called Vitali's theorem; and that's what we're applying here.)

Finally, for $(iii) \to (i)$, just by the martingale property we have

$$X_t = \mathbb{E}[X_{t_0} \mid \mathcal{F}_t] \quad \text{for all } t_0 \geq t.$$

Now we want to take $t_0 \to \infty$. Using the assumption that $X_{t_0} \to X_\infty$ almost surely and in L^1, we have that $\mathbb{E}[X_{t_0} \mid \mathcal{F}_t] \to \mathbb{E}[X_\infty \mid \mathcal{F}_t]$ in L^1. And that's basically enough — it implies that $X_t = \mathbb{E}[X_\infty \mid \mathcal{F}_t]$. (If X_t converges almost surely and in L^1, then the almost sure limit — which is X_∞ — also has to be in L^1. And if you have a sequence of random variables converging in L^1, then it's a fact that any conditional expectation also converges in L^1. And the sequence $\mathbb{E}[X_{t_0} \mid \mathcal{F}_t]$ is actually constant — it's actually equal to X_t.)

So that shows that X is closed, and it also shows the last statement (since we explicitly exhibited that X_∞ is the random variable making this collection closed). □

§7.3 Optional stopping theorems

Now that we have this concept of uniform integrability, we can finally get to optional stopping theorems.

Let (X_t) be a right-continuous martingale such that $X_t \to X_\infty$ almost surely (right now, we're not assuming any integrability conditions).

Definition 7.6. Given a stopping time T, we define

$$X_T = \mathbf{1}[T < \infty] X_T + \mathbf{1}[T = \infty] X_\infty.$$

The point is that before, we only defined X_T when $T < \infty$; but now that we have a notion of X_∞, we can also define this when $T = \infty$.

We can verify that $X_T \in \mathcal{F}_\infty$ (last time we saw that $\mathbf{1}[T < \infty] X_T$ is $\mathcal{F}_T$-measurable, and $\mathcal{F}_T \subseteq \mathcal{F}_\infty$, and you can explicitly check the second part is also $\mathcal{F}_\infty$-measurable).

Theorem 7.7 (Optional stopping theorem)

Let (X_t) be a right-continuous uniformly integrable martingale, and let S and T be stopping times such that $S \leq T$ (i.e., $S(\omega) \leq T(\omega)$ for all ω). Then $X_S, X_T \in L^1$, and

$$X_S = \mathbb{E}[X_T \mid \mathcal{F}_S].$$

In particular, $X_S = \mathbb{E}[X_\infty \mid \mathcal{F}_S]$, and $\mathbb{E}[X_S] = \mathbb{E}[X_\infty] = \mathbb{E}[X_0]$.

Note that we're not assuming S and T are finite — we have a uniformly integrable martingale, so we converge to some X_∞ almost surely and in L^1.

This essentially states that we still have the martingale property for these *random* times. The second statement comes from taking $T = \infty$ (which is certainly a stopping time); the third comes from taking expectations (we have $\mathbb{E}[X_\infty] = \mathbb{E}[X_0]$ because we know $\mathbb{E}[X_t] = \mathbb{E}[X_0]$ for all t, and we can take $t \to \infty$ because we have L^1 convergence).

So this second part follows as soon as we show the first part.

Proof. For this, it's helpful to do a discrete approximation (as usual). The point is that you want to apply the discrete-time optional stopping theorem, so that's what we're going to reduce to.

So for $n \geq 1$, let

$$T_n = \sum_{k=1}^{\infty} \mathbf{1}[(k-1)2^{-n} \leq T \leq k2^{-n}]k2^{-n} + \mathbf{1}[T = \infty]\infty$$

(this is a discrete approximation of T, where you split time into a bunch of intervals of length 2^{-n}, and take the right endpoint of the interval T lies in). Define S_n similarly.

Last time we had a lemma showing that S_n and T_n are stopping times; $S_n \searrow S$ and $T_n \searrow T$; and $S_n \leq T_n$.

Now we claim that for each fixed n, the times $2^n S_n$ and $2^n T_n$ (the multiplication by 2^n is so that these are integer-valued — we're doing that just because discrete-time stopping theorems are for integer values) are stopping times of the discrete-time filtration $\mathcal{H}_k^{(n)} = \mathcal{F}_{k2^{-n}}$ (here we've fixed n, and the time index is k). The point is that we're changing time by units of 2^{-n} — we went from units of 2^{-n} to integer values, which means you want to do that here too. (You can explicitly verify this; we're not going to do so.)

So we have this discrete-time filtration. Now let's define our discrete-time martingale

$$Y_k^{(n)} = X_{k2^{-n}}$$

(again we're fixing n, and k is our time parameter). Since our original process X was adapted to the continuous-time filtration, this process is adapted to the discrete-time one. And it's also a uniformly untegrable martingale with respect to this discrete-time filtration.

So now we've reduced to this discrete-time scenario, and now we can apply the discrete-time optional stopping theorem. So we get that

$$X_{S_n} = Y^{(n)}_{2^n S_n} = \mathbb{E}[Y^{(n)}_{2^n T_n} \mid \mathcal{H}^{(n)}_{2^n S_n}]$$

(this is where we use the discrete-time optional stopping theorem). And now it's kind of natural that given the identity $\mathcal{H}_k^{(n)} = \mathcal{F}_{k2^{-n}}$, we should also get that this is equal to

$$\mathbb{E}[Y^{(n)}_{2^n T_n} \mid \mathcal{F}_{S_n}].$$

Formally this is true because you can plug in S_n in place of k; but you actually have to verify something, because S_n is a *random* time (so this is actually a stopped σ-algebra); so let's verify why this identity is true, i.e., that

$$\mathcal{H}^{(n)}_{2^n S_n} = \mathcal{F}_{S_n}.$$

To see this, we just have to write out definitions. Recall that

$$\mathcal{H}^{(n)}_{2^n S_n} = \{A \in \mathcal{H}^{(n)}_\infty \mid A \cap \{2^n S_n \leq k\} \in \mathcal{H}^{(n)}_k \text{ for all } k\}$$

(where k is an integer). By definition, $\mathcal{H}^{(n)}_\infty = \mathcal{F}_\infty$, and $\mathcal{H}^{(n)}_k = \mathcal{F}_{k2^{-n}}$. Meanwhile, if we write out the definition of the continuous-time stopped σ-algebra, it's

$$\mathcal{F}_{S_n} = \{A \in \mathcal{F}_\infty \mid A \cap \{2^n S_n \leq t\} \in \mathcal{F}_t \text{ for all } t\}.$$

The second one comes with more conditions, so the second σ-algebra is contained in the first. To verify the reverse (i.e., that it's enough to have this discrete set of conditions), the point is that $2^n S_n$ only takes values in these discrete values — we can write

$$\{2^n S_n \le t\} = \{2^n S_n \le \lfloor t \rfloor\}.$$

The point is just that it's enough to verify your condition at integers, because this thing $2^n S_n$ is just integer-valued.

So we've now arrived at the identity

$$X_{S_n} = \mathbb{E}[Y^{(n)}_{2^n T_n} \mid \mathcal{H}^{(n)}_{2^n S_n}] = \mathbb{E}[Y^{(n)}_{2^n T_n} \mid \mathcal{F}_{S_n}].$$

By definition, we have $Y^{(n)}_{2^n T_n} = X_{T_n}$, so in summary, we get that

$$X_{S_n} = \mathbb{E}[X_{T_n} \mid \mathcal{F}_{S_n}].$$

We're trying to show that $X_S = \mathbb{E}[X_T \mid \mathcal{F}_S]$. So how do we conclude? You might try to take a limit as $n \to \infty$. But this is somewhat less clear because now the σ-algebra is also varying in n, which is kind of problematic. (If we just had $\mathcal{F}_S$, it would probably be fine.)

So to get around that, we kind of go back to the definition of conditional expectation. We want to take some $A \in \mathcal{F}_S$ (note that $S \le S_n$, so this means $\mathcal{F}_S \subseteq \mathcal{F}_{S_n}$, which means A is also in $\mathcal{F}_{S_n}$). And we want to show that

$$\mathbb{E}[X_S \mathbf{1}_A] = \mathbb{E}[X_T \mathbf{1}_A].$$

(Also, as a consequence of the discrete-time optional stopping theorem, we get that $X_{S_n}, X_{T_n} \in L^1$. So we can actually talk about conditional expectations.)

So we want to show this identity. And because we have $\mathcal{F}_S \subseteq \mathcal{F}_{S_n}$, we have this identity for the discrete approximations — that

$$\mathbb{E}[X_{S_n} \mathbf{1}_A] = \mathbb{E}[X_{T_n} \mathbf{1}_A].$$

Now we send $n \to \infty$. We have $X_{S_n} \to X_S$ pointwise (surely, by construction). But to get the convergence of expectations, we need additional integrability. So we use the fact that (X_{S_n}) is uniformly integrable — this is again by the discrete-time optional stopping theorem (we have that $\mathbb{E}[X_\infty \mid \mathcal{F}_{S_n}] = X_{S_n}$, and we can use the fact about conditional expectations being uniformyl integrable to conclude). So by combining the sure convergence with uniform integrability, you get

$$\mathbb{E}[X_{S_n} \mathbf{1}_A] \to \mathbb{E}[X_S \mathbf{1}_A].$$

And the same is true for T_n and T, so we get $\mathbb{E}[X_S \mathbf{1}_A] = \mathbb{E}[X_T \mathbf{1}_A]$, and we're done. □

Now let's look at some consequences of this result, and then we'll apply some of this stuff to Brownian motion.

Corollary 7.8

Let (X_t) be a right continuous martingale (not necessarily uniformly integrable), and let $S \le T$ be *bounded* stopping times. Then $X_S, X_T \in L^1$ and $X_S = \mathbb{E}[X_T \mid \mathcal{F}_S]$.

So we're trading the uniform integrability assumption for this boundedness assumption. In other words, if you have bounded stopping times, then you don't need uniform integrability.

Proof. Let a be such that $S \leq T \leq a$ (by assumption, there's some deterministic a which upper-bounds your stopping times). Then we apply this theorem to the process $(X_{t\wedge a})_t$ (it turns out this is also a martingale — if you stop your martingale at any deterministic time a, it's still a martingale).

We claim that this is actually uniformly integrable — in particular, it's going to be closed by X_a, meaning that

$$X_{t\wedge a} = \mathbb{E}[X_a \mid \mathcal{F}_t].$$

As soon as we show it's closed, it's a uniformly integrable martingale, so we can apply the previous theorem to get that this is true (because for instance, $X_{S\wedge a} = X_S$, and similarly $X_{T\wedge a} = X_t$; so the main statement of that theorem just applies).

So the main thing to do is show why your stopped process is actually closed by the endpoint The natural reason why this should be true is if you take $t \to \infty$, $X_{t\wedge a} \to X_a$; so if you believe your process should be closed, then it should just be closed by X_a.

So let's verify that.

Case 1 ($t \leq a$). Then we have the martingale identity $X_t = \mathbb{E}[X_a \mid \mathcal{F}_t]$. But $X_t = X_{t\wedge a}$, and that's exactly what we wanted.

Case 2 ($t > a$). Then $X_{t\wedge a} = X_a = \mathbb{E}[X_a \mid \mathcal{F}_t]$ (because X is adapted, so X_a will be $\mathcal{F}_t$-measurable because $t > a$ so $\mathcal{F}_t$ contains $\mathcal{F}_a$), so this is true. □

Corollary 7.9

Let (X_t) be a right continuous martingale, and let T be a stopping time.

(i) The stopped process $(X_{t\wedge T})_t$ is a martingale.

(ii) If (X_t) is uniformly integrable, then $(X_{t\wedge T})$ is also uniformly integrable. Moreover, for all t, we have

$$X_{t\wedge T} = \mathbb{E}[X_T \mid \mathcal{F}_t].$$

Again, if $(X_{t\wedge T})$ is uniformly integrable, then it has to be closed by X_T — this is the almost sure limit, and if we take $t \to \infty$ then it converges to X_T. So this is the only possibility.

Basically this is saying that if you consider your stopped martingale at a stopping time, it's still a martingale; and it's still UI if the original was.

Proof. We'll first show (ii). First, note that $T \wedge t$ is a stopping time (using the properties of stopping times from last week). And we also have that $X_{t\wedge T}, X_T \in L^1$ (since we showed when you have a uniformly integrable martingale and evaluate it at any stopping time, th eresult is also uniformly integrable). And $X_{t\wedge T}$ is $\mathcal{F}_{t\wedge T}$-measurable (we proved this last week, at least on the event that $t \wedge T$ is finite; and this is definitely true, since it's at most t). And $\mathcal{F}_{t\wedge T} \subseteq \mathcal{F}_T$, because $t \wedge T \leq T$.

So what's the consequence of this? At least the fact that $X_{t\wedge T}$ is $\mathcal{F}_T$-measurablee is consistent with the claim $X_{t\wedge T} = \mathbb{E}[X_T \mid \mathcal{F}_t]$. Then using the definition of conditional expectation, it remains to show that for all $A \in \mathcal{F}_t$, we have

$$\mathbb{E}[X_{t\wedge T}\mathbf{1}_A] = \mathbb{E}[X_T\mathbf{1}_A]$$

(once we show this, we're done).

And by the previous theorem, we have that

$$X_{t\wedge T} = \mathbb{E}[X_T \mid \mathcal{F}_{t\wedge T}]$$

(again, because we started with a uniformly integrable martingale X and we had the stopping times $t\wedge T \leq t$, and the previous theorem gave us the martingale property for these stopping times).

Now to use this to verify what we want, we can write

$$\mathbb{E}[X_{t\wedge T}\mathbf{1}_A] = \mathbb{E}[X_{t\wedge T}\mathbf{1}_{A\cap\{T>t\}}] + \mathbb{E}[X_{t\wedge T}\mathbf{1}_{A\cap\{T\le t\}}]$$

(all we're doing is splitting based on whether $T \le t$).

Now, if $T \le t$, then we just get X_T. To handlethe first term, we're going to use the above identity of conditional expectations — the point is that $A \cap \{T > t\}$ is $\mathcal{F}_{t\wedge T}$-measurable. Given this claim, we can basically replace $X_{T\wedge t}$ with X_T by the definition of conditional expectations; so we get

$$\mathbb{E}[X_T\mathbf{1}_{A\cap\{T>t\}}] + \mathbb{E}[X_T\mathbf{1}_{A\cap\{T\le t\}}] = \mathbb{E}[X_T\mathbf{1}_A].$$

So it remains to show why $A \cap \{T > t\} \in \mathcal{F}_{t\wedge T}$. To do so, we have to show that

$$A \cap \{T > t\} \cap \{t \wedge T \le s\} \in \mathcal{F}_{s\wedge T}.$$

We can write this event as

$$A \cap \{T > t\} \cap \{t \le s\}$$

(because if $T \ge t$, then $t \wedge T$ is just t itself). Now if $s \ge t$, then this is just $A \cap \{T > t\}$; and if $s < t$, then this is just $\emptyset$. And $\emptyset$ is definitely in any σ-algebra you care about, so we just need to show that $A \cap \{T > t\}$ is in $\mathcal{F}_{s\wedge T}$.

And it's definitely in $\mathcal{F}_t \subseteq \mathcal{F}_s$. So because $\mathcal{F}_{s\wedge T} = \mathcal{F}_s \cap \mathcal{F}_T$, we just need to show that it's also in $\mathcal{F}_T$; and this is because you can again intersect with an event like $\{T \le s\}$, so you want to show that $A \cap \{T > t\} \cap \{T \le s\} \in \mathcal{F}_s$, which we essentially just did.

So this finishes the proof of (ii) — if you have a uniformly integrable right-continuous martingale, then the stopped process is also a uniformly integrable martingale. Now let's use that to show (i).

For (i), we can appy (ii) with the process $(X_{t\wedge a})_t$. We saw that $(X_{t\wedge a})$ is always uniformly integrable, because it's closed by a. So we get that for all t,

$$X_{t\wedge a\wedge T} = \mathbb{E}[X_{a\wedge T} \mid \mathcal{F}_{t\wedge a\wedge T}].$$

(Here we're replacing t by $t \wedge a$ everywhere in (ii), and we get this identity.) In particular, for $t \le a$, we get

$$X_{t\wedge T} = \mathbb{E}[X_{a\wedge T} \mid \mathcal{F}_{t\wedge T}].$$

But this is precisely the martingale identity; so this shows why the stopped process $(X_{t\wedge T})_t$ is a martingale. □

§7.4 Applications to Brownian motion

Now that we're done with that, we can compute some properties of stopping times of Brownian motion.

Let (B_t) be a Brownian motion. It's a martingale with respect to its canonical filtration (implicitly, that's the filtration we'll be using throughout, at least in this part).

Recall that for any $a \in \mathbb{R}$, we defined

$$T_a = \inf\{t \ge 0 \mid B_t = a\}$$

as the first hitting time to a. Now let's take two levels $a < 0 < b$, and let's consider $T = T_a \wedge T_b$.

First, why is T_a a stopping time? We already showed this (you can relate it to a property about the sup or inf of B, which is measurable; or you can use the fact that $\{a\}$ is closed, and the first hitting time of

a continuous martingale for a closed set is a stopping time). And the min of two stopping times is also a stopping time, so T is too.

Now if we consider the stopped process $B_{t\wedge T}$, this is a martingale by what we just proved. And it's also bounded — B starts at 0, so if you're before the first time you hit a or b, then you have to be in $[-a, b]$. So this thing is bounded, which in particular means it's uniformly integrable.

So the stopped process $(B_{t\wedge T})_t$ is a uniformly integrable martingale, and we thus obtain that

$$\mathbb{E}[B_{\infty|T}] = \mathbb{E}[B_0 \mid T] = 0.$$

(This is one consequence of the optional stopping theorem.)

And $T < \infty$ almost surely — we proved in class that $\limsup B_t = \infty$ and $\liminf B_t = -\infty$ almost surely, so these times are both finite almost surely, and you took a min.

In particular, this means $B_{\infty\wedge T} = B_T$; and this is a if $T_a < T_b$, or b if $T_b > T_a$. so we get

$$\mathbb{E}[a\mathbf{1}[T_a < T_b] + b\mathbf{1}[T_b > T_a]] = 0,$$

which we can rewrite as

$$a\mathbb{P}[T_a < T_b] + b(1 - \mathbb{P}[T_a < T_b]) = 0.$$

(We're taking expectations of an indicator, so we get probabilities; and either $T_a < T_b$ or $T_b < T_a$ — they're both finite almost surely, and they can't be equal.)

But now you can just solve this equation, and you get that

$$\mathbb{P}[T_a < T_b] = \frac{b}{b-a}.$$

(Note that a is negative.) This also implies

$$\mathbb{P}[T_b < T_a] = \frac{-a}{b-a}.$$

So that's the first application.

Next, for $a > 0$, let $U_a = T_a \wedge T_{-a}$ — this is the first time your absolute value reaches level a, i.e.,

$$U_a = \inf\{t \mid |B_t| = a\}.$$

Now recall that we had the quadratic martingale $(B_t^2 - t)_t$ from last week. And now we can stop it at the stopping time U_a; so we get the martingale

$$(B_{t\wedge U_a}^2 - (t \wedge U_a))_t.$$

Then just using the martingale property, at any fixed time t this thing has expectation 0. If you turn that around, that tells you

$$\mathbb{E}[B_{t\wedge U_a}^2] = \mathbb{E}[t \wedge U_a] \quad \text{for all } t.$$

And now we just want to send $t \to \infty$; the right-hand side is monotone in t, so we can apply monotone convergence to handle it. To handle the left-hand side, note that $B_{t\wedge U_a} \to \pm a$ almost surely as $t \to \infty$. So naturally, you want to say the expectation converges to a^2. You can't verify that using the monotone convergence theorem, because this thing is not monotone. But you can verify it using the dominated convergence theorem — because $|B_{t\wedge U_a}| \le a$ (before you hit absolute value a, you have to have absolute value less than a). So we actually get that $\mathbb{E}[U_a] = a^2$.

So far, we've used the linear martingale of Brownian motion (which is Brownian motion itself) and the quadratic martingale; so now let's use the exponential martingale. Let $a > 0$. For $\lambda \in \mathbb{R}$, recall that we defined the exponential martingale

$$N_t^\lambda = \exp\left(\lambda B_t - \frac{\lambda^2 t}{2}\right)$$

(the way to remember what to subtract is that we want this thing to have mean-1 for all t, which means you divide by the moment generating function of B_t). We're going to take $\lambda > 0$ here. Now if we consider the stopped process $(N^\lambda_{t\wedge T_a})$, it's a martingale. Moreover, it's bounded — the second part $\lambda^2 t/2$ is fine (since it's positive), and for the first part we note that $B_{t\wedge T_a} \le a$, so $\lambda B_{t\wedge T_a} \le \lambda a$, and therefore

$$N^\lambda_{t\wedge T_a} \le e^{\lambda a}.$$

So it's bounded, which in particular means it's uniformly integrable.

And now we can apply the optional stopping theorem. So we get that

$$\mathbb{E}[N^\lambda_{T_a}] = \mathbb{E}[N_0^\lambda] = 1.$$

And now if we recall what the exponential martingale is, at time T_a we have $B_t = a$, so we get

$$e^{\lambda a}\mathbb{E}[e^{-\lambda^2 T_a/2}] = 1,$$

which implies that

$$\mathbb{E}[e^{-\lambda^2 T_a/2}] = e^{-\lambda a}.$$

And now for reparametrizing, if we replace $\lambda \mapsto \sqrt{2\lambda}$, we get

$$\mathbb{E}[e^{-\lambda T_a}] = e^{-(\sqrt{2\lambda})a}.$$

So we've basically computed the Laplace transform of T_a as a random variable. And that uniquely characterizes the law of T_a. So that's one application — if you want to find the density of T_a, in principle there's some process to invert the Laplace transform to find the density.

But another way you could've found the distribution of T_a was the reflection principle, which was on the homework — there you found the joint density of (S_t, B_t), and if you integrate out over B_t you get the density of S_t. And we can write $\{T_a \le t\} = \{S_t \ge a\}$; so that's another way you could've found the distribution of T_a.

One final application is that by reflection symmetry, we know

$$\mathbb{E}[\mathbf{1}[B_{U_a} = a]e^{-\lambda U_a}] = \mathbb{E}[\mathbf{1}[B_{U_a} = -a]e^{-\lambda U_a}]$$

(recall that U_a is the first time you get to $\pm a$) — the point is that if you hit a first, then you can reflect the picture so that you hit $-a$ first.

Eventually, we want to compute the Laplace transform of U_a. If you try to do something similar, it's not exactly going to work, because you don't exactly know what B_{U_a} is — it could be a or $-a$. But the point is that the contribution is going to be equal. And that allows you to get that these are both equal to $\frac{1}{2}\mathbb{E}[e^{-\lambda U_a}]$ (since if you added them together you'd get precisely $\mathbb{E}[e^{-\lambda U_a}]$; and they're equal, so you get the $\frac{1}{2}$).

And now we're still going to use the argument with N_t^λ — we get

$$\mathbb{E}[N^\lambda_{U_a}] = 1$$

(now the exponent is not even just bounded on one side, it's bounded on both sides by the definition of the stopping time). And if we write out what this means, it's

$$\mathbb{E}[e^{\lambda a}\mathbf{1}[B_{U_a} = a]e^{-\lambda^2 U_a/2}] + \mathbb{E}[e^{-\lambda a}\mathbf{1}[B_{U_a} = a]e^{-\lambda^2 U_a/2}].$$

And the point of the previous argument is that this is

$$\frac{e^{\lambda a}+e^{-\lambda a}}{2}\mathbb{E}[e^{-\lambda^2 U_a/2}]=1.$$

So you get a formula for the Laplace transform of U_a (which contains some 1/cosh).

This completes the unit on martingales and filtrations; for the next two weeks we'll talk about a generalization called *continuous semimartingales*, and then we'll get to stochastic integration.

§8 February 26, 2025

Today we'll start the chapter on continuous semimartingales; these will be both the integrators and integrands for our stochastic integrals, as we'll see later.

§8.1 Intuition for continuous semimartingales

The simplest example of a semimartingale is Brownian motion with drift — specifically, $\sigma B_t+\mu t$. So we have a straight line μt which is your drift, and then you have some fluctuations on top of your straight line. In some sense, continuous semimartingales are basically just locally supposed to look like Brownian motion with drift.

Continuous martingales basically *are* Brownian motions locally — in the sense that you can find a time-change so that you get a Brownian motion. So at least locally, if you actually graph the process, the fluctuations are supposed to look like BM, for a martingale.

But semimartingales are a bit more general — they don't just look like BM, but they also have a drift term (at least locally). So there's always a direction you're trending towards. It doesn't have to be a straight line — your drift can be some curvy thing, and then on top of this you have BM-type fluctuation.

So these are some pictures to have in mind. In this chapter, all processes are indexed by $\mathbb{R}_+=[0,\infty)$ and are real-valued.

§8.2 Finite variation functions

You can always write a semimartingale as basically a martingale part plus a drift part; we'll start by discussing the drift part.

The first part of the discussion is just some analytic facts, so probability isn't coming in yet.

Definition 8.1. A *signed measure* on a finite interval $[0,T]$ is a difference between two finite positive measures on $[0,T]$.

Definition 8.2. Let $T\geq 0$. We say a continuous function $a:[0,T]\to\mathbb{R}$ with $a(0)=0$ has *finite variation* (FV) if there exists a signed measure μ such that

$$a(t)=\mu([0,t])\quad\text{for all } t\in[0,T].$$

We're being maybe less general than you have to be — in principle you don't have to assume your function is continuous, but we're going to.

By general arguments (for instance, using the π–λ theorem), one can show that a determines μ. Indeed, if μ were a probability measure, then a is the cumulative distribution function of it, and your CDF determines

the measure. Here we're being slightly more general (the measure doesn't have to be a probability measure, or even positive), but the same proof just works — your 'CDF' (or really, the analog of the CDF) determines the measure.

And since we're always assuming a is continuous and starts at 0, the masure μ has no atoms — basically, you can't have jumps in your function. (You don't necessarily have to assume a is continuous; if you allowed discontinuous functions, then the associated measure might have atoms. But we'll assume this is never the case.)

One analytic lemma: By definition, a signed measure means you can decompose it as a difference of two positive measures. You might wonder when that decomposition is unique. In principle it never is, because you can always add one piece to both measures — if you had $\mu = \mu_1 + \mu_2$, you could always write $\mu = (\mu_1 + \nu) - (\mu_2 + \nu)$. So there's no way it can be unique. But the next lemma says that if you assume some further things about μ_1 and μ_2, then it *does* have to be unique.

Lemma 8.3

Let a be a finite variation function on some interval $[0, T]$, with associated measure μ. Then there is a unique decomposition

$$\mu = \mu_+ - \mu_-$$

such that μ_+ and μ_- are supported on disjoint Borel sets.

(We're not writing it, but here μ_+ and μ_- are finite positive measures.) So you just need to add this additional condition that they're disjointly supported to get the unique decomposition.

Proof. Let's start with some arbitrary decomposition $\mu = \mu_1 - \mu_2$ (we know one exists by the definition of a signed measure). Let $\nu = \mu_1 + \mu_2$. One can check that μ_1 and μ_2 are both absolutely continuous with respect to ν — if $\nu(A) = 0$, then $\mu_1(A) = \mu_2(A) = 0$ (this is just by definition, since μ_1 and μ_2 are positive). So the Radon–Nikodym theorem gives measurable functions h_1 and h_2 which are basically the densities of μ_1 and μ_2 with respect to ν, meaning that $d\mu_1 = h_1\, d\nu$ and $d\mu_2 = h_2\, d\nu$.

Now let $h = h_1 - h_2$. Then you have that h is the density of your starting measure μ with respect to ν, i.e., $d\mu = h\, d\nu$.

So now we've found this density of our starting measure with respect to some base measure. And the natural thing to do is to split based on the sign of h — we can decompose $h = h^+ - h^-$. Clearly these are supported on different sets — at any given point, at most one of h^+ and h^- is nonzero. So we can let $d\mu_+ = h^+\, d\nu$ (so the density of μ_+ is h^+), and similarly $d\mu_- = h^-\, d\nu$. Then we get $\mu = \mu_+ - \mu_-$ (again, because the densities are decomposed in this way).

So we've found a decomposition into measures supported on disjoint sets (because h^+ and h^- are) — we have $\operatorname{supp}(\mu_\pm) \subseteq \{t \in [0, T] \mid h^\pm(t) > 0\}$; we'll denote these sets by D_+ and D_-, with $D_+ \cap D_- = \emptyset$.

Now we'll show uniqueness. For this, we claim that you can find μ_+ purely in terms of μ itself — we must have

$$\mu_+(A) = \sup\{\mu(C) \mid C \in \mathcal{B}([0, T]),\, C \subseteq A\}.$$

So the claim is that for any such decomposition into μ_+ and μ_-, this has to be true. And the right-hand side doesn't really see your decomposition, so that would show uniqueness.

To see this, suppose we have some decomposition $\mu = \mu_+ - \mu_-$ (not necessarily the one we constructed above). By definition we have

$$\mu(C) = \mu_+(C) - \mu_-(C) \le \mu_+(C) \le \mu_+(A)$$

(since μ_- is a positive measure and $C \subseteq A$). So that shows one direction of the equality — that $\mu_+(A) \ge \sup\{\mu(C)\}$. To finish, we just need to show the other direction. Basically, we want to exhibit C for which

$\mu_+(A) = \mu(C)$. For this, we know μ_+ is supported on some set D_+, so $\mu_+(A) = \mu_+(A \cap D_+)$. (Note that D_+ is not necessarily the one we constructed; it's just the support of μ_+, which by hypothesis is disjoint from the support of D_-.) And we can write this as

$$\mu_+(A \cap D_+) - \mu_-(A \cap D_+),$$

using the fact that μ_+ and μ_- have disjoint supports (so the second term is just 0). And this is precisely $\mu(A \cap D_+)$ — and $A \cap D_+$ is some set C contained in A, so that shows the reverse inequality. □

So we're able to decompose these finite variation functions, more or less.

Notation 8.4. We write $|\mu| = \mu_+ + \mu_-$; we call this the *total variation* of μ.

If μ is a signed measure, then this is a finite positive measure.

Fact 8.5 — We have $|\mu(A)| \leq |\mu|(A)$ for all $A \in \mathcal{B}([0,T])$.

This is because the left-hand side is $|\mu_+(A) - \mu_-(A)|$.

Moreover, since we have

$$\mu(A) = \mu_+(A \cap D_+) - \mu_-(A \cap D_-)$$

(where $D_\pm$ are the supports of $\mu_\pm$), this is bsaically saying that

$$\frac{d\mu}{d|\mu|} = \mathbf{1}_{D+} - \mathbf{1}_{D_-}$$

(where the left-hand side denotes the Radon–Nikodym derivative of μ with respect to $|\mu|$) — because by definition, what it means to be a density is basically the above equation. So μ has a density with respect to $|\mu|$, and it's basically given by this difference of indicators on its positive and negative parts.

As some more remarks, if we have a finite variation function a, then by definition

$$a(t) = \mu([0,t]) = \mu_+([0,t]) - \mu_-([0,t]).$$

So we get that a is the difference of two nondecreasing continuous functions which vanish at 0. Continuity follows because μ has no atoms, so neither can μ_+ or μ_-; so $\mu_\pm([0,t])$ also define continuous functions.

Another remark is that the converse is also true — if you had two continuous nondecreasing functions vanishing at 0, then when you take the difference, you get a FV function. To do that, you kind of need to use Caratheodory's extension theorem — if you have $a(t) = a_1(t) - a_2(t)$ (where a_1 and a_2 are nondecreasing continuous functions vanishing at 0), you want to go from this to a signed measure; so you want a_1 and a_2 to be the CDFs of some positive measure. And to do that, you need to use Caratheodory's extension theorem.

So these are all different ways of thinking of FV functions.

Definition 8.6. Let $f : [0,T] \to \mathbb{R}$ be a measurable function such that $\int_{[0,T]} |f(s)|\,|\mu|\,(ds) < \infty$. Then we define

$$\int_{[0,T]} f(s)\,da(s) = \int_{[0,t]} f(s)\mu(ds).$$

Similarly, we define $\int_{[0,t]} f(s)\,|da(s)| = \int_{[0,T]} f(s)\,|\mu|\,(ds)$.

Here in the background we're assuming we have a FV function, which gives us some μ; and we take its total variation $|\mu|$, and we assume f is integrable with respect to that.

What does this mean? The point is that da is basically a measure; so integrating against da basically means integrating against the associated measure.

Example 8.7

If $a(t) = t$, then what would you want to define $\int_0^T f(s)\,da(s)$? Here da is just ds, so naturally you'd want to define this as $\int_0^T f(s)\,ds$. And that's precisely what this integral does — we're basically integrating our function with respect to the Lebesgue measure on this interval.

So this example is at least consistent with what you would expect.

Fact 8.8 — We have $\left|\int_0^T f(s)\,da(s)\right| \le \int_0^T |f(s)|\,|da(s)|$.

You basically use the definition of what da is and split $\mu = \mu_+ - \mu_-$. On the left you have a difference, and on the right you have a sum of positive things; so that's why you get this inequality.

As another remark, by restricting your finite variation a to a smaller interval $[0,t] \subseteq [0,T]$, you can always define these smaller integrals

$$\int_0^t f(s)\,da(s) \quad \text{and} \quad \int_0^t f(s)\,|da(s)|\,.$$

(You basically just restrict μ to this smaller interval and view this thing as a Lebesgue integral.)

Fact 8.9 — The function $t \mapsto \int_0^t f(s)\,da(s)$ is finite variation.

This is under the assumption that f is integrable against $|\mu|$. And the reason is that the associated measure is just going to be $f\,d\mu$ — you can define a measure whose density with respect to μ is f; and unwinding the definitions, you'll see this is precisely the measure associated to this. It starts at 0 and is continuous (e.g., by dominated convergence), and it comes from an actual signed measure, namely this one.

Now let's get to a proposition, which basically tells you how to approximate the total variation measure.

Proposition 8.10

For all $t \in [0,T]$, we have

$$\int_0^t |da(s)| = \sup\left\{\sum_{i=1}^{p} |a(t_i) - a(t_{i-1})|\right\},$$

where the sup is over subdivisions of $[0,t]$. More precisely, for any increasing sequence of subdivisions $0 = t_0^n < \cdots < t_{p_n}^n = t$ whose mesh size tends to 0, we have

$$\lim_{n\to\infty} \sum_{i=1}^{p} \left|a(t_i^n) - a(t_{i-1}^n)\right| = \int_0^t |da(s)|\,.$$

The integral on the left is just the total variation of the measure on the interval $[0,t]$. And the natural discrete approximation is as these finite Riemann sums. So the claim is that you can evaluate this as some sort of Riemann approximation procedure.

When we say 'increasing,' we mean in the sense of increasing sets; and the mesh size is the maximum difference between two consecutive points.

So this is basically saying you can find a value of your measure on some interval as the limit of sums of values of the measure on partitions of the interval, or something like that.

So far, nothing has been random. But eventually we're going to define a finite variation *process* to be a stochastic process whose sample paths are finite variation (and which is adapted); and when proving various things about measurability, it's useful to have this discrete approximation.

The proof of this is interesting in that it uses martingales — it's an analysis statement, but you can prove it by introducing probability.

Proof. Without loss of generality we can assume $t = T$. Let's first prove the easier direction of th equality, that $\int_0^t |da(s)|$ is greater than or equal to the sup. To see that, let's just evaluate this Riemann sum; by definition

$$\sum_{i=1}^{p} |a(t_i) - a(t_{i-1})| = \sum_{i=1}^{p} |\mu((t_{i-1}, t_i])| .$$

And we can take the absolute value inside, so we get that this is at most

$$\sum_{i=1}^{p} |\mu| ((t_{i-1}, t_i]).$$

And this is a partition of the interval, so this is equal to $|\mu| ((0, T])$. And by definition, this is just $\int_0^T |da(s)|$ (we're just taking $f = 1$ in the definition). So for *any* subdivision of the interval, you have this inequality.

Now we have to prove the other direction — we have to show that for any increasing sequence of subdivisions with mesh tending to 0, you actually have this convergence.

To do that, let's take some increasing sequence of subdivisions $(\{t_i^n\})_n$. Without loss of generality assume $|\mu| ((0, T]) \neq 0$ (if it's 0, then both sides have to be 0, so the statement is trivial).

Now let's define a probability space — let $\Omega = [0, T]$ and $\mathcal{F} = \mathcal{B}([0, T])$. And let's define a discrete-time filtration

$$\mathcal{F}_n = \sigma((t_{i-1}^n, t_i^n] \mid 1 \leq i \leq p_n)$$

(where p_n is the number of points in the nth subdivision), and $\mathcal{F}_\infty = \sigma(\mathcal{F}_n \mid n \geq 1)$. So we have this filtered probability space.

Claim 8.11 — We have $\mathcal{F}_\infty = \mathcal{B}([0, T]) = \mathcal{F}$.

As a couple of words, it's true because you assumed the mesh size tends to 0. So the point is for instance you want to be able to write any given interval $(s, t) \subseteq [0, T]$ as a countable union of intervals in the $\mathcal{F}_n$'s. But because your mesh size tends to 0, you can do this. And once you show (s, t) is in the generated σ-algebra, it has to be contained in the Borel one (which is contained in intervals of this form). So that's the idea — because the mesh tends to 0, you can write this interval (s, t) as a countable union of intervals you do have.

Now let's define the following random variable — we let $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$ be defined as the density of μ with respect to its total variation — so

$$X = \frac{d\mu}{d|\mu|}.$$

Recall that this density is a difference of indicators, so

$$X = \mathbf{1}_{D_+} - \mathbf{1}_{D_-}.$$

So this is certainly a measurable function on $(\Omega, \mathcal{F})$.

Now let's define a probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$ by

$$\mathbb{P}[A] = \frac{|\mu| (A)}{|\mu| ([0, T])}.$$

We have to divide for this to be a probability measure (so that Ω itself has size 1).

So we have all this setup now. The key computation to do, which will probably be on the next homework:

Exercise 8.12. We have

$$\mathbb{E}[X \mid \mathcal{F}_n] = \sum_{i=1}^{p_n} \frac{\mu((t_{i-1}^n, t_i^n])}{|\mu|\left((t_{i-1}^n, t_i^n]\right)} \mathbf{1}_{(t_{i-1}^n, t_i^n]}.$$

Basically, we're defining this Doob martingale that's going to be closed by X itself (in discrete time). We know that X is bounded, so it's integrable; so that's fine. Then by martingale convergence, we're going to get that $\mathbb{E}[X \mid \mathcal{F}_n]$ converges almost surely to X (we know X is $\mathcal{F}_\infty$-measurable — *a priori* Radon–Nikodym tells you X is $\mathcal{F}$-measurable, but $\mathcal{F} = \mathcal{F}_\infty$). And martingale convergence, when you unwrap it, is going to give exactly the convergence we want.

And what is this martingale going to be? It has to be a linear combination of indicator functions. Because we defined our filtration at finite times, it's discrete; so any measurable function with respect to $\mathcal{F}_n$ has to be a linear combination of these indicators. This means all we have to do is figure out the coefficients in this linear combination. And if you use the definition of conditional expectation, you'll see that the coefficients have to be as above. (The way you come up with this is by putting in general coefficients, and then solving using the definition of conditional expectation.)

Note that $\mu((t_{i-1}^n, t_i^n]) = a(t_i^n) - a(t_{i-1}^n)$ by definition.

So now by the uniformly integrable martingale convergence theorem, we get

$$\mathbb{E}[X \mid \mathcal{F}_n] \to \mathbb{E}[X \mid \mathcal{F}_\infty] = X \quad \text{in } L^1.$$

(The latter equality is because X is $\mathcal{F}_\infty$-measurable.)

The convergence in L^1 implies convergence of L^1 norms, so

$$\mathbb{E}[|\mathbb{E}[X \mid \mathcal{F}_n]|] \to \mathbb{E}[|X|].$$

(If you have variables converging in L^1, then their norms converge as real numbers; this is a consequence of the triangle inequality.)

But X is a difference between two indicator functions on disjointly supported sets, so you can show $\mathbb{E}[|X|] = 1$. (There's some argument you actually have to do — in principle if $D_+ \cup D_- = [0, T]$ then $|X|$ is always equal to 1. But the supports of your measures in principle don't have to partition the interval. But the point is that $|X| = 1$ $\mathbb{P}$-almost surely, where $\mathbb{P}$ is the measure you defined — in other words, basically $X = 1$ on the support of your probability measure.)

And now let's compute the left-hand side. Again, we just have a linear combination of indicators, so the expectation shouldn't be too hard to compute — we get

$$\mathbb{E}[|\mathbb{E}[X \mid \mathcal{F}_n]|] = \sum_{i=1}^{p_n} \frac{\left|a(t_i^n) - a(t_{i-1}^n)\right|}{|\mu|\left((t_{i-1}^n, t_i^n]\right)} \cdot \frac{|\mu|\left((t_{i-1}^n, t_i^n]\right)}{|\mu|\left([0, T]\right)}$$

(because we're taking the expectation of an indicator, which is just the probability of that interval,a nd you plug in the formula for $\mathbb{P}[A]$). Now we have this cancellation, and we're basically done — we get that this sum converges to 1, and if we move the denominator to the right-hand isde, we get

$$\sum \left|a(t_i^n) - a(t_{i-1}^n)\right| \to |\mu|\left([0, T]\right) = \int_0^t |da(s)|. \qquad \square$$

So that'll be one thing that's useful. Now let's prove another approximation result. (Here we're again assuming we have some FV function a.)

Proposition 8.13

Let $f : [0,T] \to \mathbb{R}$ be continuous, and let $(\{t_i^n\})_n$ be a sequence of subdivisions of $[0,T]$ with mesh size tending to 0. Then we can write

$$\int_0^T f(s)\, da(s) = \lim_{n\to\infty} \sum_{i=1}^{p_n} f(t_{i-1}^n)(a(t_i^n) - a(t_{i-1}^n)).$$

This is kind of the first thing you'd guess as how to define an integral — as the limit of these types of approximations (a sort of Riemann approximation). And indeed it's true.

Proof. First, f is continuous, so it's bounded on any finite interval, which means it's definitely measurable with respect to $|\mu|$; so the left-hand side is well-defined.

Now let's approximate f as a sum of indicators — so we let

$$f_n = \sum_{i=1}^{p_n} f(t_{i-1}^n)\mathbf{1}_{(t_{i-1}^n, t_i^n]}.$$

The idea is that you kind of want to view the sum on the right-hand side as the integral against da of some linear combination of indicators; and it's basically going to be the above.

Before we see that, first note that $f_n \to f$ by continuity (and the fact that the mesh size tends to 0). We also have that

$$\sup_n \|f_n\|_{L^\infty} \le \|f\|_{L^\infty} < \infty$$

(where this denotes the sup norm on the finite interval $[0,T]$). To finish, the right-hand side of the lemma statement can be written as

$$\text{RHS} = \int_{[0,T]} f_n(s)\, \mu(d_s).$$

We basically want to apply some convergence theorem; it's more convenient with a positive rather than signed measure, so we can write this as

$$\int_{[0,T]} f_n(s)\mu_+(ds) - \int_{[0,T]} f_n(s)\mu_-(ds).$$

Now we can apply the bounded convergence theorem to each piece (they're bounded by the above argument) to get that this converges to $f\, d\mu$, which is precisely the left-hand side. □

Now, if you have a function over *all* times:

Definition 8.14. We say a function $a\colon \mathbb{R}_+ \to \mathbb{R}$ is *finite variation* if the restriction of a to any finite interval has finite variation.

Remark 8.15. If a is finite variation on $\mathbb{R}_+$, then there exists a unique σ-finite positive measure μ on $\mathbb{R}_+$ (σ-finite means that you can write $\mathbb{R}_+$ as an increasing union of bounded sets, and your measure is finite on all of them) whose restriction to any finite interval $[0,T]$ is the total variation of the function $a|_{[0,T]}$.

Definition 8.16. For any positive measurable function f, we define

$$\int_0^\infty f(s)\, |da(s)| = \int f\, d\mu$$

(where the integral is over the entire interval).

Because f is nonnegative, this is always defined, though it might be ∞.

Definition 8.17. For integrable f (i.e., f such that $\int_0^\infty |f(s)|\,|da(s)| < \infty$), we define

$$\int_0^\infty f(s)\,da(s) = \lim_{T\to\infty} \int_0^T f(s)\,da(s).$$

One can check (by e.g., dominated convergence) that this limit exists, and basically equals $\int f\,d\nu$.

§8.3 Finite variation processes

Now we can get onto finite variation processes — so far, we've just seen some deterministic facts.

Throughout, let's fix a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{P})$.

Definition 8.18. An adapted process (A_t) is a *finite variation process* if it has finite variation sample paths on $\mathbb{R}_+$. If additionally the sample paths are nondecreasing, we say that A is a *increasing process*.

As a consequence, A has continuous sample paths and $A_0 = 0$ (because any finite variation process is continuous and is 0 at $t = 0$).

Definition 8.19. If (A_t) is a finite variation process, we define its *total variation process* as

$$V_t = \int_0^t |dA_s|\,.$$

Then (V_t) is going to be an increasing process. Why is (V_t) adapted? (The rest is certainly true — it has finite variation increasing sample paths.) This follows from the theorem that we can write

$$\int_0^t |dA_s| = \lim_{n\to\infty} \sum_{i=1}^{p_n} \left|A(t_i^n) - A(t_{i-1}^n)\right|$$

as a limit of discrete approximations. Each of these discrete approximations is certainly $\mathcal{F}_t$-measurable (because the t_i^n form a subdivision of $[0,t]$), and the limit of $\mathcal{F}_t$-measurable functions is $\mathcal{F}_t$-masurable.

Remark 8.20. Any finite variation process can be written as a difference of two increasing ones (similarly to fininte variation functions).

Proposition 8.21

Let A be a finite variation process, and let H be a progressive process such that

$$\int_0^t |H_s(\omega)|\,|dA_s(\omega)| < \infty \quad \text{for all } t \geq 0 \text{ and } \omega \in \Omega.$$

Then the process $H \cdot A$ given by

$$(H \cdot A)_t = \int_0^t H_s\,dA_s$$

is also a finite variation process.

This is kind of a special case of stochastic integrals — where you're integrating against a finite variation process. Here A is going to be our integrator (since we've set up all this notion of integrating against A),

and our integrand is going to be a progressive process. We need to make some integrability assumptions so that this is well-defined.

We usually omit ω in our notation, but here we're writing this to emphasize that it's true for all $\omega \in \Omega$.

This is not really the main case of stochastic integration — the main case is when you integrate against things like Brownian motion or martingales — but to be able to integrate semimartingales, you'll need to be able to integrate against both Brownian motion and finite variation processes. So we're doing this case first.

Proof. The fact that sample paths are finite variation functions is a remark we already made — if μ is the measure associated to A (so we fix a given $\omega \in \Omega$, so now we have a finite variation function and this associated measure), then this integral has associated measure $H\,d\mu$. (We basically wrote this remark earlier — when you integrate against nice enough functions, at least integrable ones, you get another finite variation function.)

The main point is you have to show this thing is adapted. The fact that it has finite variation sample paths is basically immediate (there's no probability theory involved, just analysis), but you need to show it's actually an adapted process. In other words, we need to show $(H \cdot A)_{t_0} \in \mathcal{F}_{t_0}$.

To show this, it suffices to show that given a measurable function $h : \Omega \times ([0,t], \mathcal{F}_{t_0} \otimes \mathcal{B}([0,t_0])) \to (\mathbb{R}, \mathcal{B})$ such that $\int_0^{t_0} |h(\omega,s)|\,|dA_s(\omega)| < \infty$ for all ω, we have that the random variable

$$\int_0^{t_0} h(\omega,s)\,dA_s(\omega)$$

(the integral without absolute values) is $\mathcal{F}_{t_0}$-measurable. Why does this suffice? If you recall the definition of a progressive process, by definition a progressive process is measurable in the form of h above. So now it suffices to show the claim is true for any measurable function h like this; and then we'll apply it to this progressive process. (And the second assumption will be true for our H because that's part of the assumption of the proposition.)

How do we show this? Any time you want to show some claim like this, you start with the simplest possible thing. What are measurable functions with respect to $\mathcal{F}_{t_0} \otimes \mathcal{B}([0,t_0])$? They're basically going to be combinations of indicator functions of events in this product σ-algebra, and the simplest ones of those are product events.

So let's first consider an event of the form $E \times (s_0, s_1]$, where $E \in \mathcal{F}_{t_0}$ and $(s_0,s_1] \subseteq [0,t_0]$. And let h be the indicator of the product event, i.e.,

$$h_t(\omega) = \mathbf{1}_E(\omega)\mathbf{1}_{(s_0,s_1]}(t).$$

Then h is bounded by 1, so the integrability assumption is certainly okay. And in this case, we have

$$\int_0^{t_0} h_s\,dA_s = \mathbf{1}_E(A_{s_1} - A_{s_0})$$

(here we're no longer writing the ω's) — you're more or less computing the integral of an indicator. The point is we're integrating with respect ot s, so the $\mathbf{1}_E$ doesn't matter and we can pull it out; then we're integrating the indicator of an interval, and that gives you the difference of endpoints.

And $\mathbf{1}_E \in \mathcal{F}_{t_0}$, and A is adapted and $s_1, s_0 \le t_0$; so this is all in $\mathcal{F}_{t_0}$.

So basically, we proved the claim for the simplest possible measurable functions we could think of. Now, as always, you have to find the right way to make this approximation argument work. For instance, let's let $\mathcal{G} \subseteq \mathcal{F}_{t_0} \otimes \mathcal{B}([0,t_0])$ be the collection of $F \in \mathcal{F}_{t_0} \otimes \mathcal{B}([0,t_0])$ such that the map

$$\omega \mapsto \int_0^{t_0} \mathbf{1}_F(\omega,s)\,dA_s(\omega)$$

is $\mathcal{F}_{t_0}$-measurable. So it's the collection of events for which you get a random variable which is $\mathcal{F}_{t_0}$-measurable.

We already showed that $\mathcal{G}$ contains all product events of the form $E \times (s_0, s_1]$ with $E \in \mathcal{F}_{t_0}$ and $(s_0, s_1] \in [0, t_0]$ (that was the first part). Now you want to use some sort of monotone class argument to show that $\mathcal{G}$ has to contain the generated σ-algebra.

This is going to be on the next homework; it's an exercise. Le Gall points you to the right monotone class theorem to use. But basically, you want to show that in fact

$$\mathcal{G} = \mathcal{F}_{t_0} \otimes \mathcal{B}([0, t_0])$$

via some monotone class argument.

And if we now know this, then we know the indicator function of anything in this product σ-algebra gives me a $\mathcal{F}_{t_0}$-measurable function, then we're basically done — because you can approximate any general h as a linear combination of indicators.

So given a general function $h \geq 0$, we can approximate h by a linear combination of indicators $h_n \nearrow h$. (We're not going to write this out, but it's some usual approximation thing where you split based on the values of h — whether it takes values in some interval of length $1/n$ — and you take the smaller point in that interval, so you increase up to h.) Then the monotone convergence theorem tells you that

$$\int_0^{t_0} h_n \, dA_s \to \int_0^{t_0} h \, dA_s.$$

And each h_n is a linear combination of indicators, so these are measurable by the above claim; and the limit of measurable things is measurable. So that tells you the right-hand side is also $\mathcal{F}_{t_0}$-measurable.

This deals with nonnegative h; for general h you can write $h = h^+ - h^-$ and apply this to both the positive and negative parts. □

Remark 8.22. Often, we only have a weaker condition that *almost surely*, for all $t \geq 0$, your process H is integrable up to time t, meaning that

$$\int_0^t |H_s(\omega)| \, |dA_s(\omega)| < \infty.$$

The proposition assumed it's true for all ω, but in principle sometimes you'll only have it for *almost* all ω.

Here, it's useful to have a *complete* filtration. Assuming $(\mathcal{F}_t)$ is complete, you can define a modified process H' by

$$H'_t(\omega) = \mathbf{1}_E H_t(\omega),$$

where you basically modify it off the 'good event'

$$E = \left\{ \int_0^n |H_s(\omega)| \, |dA_s(\omega)| < \infty \text{ for all } n \right\}.$$

The 'almost sure' assumption says $\mathbb{P}[E] = 1$; and the point is we're just setting H to 0 off this good event.

And H' is still progressive — because we started with a progressive process H and modified it on a measure-0 event, and the fact that the filtration is complete lets you do that.

And now we can define

$$H \cdot A = H' \cdot A,$$

because H' is literally integrable for all ω by definition.

§9 March 3, 2025

§9.1 Review

To quickly recap what we did last time, we introduced finite variation processes, which we usually denote by A. The main result from last time was that given a FV process A and a suitable integrand H (which is progressive and satisfies an integrability condition), you can define

$$(H \cdot A)_t = \int_0^t H_s \, dA_s,$$

and this is also a FV process. This didn't really require much probability theory at all — you're just integrating an integrable function with respect to a finite measure. The main thing in the proof was the fact that this actually gives you an adapted process; so there was some approximation argument we did last time to show why that was the case.

As a heuristic remark, another reason this has to be adapted is you can kind of approximate this integral — or you would expect to be able to approximate it — by some sort of Riemann sum approximation

$$\sum_{i=1}^{n} H_{t_i^n}(A_{t_i^n} - A_{t_{i-1}^n}).$$

And in the case where H is continuous, we showed that when you take the mesh to 0, you actually recover this integral. We didn't show it when H is just a progressive process. But you would believe that under some nice assumptions, you *should* be able to approximate the integral as a limit of this finite sum. And this finite sum is certainly going to be $\mathcal{F}_t$-measurable. (But the proof in the book was by this monotone class argument; but this is maybe the heuristic way to see things.)

Another remark we'll stress is that basically you should think of FV processes as varying like dt, i.e.,

$$dA_t \sim dt.$$

This is almost the definition of a FV process — we defined it based on a measure, but this is how you should think about it. What this means is that $A_{t+\varepsilon} - A_\varepsilon \sim \varepsilon$ — the variation on a small interval looks like the interval itself. In particular Brownian motion is not a FV process — heuristically we have $B_{t+\varepsilon} - B_t \sim \varepsilon^{1/2}$ instead. Basically you think of FV processes as Lipschitz functions, for instance — any Lipschitz function definitely satisfies this condition. So that's the heuristic picture of a FV process — your variation over a time interval is like dt.

Remark 9.1. If H and K are progressive such that $\int_0^t |H_s| \, |dA_s| < \infty$ and $\int_0^t |H_s K_s| \, |dA_s| < \infty$ for all t, then you get the associativity property that

$$K \cdot (H \cdot A) = (KH) \cdot A.$$

So you have this associativity of integrals. (The integrability assumptions are just so that everything is well-defined.) This is an example of the associativity property of stochastic integrals, as we'll see later, in the special case where you just have FV processes.

The proof is basically that if you define a density $k(h \, d\mu)$ — so you have some base measure μ (which you think of as A) and two densities k and h (which are integrable functions), then you can first define this new measure $h \, d\mu$, and then you can find this new measure with respect to $h \, d\mu$. And this is the same as taking the density $kh \, d\mu$. So that's basically all this is saying.

§9.2 Continuous local martingales

That concludes our discussion of FV processes; now we'll move to the other part of the discussion of semimartingales, which is continuous local martingales.

For some setup, throughout this discussion we fix a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{P})$.

§9.2.1 Heuristics

For some heuristic remarks, you should basically think of a continuous local martingale M — first, it's some generalization of a martingale. You should think of them as *locally* like Brownian motion — in particular, you should have $dM_t \sim d\sqrt{t}$, whatever this means, in the sense that at a heuristic level,

$$M_{t+\varepsilon} - M_t \sim \varepsilon^{1/2}.$$

This is one fundamental property of continuous local martingales, and it's quite different from FV processes. So basically you can't be both at the same time (unless you're just 0), and we're going to prove that soon.

You should have this picture where in a FV process you might have some continuously differentiable function, which is basically the drift of your process. And then around this function, you have some small fluctuations; and that's what this local martingale part is doing. We'll see you can decompose a semimartingale into these two parts — the drift and variance.

Note that when ε is small, $\varepsilon^{1/2}$ is much larger than ε — so continuous local martingales vary much more wildly (at small levels) than FV processes.

§9.2.2 Definition

Notation 9.2. If T is a stopping time and (X_t) is a continuous stochastic process (i.e., a stochastic process with continuous sample paths), we write

$$X_t^T = X_{t \wedge T}.$$

We call X^T the *stopped process*.

This is because we're only looking up to time T, and after that it's constant.

We proved before that this is actually a random variable; and if X is adapted, then by continuity it's progresive, so this thing will actually be adapted as well.

Fact 9.3 — If S is a stopping time, then $(X^S)^T = X^{S \wedge T} = (X^T)^S$.

This is just from plugging things right into the definition.

Now we'll define continuous local martingales. The point is that the L^1 condition — that your process be integrable at all times — is kind of restrictive. So you kind of want to remove it but still have some nice properties about your process, and that's the motivation.

Definition 9.4. An adapted process (M_t) with continuous sample paths and such that $M_0 = 0$ almost surely is a *continuous local martingale* if there exists a sequence of stopping times (T_n) such that $T_n \nearrow \infty$ and for each n, the stopped process M^{T_n} is a uniformly integrable martingale.

More generally, we say that M is a *continuous local martingale* if it's an adapted continuous process and $N_t = M_t - M_0$ is a continuous local martingale.

By $T_n \nearrow \infty$ we mean $T_n \le T_{n+1} \le T_{n+2} \le \cdots$, and they converge to ∞ (for every $\omega \in \Omega$).

Definition 9.5. We say that a sequence of stopping times (T_n) *reduces* M if $T_n \nearrow \infty$ and for all n, the stopped process M^{T_n} is a uniformly integrable martingale.

In other words, this condition of a continuous local martingale means you have a sequence of increasing stopping times going to ∞ which reduces M.

Remark 9.6. Note that we don't require $M_t \in L^1$. The point is we want to loosen the martingale condition to remove this restriction.

Remark 9.7. Any continuous martingale is a continuous local martingale — you can just take the sequence of stopping times $T_n = n$. In the martingale module, we proved that if you stop your process at a deterministic time, then $(M_{t\wedge n})$ is always uniformly integrable. And as a consequence, if you start with a continuous martingale, then $T_n = n$ reduces your martingale.

The proof of this was that $(M_{t\wedge n})$ is closed by the endpoint of your interval — we have

$$\mathbb{E}[M_n \mid \mathcal{F}_t] = M_{n\wedge t}.$$

However, the converse is not true — otherwise what would be the point of this definition? To see that, you can consider (for instance) $Z + B_t$, where B_t is a Brownian motion and $Z \notin L^1$ (and Z is independent of (B_t)). This cannot be a martingale because this thing is not in L^1 (it starts at Z because $B_0 = 0$). But one can check that you can actually find a sequence of stopping times (T_n) which reduces this, so it's actually a continuous local martingale.

This is a somewhat contrived example, but later we'll see some more natural examples — it might come out of looking at $1/|B_{t+x}|$, where you have a 3-dimensional Brownian motion (we add some fixed time x because we don't want to have ∞ at time 0). It'll probably be an exercise (but you have to use Ito's formula for it) that this is a continuous local martingale, but it's not going to be a martingale. But that's for later.

§9.3 Some properties of continuous local martingales

We'll now talk about some properties of continuous local martingales. We already saw one property — any continuous martingale is also a continuous local martingale.

Proposition 9.8

In the definition of a continuous local martingale, we can replace 'uniformly integrable martingale' with 'martingale.'

This says you don't need 'uniformly integrable' in the definition of a continuous local martingale — it suffices to display some sequence of times for which the stopped process is just a martingale. Indeed, if you look at other sources than Le Gall, often they'll just define it with martingale instead of UI martingale. The proof is that if (M^{T_n}) is a martingale, then $(M^{T_n\wedge n})$ is still a martingale, and it's going to be uniformly integrable for the same reason as above. (It's also a consequence of something we proved last week — $T_n \wedge n$ is a bounded stopping time.) (The other part is that $T_n \nearrow \infty$, so $T_n \wedge n \nearrow \infty$ as well.)

Proposition 9.9

If M is a continuous local martingale and T is a stopping time, then the stopped process M^T is also a continuous local martingale.

The reason why you have this is that if (M^{T_n}) is a UI martingale, then so is $(M^{T\wedge T_n})$ — again, this was a consequence of some proposition we proved last time. And $M^{T\wedge T_n} = (M^T)^{T_n}$ (by the commutation property we mentioned earlier). So if (T_n) reduces M, it also reduces the stopped process M^T.

Proposition 9.10

If a sequence of stopping times (T_n) reduces M and (S_n) is a sequence of stopping times with $S_n \nearrow \infty$, then $(T_n \wedge S_n)$ also reduces M.

For this, we have that $T_n \wedge S_n \nearrow \infty$, so we just need to check why you get a UI martingale. And the point is that as before, if (M^{T_n}) is a UI martingale, then by the optional stopping theorem you can stop it at another stopping time S_n and still get a UI martingale.

Proposition 9.11

The space of continuous local martingales is a vector space.

The reason this is true is basically a consequence of Proposition 9.10 — if you have two continuous local martingales M and M' reduced by (T_n) and (T_n'), you need to find a sequence of stopping times reducing $M + M'$. And the natural sequence of stopping times is $(T_n \min T_n')$. That reduces both M and M' by Proposition 9.10, so it also reduces $M + M'$ (if you add two UI martingales, it's still a UI martingale).

Now let's get into some results about continuous local martingales.

Proposition 9.12

(i) A nonnegative continuous local martingale such that $M_0 \in L^1$ is actually a supermartingale.

(ii) A continuous local martingale M such that there exists $Z \in L^1$ which dominates the entire process (meaning $|M_t| \leq Z$ for all t) is a uniformly integrable martingale.

(iii) If M is a continuous local martingale with $M_0 = 0$, or more generally $M_0 \in L^1$, then the sequence of stopping times

$$T_n = \inf\{t \geq 0 \mid |M_t| \geq n\}$$

reduces M.

In particular, (i) says that M_t is integrable for all t. We're not assuming *a priori* that's the case, but actually it will be the case if we assume $M_0 \in L^1$. (The nonnegativity assumption is important for this.)

For (ii), maybe the most common case we'll apply it is where Z is just constant. In particular, a *bounded* continuous local martingale is actually a uniformly integrable martingale.

For (iii), you're just looking at the first time your martingale goes above n; the claim is that this reduces M.

Proof of (i). Let (T_n) be a sequence of stopping times reducing M. Then for all times $s \leq t$ and all n, we have the martingale property

$$\mathbb{E}[M_t^{T_n} \mid \mathcal{F}_s] = M_s^{T_n}.$$

As a consequence, in particular, we have

$$\mathbb{E}[M_t^{T_n}] = \mathbb{E}[M_0^{T_n}] = \mathbb{E}[M_0]$$

(because you can apply this with $s = 0$).

Now we use the fact that $T_n \nearrow \infty$; so for any fixed t, as you send $n \to \infty$, we have $M_t^{T_n} \to M_t$; and similarly we have $M_s^{T_n} \to M_s$.

We want to apply Fatou's lemma. So first we want to say why this thing is integrable — we claim it's a supermartingale, so why is it integrable? We have $M_t = \liminf M_t^{T_n}$ (because it's actually just the limit); Fatou's allows us to take the lim inf outside and say

$$\mathbb{E}[M_t] \leq \liminf_{n\to\infty} \mathbb{E}[M_t^{T_n}].$$

(Here's where we use the nonnegativity assumption, which is required to apply Fatou.) And this is equal to

$$\liminf_{n\to\infty} \mathbb{E}[M_0^{T_n}] = \mathbb{E}[M_0].$$

This shows M is integrable for all t (because M is nonnegative, and we've shown its expectation is finite). So $M_t \in L^1$.

Now we need to show the supermartingale property. So we start with the martingale property at finite n and also apply Fatou's, but *conditional* Fatou's. First, we have to say M_t is integrable in order to say $\mathbb{E}[M_t \mid \mathcal{F}_s]$ is well-defined (actually this is not so important, because M_t is nonnegative); we've done this, so we can apply conditional Fatou's to get that

$$\mathbb{E}[M_t \mid \mathcal{F}_s] \leq \liminf_{n} \mathbb{E}[M_t^{T_n} \mid \mathcal{F}_s].$$

And now we can use the martingale property to say this is equal to

$$\liminf_{n} M_s^{T_n}.$$

But we have $\lim_n M_s^{T_n} = M_s$; so this shows the supermartingale property. □

Proof of (ii). Part (ii) says that if we have this domination condition, then M is actually a UI martingale. The fact that it's UI just follows — any collection of random variables satisfying this uniform domination condition is UI (you can verify that directly from the definition). So the main thing is to show why it's a martingale.

So let (T_n) reduce M. We had the martingale property

$$\mathbb{E}[M_t^{T_n} \mid \mathcal{F}_s] = M_s^{T_n}.$$

And what we want to do is take limits on both sides. As a consequence of the assumption, we have $M_t, M_s \in L^1$, which means the conditional expectation after taking the limit is well-defined. And you just want to apply the conditional dominated convergence theorem — the right-hand side is fine because it converges almost surely (or even surely) to M_s. And on the left-hand side, you apply conditional dominated convergence to show this converges to $\mathbb{E}[M_t \mid \mathcal{F}_s]$. So we get

$$\mathbb{E}[M_t \mid \mathcal{F}_s] = M_s,$$

which is precisely the martingale property. □

Proof of (iii). First, let's say why this process is uniformly integrable. By the definition of T^n, you might think that $|M_t^{T_n}| \leq n$. But there's also the case where you have to use the initial starting value instead (this is if n is smaller than your initial value, because then t is just 0); so we actually have

$$|M_t^{T_n}| \leq n + |M_0|.$$

And now you have that $(M_t^{T_n})$ is UI, because you've dominated it by an integrable random variable. And one can check that $T_n \nearrow \infty$ (it's clearly monotone in n, and it increases to ∞ because M_t is continuous — this is just some fact about continuous functions). □

Remark 9.13. As one word of caution, we saw that the uniform domination condition in (ii) suffices to say a continuous local martingale is actually a UI martingale. And part of the proof was saying that this condition implies your process is uniformly integrable.

There's other conditions that also imply your process is uniformly integrable. In particular, a continuous local martingale which is itself uniformly integrable — or even more, bounded in L^p for $p > 1$ (one can use Hölder's inequality to show that if you're L^p-bounded for $p > 1$, you're UI) — is actually not necessarily a martingale. This is quite counterintuitive. You might think you just need this condition $|M_t| \leq Z$ to say you're UI, and then you have to be a martingale. But that's not actually the case — your process itself can be UI, but not a martingale.

It's hard to give intuition for why this is true, but the example we'll see later is $1/|B_t + x|$ — basically the inverse distance of a 3D Brownian motion. We'll see that this is a continuous local martingale but not a martingale (as mentioned earlier), but it's actually L^p bounded for all p.

§9.4 Continuous local martingales vs. FV processes

Now let's move on to proving another thing about continuous local martingales. This theorem is basically what we said near the beginning — you cannot be both a continuous local martingale and a FV process — the conditions are incompatible with each other.

Theorem 9.14

Let M be a continuous local martingale and a finite variation process. Then M is indistinguishable from 0.

In other words, almost surely, for all times t we have $M_t = 0$. We basically don't care about processes up to indistinguishability.

The intuitive reason why this is the case is that on one hand a finite variation process varies like $dA_t \sim dt$, but you should think of a continuous local martingale as $dM_t \sim d\sqrt{t}$. These are very different conditions, so they can't both be satisfied.

Proof. In some sense, we already saw this proof for Brownian motion. Define the stopping times

$$\tau_n = \inf\left\{t \geq 0 \;\middle|\; \int_0^t |dM_s| \geq n\right\}.$$

(Note that M starts at 0 as a consequence of it being a FV process — all FV processes start at 0. And given a FV process, you can consider this total variation process $\int_0^t |dM_s|$; it's going to be adapted and all that so τ_n is actually a stopping time.) And $\tau_n \nearrow \infty$.

Also note that the stopped process $M_t^{\tau_n}$ satisfies

$$|M_t^{\tau_n}| = |M_{t\wedge\tau_n} - M_{0\wedge\tau_n}| \leq \int_0^{t\wedge\tau_n} |dM_s|$$

(we proved this total variation is the sup over all partitions of your interval of the sum of difference increments; and this is a partition of the interval, where we just take the two endpoints). And by the definition of τ_n, this is at most n.

So the sequence (τ_n) reduces M — because this stopped process M^{τ_n} is bounded and it's a continuous local martingale, so it's actually a UI martingale. In other words, we've shown (M^{τ_n}) is a UI martingale, which means (τ_n) reduces M.

Now consider a subdivision of our interval $[0,t]$ — say we fix a time t. We want to show M_t is 0 almost surely — that's what we're working towards. So we fix a subdivision

$$0 = t_0^k < t_1^k < \cdots < t_{p_k}^k = t.$$

(These subdivisions don't have to be nested for this proof, but they do need to have mesh size tending to 0 as $k \to \infty$.)

When you have a bounded martingale (or more generally, a L^2 martingale), you want to look at the second moment — because it has this nice decoupling property where the increments are uncorrelated. So you can write

$$\mathbb{E}[(M_t^{\tau_n})^2] = \mathbb{E}\left[\sum_{i=1}^{p_k} (M_{t_i^k}^{\tau_n} - M_{t_{i-1}^k}^{\tau_n})^2\right]$$

(we showed this last week, even at the level of conditional expectations — it just follows because different increments are decorrelated). And now you can bound this by the sup and sum — you can basically bound L^2 by L^1 times L^∞ — and you get that this is at most

$$\mathbb{E}\left[\sup_{1\le i\le p_k} |M_{t_i^k}^{\tau_n} - M_{t_{i-1}^k}^{\tau_n}| \left|\sum_{i=1}^{p_k} \left|M_{t_i^k}^{\tau_n} - M_{t_{i-1}^k}^{\tau_n}\right|\right|\right].$$

(We did something similar to this with Brownian motion, where we saw that this thing converged to t; the point is that continuous local martingales are supposed to be kind of similar to Brownian motion, so we're basically following the same proof.)

Now we can further bound the above. The sum of the increments is also bounded by n — because it's bounded by the total variation on your time interval. So we get that the right-hand side is bounded by

$$n \cdot \mathbb{E}\left[\sup_{1\le i\le p_k} \left|M_{t_i^k}^{\tau_n} - M_{t_{i-1}^k}^{\tau_n}\right|\right].$$

(This is by what we proved last time, that the total variation is the sup over all subdivisions of the sum of increments, and by how we defined our stopping time.)

And this goes to 0 as $k \to \infty$. Why? The process M is bounded, and this is bounded by $2M$. And M is continuous and the gap inside is going to 0. So you can use dominated convergence to say that the expectation goes to 0.

So what have we shown? The left-hand side doesn't see k at all — we just introduced this parameter k tracking our subdivision, but the left-hand side was independent of k. So we've shown that

$$\mathbb{E}[(M_t^{\tau_n})^2] = 0.$$

We wanted to show this for M_t, not the stopped process; so we can just apply Fatou and get

$$\mathbb{E}[M_t^2] \le 0$$

(because $M_t^2 = \liminf M_t^{\tau_n}$ — it's actually the limit — and you can pull the $\liminf$ outside). And that implies $M_t = 0$ almost surely.

This is true for any fixed t. But M is continuous, so you can take a countable dense sequence of times; and by continuity you're going to get that M is indistinguishable from 0 (if you're 0 at a countable dense sequence of times and you're continuous, then you're actually 0 everywhere). □

Student Question. *Where did we use that M was a FV process?*

Answer. Otherwise τ_n might always be 0. Here we used the fact that $\tau_n \nearrow \infty$; and this comes from the fact that the total variation process is a continuous function, so in particular it's finite.

It's kind of like — going back to this other property of Brownian motion, where we showed that this sum of absolute increments is *infinite* almost surely. And we actually showed this by starting from the fact that the sum of squared increments is basically t. And these two things are incompatible. At the infinitesimal level, it's basically saying this — imagine I break my interval into pieces of $1/n$. For a FV process, we have $\sum_{i=1}^n 1/n = 1$. And that means $\sum_{i=1}^n (1/\sqrt{n})^2 = 1$. But that has to mean $\sum_{i=1}^n 1/\sqrt{n} = \sqrt{n} \to \infty$. That's the basic reason for why for Brownian motion, this sum of absolute increments has to be infinite (if you take a finer and finer mesh size).

§9.5 Quadratic variation

Now let's get to defining probably the most important thing about continuous local martingales. This is precisely this type of thing — the sum of squared increments. As we said, for continuous local martingales we expect the sum of squared increments to be $O(1)$ as we take the mesh size tending to 0 — so that thing is going to be called the quadratic variation.

From now on, we assume $(\mathcal{F}_t)$ is complete — that means $\mathcal{F}_0$ contains all null sets. This completeness assumption is needed so that certain processes you define later on are still adapted (you want to be able to modify processes on measure 0 sets, and when you modify a process you need to know that measure 0 set is measurable, or adapted; and this completeness says it's $\mathcal{F}_0$-measurable, so it's also $\mathcal{F}_t$-measurable).

In the notes, this is highlighted as the most important theorem of the chapter. It's also probably going to be the most difficult or involved proof in this entire course. It takes 30 pages in the notes, because you have to do this long computation.

Theorem 9.15

Let M be a continuous local martingale. Then there exists an increasing process $(\langle M, M\rangle_t)$, unique up to indistinguishability, such that $M_t^2 - \langle M, M\rangle_t$ is a continuous local martingale. Moreover, for any $t > 0$ and a sequence of increasing subdivisions $0 = t_0^n < \cdots < t_{p_n}^n = t$ with mesh size tending to 0, we have

$$\langle M, M\rangle_t = \lim_{n\to\infty} \sum_{i=1}^{p_n} (M_{t_i^n} - M_{t_{i-1}}^2).$$

We call $\langle M, M\rangle$ the *quadratic variation* of M.

Recall we defined an increasing process as a FV process whose sample paths are nondecreasing.

§9.5.1 Some big-picture stuff

The fact that you have nontrivial quadratic variation — this is in principle not just 0 — is encoding the fact that you vary like $\sqrt{t}$ on small intervals. If this were 0, then you're actually going to get that M is 0 almost surely — we'll probably see that later. For instance, if $\langle M, M\rangle = 0$, which says you don't really vary like $\sqrt{t}$ on small intervals, then M_t^2 is a continuous local martingale, and it's nonnegative. And we have this result that a nonnegative continuous martingale starting in L^1 is a supermartingale. So we can apply that here; then you have a supermartingale starting at 0 which is nonnegative, and that implies it's 0.

Then you might ask, what is this martingale $M_t^2 - \langle M, M\rangle$ supposed to be? Do I have a representation of it? That's exactly stochastic integration — in particular, this is going to be a special case of Ito's formula

— that whatever this difference is, you can write it as a stochastic integral. In fact, one consequence of Ito's formula is that you can write

$$M_t^2 = \int_0^t M_s \, dM_s + \langle M, M \rangle_t,$$

whatever this means (when we talk about stochastic integration, we'll define what the integral term means).

So this is sort of the first instance of Ito's formula we're seeing. Why is it all going to be true? It kind of comes from the fact that if you have a (local) martingale, oftentimes what you want to do is take some continuous function f of your martingale, e.g., the square, and you want to take derivatives — so you want to look at

$$\frac{d}{dt} f(a(t)).$$

Classically in calculus, we know that if a is a continuously differentiable function of t and we apply some differentiable function f to it, the chain rule tells me

$$\frac{d}{dt} f(a(t)) = f'(a(t))a'(t).$$

If you try applying the same logic to looking at the time-derivative of a continuous local martingale, you might think that

$$\frac{d}{dt} M_t^2 = 2M_t \, dM_t$$

(taking $f(x) = x^2$, whose derivative is $2x$ — whatever dM_t means, we'll talk about it later). But this is actually not true. That's what's new about Ito's formula, and what sets it apart from regular calculus. The reason why it's not true, we can basically already see from the statement of this theorem. Because what does the derivative of M_t^2 mean? It means I take a small increment and try to see what it looks like, so I consier $M_{t+\varepsilon}^2 - M_t^2$. And I can write this as

$$(M_t + (M_{t+\varepsilon} - M_t))^2 - M_t^2.$$

Basically I want to do Taylor expansion on a quadratic function, so let's see if I can do that. When you look at a quadratic function and take this difference, in principle you get a cross-term $2M_t(M_{t+\varepsilon} - M_t)$. But then you also get this square term — you actually get

$$2M_t(M_{t+\varepsilon} - M_t) + (M_{t+\varepsilon} - M_t)^2.$$

And the whole point is in classical calculus, if $M_{t+\varepsilon} - M_t \sim \varepsilon$ (e.g., if you had a FV process instead of a martingale), then when we take the square, you're really summing $\sum^{\varepsilon^{-1}} \varepsilon^2 = \varepsilon$. So we're basically proving the chain rule for this quadratic function, if a is differentiable — you only need to expand to first order, because the second order terms are negligible.

But this is no longer true if you have a martingale — $M_{t+\varepsilon} - M_t$ varies like $\sqrt{\varepsilon}$, so when you square it, it varies like ε^2. And when you sum up ε^{-1} terms of order ε, you get $O(1)$.

So that's why in stochastic calculus you get this correction term — because you're not actually finite variation, you're like a martingale. And this means you get the correction term $\langle M, M \rangle_t$ in the chain rule.

This is the first instance of that.

§9.5.2 Proof of uniqueness

We'll start the proof; we definitely won't finish today, but we'll see how far we'll get.

First let's talk about uniqueness. That's going to be an easy consequence of what we showed before — that you can't be both FV and a continuous local martingale. So if A^1 and $A^{(2)}$ are both such that

$$M_t^2 - A_t^{(1)} \text{ and } M_t^2 - A_t^{(2)}$$

are both continuous local martingales — I can add continuous local martingales and still get a continuous local martingale, so that means

$$A_t^{(1)} - A_t^{(2)} = (M_t^2 - A_t^{(2)}) - (M_t^2 - A_t^{(1)})$$

is a continuous local martingale. But it's also a FV process, because sums and differences of FV processes are also FV processes. This means $A^{(1)} - A^{(2)} = 0$ up to indistinguishability.

So uniqueness is basically immediate; it's really about constructing such a process.

§9.5.3 Proof of existence for bounded M

Whenever you try to prove a theorem like this, you first make a bunch of simplifying assumptions — because if you can't prove the simpler case, how can you prove the general case? So we'll assume $M_0 = 0$ and that M is bounded — there is some constant a such that M is bounded uniformly in time by a. (The reason we'll be able to remove these assumptions is we have a continuous local martingale, so you can always take some sequence of stopping times where your process is bounded; so we'll start with this simpler situation where your process is bounded to begin with.)

Now fix a time $K > 0$ and an increasing sequence of subdivisions $0 = t_0^n < \cdots < t_{p_n}^n = K$ with mesh size tending to 0.

Where's all this going to come from? We always just discretize and see what happens in discrete time, so that's what we're going to do. We can write our martingale M_t — for $t \in [0, K]$ — as a sum of telescoping increments

$$M_t = \sum_{i=1}^{p_n} (M_{t \wedge t_i^n} - M_{t \wedge t_{i-1}^n})$$

(this is where we use the fact that $M_0 = 0$ — otherwise we'd have to insert it into this sum, because it doesn't telescope with anything). Now when we take the square, we're going to get

$$\left(\sum_{i=1}^{p_n} (M_{t \wedge t_i^n} - M_{t \wedge t_{i-1}^n}) \right)^2.$$

And as usual, when you expand out the square of a sum, you get the diagonal terms and the cross-terms. The diagonal term is going to precisely be the sum of squared increments

$$\sum_{i=1}^{p_n} (M_{t \wedge t_i^n} - M_{t \wedge t_{i-1}^n})^2.$$

And for the cross-terms, we're going to get

$$2 \sum_{i=1}^{p_n} (M_{t \wedge t_i^n} - M_{t \wedge t_{i-1}^n}) \sum_{j<i} (M_{t \wedge t_j^n} - M_{t \wedge t_{j-1}^n}).$$

And when you write it in this particular way, you see that this inner sum over j also telescopes. So the second part is going to be equal to

$$2 \sum_{i=1}^{p_n} (M_{t \wedge t_i^n} - M_{t \wedge t_{i-1}^n})(M_{t \wedge t_i^n})$$

(when the sum over j telescopes, we get the two endpoints — one is at the right endpoint $j = i - 1$, and the other is at the left endpoint, where we get $M_0 = 0$).

And what's the point of all this algebra? Now I have this process which I can define as $2X_n^t$ (for fixed n, the process in t is precisely this sum).

Claim 9.16 — (X_t^n) is a martingale.

This is where you start believing that when you subtract out the quadratic variation you get a martingale (or that the QV should be of this form) — because on a discrete level, this is true. And actually this is a martingale even in continuous time. The only problem is that if you look at the sum of squared increments at continuous times, it's not necessarily increasing — it's increasing on your fixed points in the mesh, but in between it can vary. So you eventually want to take a fine-mesh limit and get an increasing process in continuous time.

So that's kind of what we're working towards; we'll continue with the proof next time.

§10 March 5, 2025

§10.1 Quadratic variation

Today we'll continue with the proof of the main theorem defining the quadratic variation.

Theorem 10.1

Let M be a continuous local martingale. Then there exists an increasing process $(\langle M, M\rangle_t)$, unique up to indistinguishability, such that $M_t^2 - \langle M, M\rangle_t$ is a continuous local martingale. Furthermore, for any $t > 0$ and an increasing sequence of subdivisions $0 = t_0^n < \cdots < t_{p_n}^n = t$ with mesh size tending to 0, we have that

$$\langle M, M\rangle_t = \lim_{n\to\infty} \sum_{i=1}^{p_n} (M_{t_i^n} - M_{t_{i-1}^n})^2$$

in probability.

This states that if there's a continuous local martingale, there's a unique increasing process such that M_t^2 minus this process is itself a continuous local martingale; and the way you compute it is by this limit. Last time we gave some intuition for why this should be the right thing.

Remark 10.2. We're assuming $(\mathcal{F}_t)$ is complete. That's an assumption in this theorem; we'll see why we need it. Basically, in principle we're going to get an increasing process on an almost sure event; and you want to modify it to be almost surely increasing, not just increasing. But when you modify it, you need to keep the condition that it's adapted to your filtration. So you want to be able to say that when you modify on a measure-zero event, you're still adapted. And that'll be true if your filtration is complete, because any measure-zero event is in $\mathcal{F}_0$.

§10.2 Some remarks

Another heuristic way to see this — recall the quadratic martingale (of e.g., Brownian motion). In the section on martingales, we had some general principles of how you generate martingales starting from a martingale itself, or some process with independent increments. In particular, if you take the square of the process, the thing to subtract is basically the variance. Here you have a local martingale instead of a martingale, but it's natural you want to subtract some variance-like thing; and infinitesimally the variance is supposed to look like this.

As another remark, you should basically think of the quadratic variation as the 'variance' of M. Of course, this doesn't make formal sense because M is a continuous local martingale, so it doesn't even have to

be integrable — so the variance isn't well-defined. But this is somewhat true on an infinitesimal level — basically, increments of $\langle M, M\rangle_t$ are telling you locally how much your variance is. In particular, you should think of M_t as in some sense a Gaussian random variable $\mathcal{N}(0, \langle M, M\rangle_t)$ — so you should basically think of continuous local martingales as sort of normal random variables. In some way, they're like Brownian motion, just with time changed.

One way to make this precise is that if at some time t you have $\langle M, M\rangle_t = \sigma^2$ almost surely (in principle it's a random variable, but we're supposing it's actually a constant), then it is actually true that $M_t \sim \mathcal{N}(0, \sigma^2)$. So this lends some motivation to heuristics like this — if the thing literally is a constant, then you literally do have a standard normal distribution. (We will need Ito's formula to prove this, but we'll prove it later.)

So basically you think of local martingales as mixtures of Gaussian distributions where the variance term could be random. In particular, maybe the reason M_t might not be integrable is that $\langle M, M\rangle_t$ can have very poor tail behavior — maybe it's super large on average, so your variance tends to be super large, and then when you compute this mixture the variance actually explodes.

As a third remark, in fact if B is a $(\mathcal{F}_t)$-Brownian motion, then one computes tha t$\langle B, B\rangle_t = t$ — basically the simplest possible increasing process. In particular, here the quadratic variation is not random, it's deterministic. (This is also consistent with what we just said about variance.) So you should always think of BM as the simplest, most concrete case.

Why is this true? We already showed previously that $B_t^2 - t$ is a martingale (the quadratic martingale of Brownian motion). And martingales are continuous local martingales; so by the uniqueness part, you've computed its quadratic variation.

§10.3 Proof of existence

Last time we had a short argument for why uniqueness is true. This was basically just that you can't be both a local martingale and a FV process (unless you're 0) — those two are incompatible with each other. And then we got to the following point: we fix $K > 0$ (think of this as an integer), and an increasing sequence of subdivisions $0 = t_0^n < \cdots < t_{p_n}^n = K$. We also started by making some simplifying assumptions: we assumed that $M_0 = 0$ (almost surely) and that M is bounded. Again, any time you prove a result, you start in the simplest possible case (because if you can't prove it in the simplest case, you probably can't prove it in full generality). This is similar to how in analysis, any time you prove some inequality (like Sobolev), you start with smooth functions (rather than a general function with finite Sobolev norm). So for us, we start with bounded martingales — now it's not just a local martingale, but a martingale.

And we wrote our local martingale at time t as the sum of its increments (basically, you can telescope), as

$$M_t = \sum_{i=1}^{p_n} (M_{t\wedge t_i^n} - M_{t\wedge t_{i-1}^n}).$$

(In principle you might need M_0, but here we don't have that because $M_0 = 0$.)

Then we had this computation — what I want to look at is M_t^2. And when you expand out the square of this sum, as usual you get the diagonal terms and the cross-terms. The diagonal term looks exactly like what we're claiming should be the quadratic variation, more or less — it's

$$\sum_{i=1}^{p_n} (M_{t\wedge t_i^n} - M_{t\wedge t_{i-1}^n})^2.$$

And for the cross-terms, there's this algebraic thing where you can actually write it as

$$2\sum_{i=1}^{p_n} M_{t\wedge t_{i-1}^n}(M_{t\wedge t_i^n} - M_{t\wedge t_{i-1}^n}).$$

The reason is in principle the cross-term is

$$2\sum_{j<i}(\text{increment at } i)(\text{increment at } j).$$

And you reorganize this as first a sum over i, and then inside, a sum over $j < i$. The increment over i is going to be the second term. And when you sum over $j < i$, you have this telescoping thing; so you just pick up the right endpoint and the left endpoint. The right endpoint is going to be $M_{t\wedge t_{i-1}^n}$, and the left endpoint is $M_0 = 0$. So you get this algebraic identity. And it's going to be quite important — we'll use it elsewhere in the proof as well (that you can group the cross-terms and get this telescoping sum).

Then we defined this sum as

$$\sum_{i=1}^{p_n} M_{t\wedge t_{i-1}^n}(M_{t\wedge t_i^n} - M_{t\wedge t_{i-1}^n}) = X_t^n.$$

So we get that

$$M_t^2 = \sum_{i=1}^{p_n}(M_{t\wedge t_i^n} - M_{t\wedge t_{i-1}^n})^2 + 2X_t^n$$

(where the first term is what we should eventually be thinking of as the quadratic variation).

The first thing to note in this proof is:

> **Claim 10.3 —** (X_n^t) is a bounded martingale.

In the end, this is where the statement comes from — we want to figure out what we need to subtract from M^2 to get a martingale. And we wrote M^2 in this way and saw the second term is a martingale, so we should just be subtracting the first.

Proof. The short reason is any time you have a Riemann sum approximation like this, with increments of your martingale and something measurable with respect to the left endpoint of your increment, you get a martingale. For example, when proving Doob's upcrossing lemma you define a submartingale out of your submartingale, and it's something like this — you always keep your thing you're multiplying the increment by measurable with respect to the left endpoint.

It really suffices to check that each term inside the sum is a martingale. If you ignore the sum, each summand is a process in t; so we just want to check that each summand is a martingale. More generally, it suffices to check that for any $r \le s$ and some bounded $\mathcal{F}_r$-measurable random variable Z, the process

$$V_t = Z(M_{t\wedge s} - M_{t\wedge r})$$

is a martingale. (It suffices to check this because it applies to each summand, and a sum of martingales is a martingale.)

(The fact that it's bounded comes from the fact that each of the M's is bounded.)

To see this, let $t_0 \le t_1$. We basically just want to check the martingale property

$$\mathbb{E}[V_{t_1} \mid \mathcal{F}_{t_0}] = V_{t_0}.$$

We'll do some case analysis.

Case 1 ($t_1 \le r$). Then by definition, $V_{t_0} = V_{t_1} = 0$ (because you're taking $t \wedge s$ and $t \wedge r$; so in V_{t_1} you get $M_{t_1} - M_{t_1} = 0$, and similarly with V_{t_0}).

Case 2 ($t_0 \le r \le t_1$). In this case, you can use the tower property to say that

$$\mathbb{E}[V_{t_1} \mid \mathcal{F}_{t_0}] = \mathbb{E}[V\mathbb{E}[M_{t_1\wedge s} - M_r \mid \mathcal{F}_r] \mid \mathcal{F}_{t_0}].$$

(We want to show the right-hand side is 0 — we still have $t_0 \le r$, so V_{t_0} is still 0.) But

$$\mathbb{E}[M_{t_1 \wedge s} - M_r \mid \mathcal{F}_r] = 0$$

by the martingale property. So this case is done.

Case 3 ($r \le t_0 \le t_1$). This may not be the most efficient way to do it, but now let's compute $\mathbb{E}[V_{t_1} \mid \mathcal{F}_{t_0}]$. Because Z is $\mathcal{F}_r$-measurable and therefore $\mathcal{F}_{t_0}$-measurable, we can pull it out, and we get

$$Z\mathbb{E}[M_{t_1 \wedge s} - M_{t_1 \wedge r} \mid \mathcal{F}_{t_0}].$$

And because of our assumption, $t_1 \wedge r = r$; and this is $\mathcal{F}_{t_0}$-measurable, so we don't need the conditional expectation, and so we get

$$Z(\mathbb{E}[M_{t_1 \wedge s} \mid \mathcal{F}_{t_0}] - M_{t_1 \wedge r}).$$

To finish, we just need to show that this first conditional expectation is $M_{t_0 \wedge s}$. We again split into cases — if $t_0 \ge s$, then we have

$$\mathbb{E}[M_{t_1 \wedge s} \mid \mathcal{F}_{t_0}] = M_s$$

(because $t_1 \ge t_0 \ge s$, so $t_1 \wedge s$ is just s, and M_s is $\mathcal{F}_{t_0}$-measurable). And if $t_0 < s$, then

$$\mathbb{E}[M_{t_1 \wedge s} \mid \mathcal{F}_{t_0}] = M_{t_0}$$

by the martingale property of M, which you can write as $M_{t_0 \wedge s}$. □

So in summary, we've shown this claim that for any n, this process (X_t^n) is a martingale. You see it just reduces to a bunch of case analysis. But the way you remember it is any time you see a sum of increments of your martingale multiplied by something measurable with respect to the left endpoint, you're going to get a martingale. In fact, this is basically how you define stochastic integrals.

You might think we've written M_t^2 minus this sum of squared increments as a martingale. Are we done? No. If you think about this thing

$$\sum_{i=1}^{p_n} (M_{t \wedge t_i^n} - M_{t \wedge t_{i-1}^n})^2$$

as a process in t, it's not necessarily increasing — it's only increasing on a mesh of points. On your mesh it's increasing because you're summing nonnegative things. But in between, it's not necessarily increasing — this thing may increase or decrease in between your mesh. So this itself is not an increasing process — it's only an increasing process if you restrict to your discrete set of points. So we're not done.

But this is why you want your mesh size tending to 0. We have a mesh basically tending to the continuum, so in the limit you might think that whatever limit you get out of this process in n is going to be increasing. And that's exactly what we want to show. So we want to take this limit in n and show these process converge. And whatever this process converges to, it's going to be increasing — because it's increasing on your mesh of points, which is getting dense in $[0, K]$. So that's what we're working towards.

How do you show this process $\sum_{i=1}^{p_n} (M_{t \wedge t_i^n} - M_{t \wedge t_{i-1}^n})^2$ converges? We're actually going to show the second process (X_t^n) converges (as $n \to \infty$) — because M_t^2 doesn't depend on n. The reason you want to do this is because it's a martingale, and martingales have special properties, so if you can work with a martingale you probably should.

What does it mean for a martingale to converge? I want to define a limiting process out of this, but I don't know what the limit is yet. So you basically show this is going to be a Cauchy sequence, and then appeal to some completeness about spaces to show any Cauchy sequence converges.

So basically, we want to show:

Goal 10.4. Show that

$$\lim_{n\to\infty} \sup_{m\geq 1} \mathbb{E}[(X_K^m - X_K^n)^2] \to 0.$$

Basically what this is saying is that your processes X^n, evaluated at the endpoint K, are Cauchy in $L^2(\Omega, \mathcal{F}, \mathbb{P})$ as a sequence in n.

This is where the true difficulty of the theorem lies — in showing this. We said last time that this is probably going to be the hardest proof in the class. The construction of stochastic integration and Ito's formula are relatively less involved. But to prove this, you actually have to do some long calculations, and in the middle it's kind of hard to see what the end result is going to be. So we're going to defer this. (And for the complete proof, there are some parts we might skip and assign as homework — because one absorbs a long proof better that way.)

Why is this going to be enough to eventually get what we want? We're going to assume this for now. We're saying this thing is hard, but modulo the hard thing, it's going to be relatively easy.

We said we wanted to show $\sum_{i=1}^n (M^2_{t\wedge t_i^n} - M_{t\wedge t_{i-1}^2})^2$ converges as a process; but we're saying the second term converges. But in this claim, we took the right endpoint, not the entire thing itself.

But the reason we can do this is because it's a martingale. For L^2 martingales, you can always convert a L^2 bound on the endpoint into a L^2 sup bound over t — because of the L^2 maximal inequality, which says that you actually have

$$\mathbb{E}\left[\sup_{t\in[0,K]} |X_t^m - X_t^n|^2\right] \leq 4\mathbb{E}[|X_K^m - X_K^n|^2]$$

(because each X is a martingale, if you take a difference you still get a martingale; so we're applying the L^2 maximal inequality to this difference, which is a martingale). (The constant 4 doesn't matter; but this should be the right constant for $p = 2$.) To change notation, we're going to write

$$\mathbb{E}[\|X^m - X^n\|^2_{\mathcal{C}([0,K])}]$$

to denote the sup on the left-hand side.

So you're going to apply this. And then this convergence is going to allow you to obtain a subsequence $\{n_j\}$ such that for all j, we have

$$\mathbb{E}[\left\|X^{n_j+1} - X^{n_j}\right\|^2_{\mathcal{C}([0,K])}] \leq 2^{-j}.$$

(This is similar to what you do when you show that convergence in probability means you have a subsequence converging almost surely — 2^{-j} doesn't matter much, you just want it to be summable in j). This implies

$$\mathbb{E}\left[\sum_{j=1}^\infty \left\|X^{n_j+1} - X^{n_j}\right\|_{\mathcal{C}([0,K])}\right] \leq \sum_{j=1}^\infty 2^{-j/2} < \infty$$

(the reason for $j/2$ is we want the L^1 norm instead of the L^2 norm, so we use Hölder and get a square root). This implies

$$\sum_{j=1}^\infty \left\|X^{n_j+1} - X^{n_j}\right\|_{\mathcal{C}([0,K])} < \infty \quad \text{almost surely.}$$

And that's going to imply that the subsequence (X^{n_j}) is Cauchy in the space of continuous functions on your interval $[0, K]$.

So let E be the event that the above sum is finite (so E is probability 1). Then on this event E, we have what we just said — the sequence (X^{n_j}) is Cauchy in $\mathcal{C}([0, K])$ (the space of continuous functions in my

interval). And this space is complete, so this means there exists some limiting continuous function. So we can define a process

$$Y_t(\omega) = \mathbf{1}_E(\omega) \cdot \lim_j X_t^{n_j}(\omega).$$

Basically, on the event E, I want to define Y to be the limit of the X^{n_j}, which exists because of what we just said (it's Cauchy); and off the event we define Y to be 0. Then Y is always going to be a continuous function.

So in particular, we obtain a stochastic process Y with continuous sample paths. Moreover, Y is adapted, because $(\mathcal{F}_t)$ is complete. (Why is Y_t going to be $\mathcal{F}_t$-measurable? In this limit, each $X_t^{n_j}$ is $\mathcal{F}_t$-measurable, just by the way we defined X^n. And you take a limit, and it's still going to be $\mathcal{F}_t$-measurable. And E is also $\mathcal{F}_t$-measurable, because it's in $\mathcal{F}_0$ by the completeness assumption. This is one of the places we need this completeness assumption — because you want a process which always has continuous sample paths, not just on some almost sure event.)

Moreover, for all times $t \in [0, K]$, we have that

$$\mathbb{E}[\left|X_t^{n_j} - Y_t\right|^2] \to 0.$$

(i.e., we get convergence in L^2). The point is we had almost sure convergence, so we need to say why that implies convergence in L^2. And that's just because everything is bounded — X is bounded, so Y is also bounded, and this is just going to be true.

Why did we want this almost sure convergence? We want to say why Y is itself a martingale. It's a limit of martingales, and as long as we have some sort of convergence like this in L^2, it's itself also going to be a martingale. So we can show that

$$\mathbb{E}[Y_t \mid \mathcal{F}_s] = Y_s,$$

i.e., Y is a martingale — because in the pre-limit this is true when you replace Y by X^{n_j}. And then you want to take a limit and say the conditional expectations converge, and that'll be true by the above convergence thing.

So we get a martingale Y, at least on our interval $[0, K]$. Then we obtain that $(Y_{t \wedge K})$ is a martingale on the entire real line — because if you have a martingale on this finite interval, and you just make it constant after that interval, then it's going to be a martingale for all time.

So now we're basically done with the martingale part — we succeeded in showing that the limit of X^n gives a martingale. Now we want to say why, when we subtract $2X^n$ from M_t^2, we get an increasing process.

> **Goal 10.5.** Show that $M_t^2 - 2Y_t$ is increasing.

This will basically finish the construction of the quadratic variation, at least on a bounded interval $[0, K]$ — because we get this increasing process, and then when we subtract it from M_t^2 we get $2Y_t$, which is a martingale.

Why is this increasing? Well, for each n, this sum of squared increments

$$t \mapsto \sum_{i=1}^{p_n} (M_{t_i}^n - M_{t_{i-1}^n})^2$$

is increasing on our mesh of points (t_i^n) — so if you restrict your times onto this discrete mesh of points, you get an increasing function.

And then we took the limit. Because the X^n's converge in $\mathcal{C}([0, k])$, so does this function (because their sum M_t^2 does not depend on n). So we obtain that (at least on the event E — off this event we don't have convergence)

$$M_t^2 - 2Y_t = \lim_j M_t^2 - 2X_t^{n_j}$$

is increasing on a countable dense subset of $[0, K]$. That's because if you take t to be any point in your mesh, then it exactly coincides with this discretized function (the pre-limit on your mesh). So when you take the limit, you get this increasing function on the union of all your meshes; and because you took the mesh size to 0, that union is countable dense.

And because it's continuous, that implies $M_t^2 - 2Y_t$ is increasing on the entire interval $[0, K]$.

So now we just define our increasing process A^K by

$$A_t^K = \mathbf{1}_E \cdot (M_t^2 - 2Y_t).$$

And based on what we just said, when E happens A is increasing; and if E does not happen then A is just always 0, so it's always increasing.

So A^K is an increasing process, and the process

$$(M_{t\wedge K}^2 - A_{t\wedge K}^k)_{t\geq 0}$$

is a martingale. (We're adding in these $t \wedge K$'s to extend the process to the entire interval; but after K it's just a constant, so I only have to worry about whether it's a martingale before time K. And before time K it's just going to be $2Y$, at least on an almost sure event; so I get that this is a martingale.)

§10.3.1 Stitching things together

Now what it remains to do is — I kind of constructed this process on any finite interval, so now I just want to stitch these processes together to get a process on the infinite interval (basically, I want to remove this min with K). So we apply what we've obtained with K being any integer $\ell \geq 1$, and we get that for each ℓ, we have a process $(A_t^\ell)_{t\in[0,\ell]}$. And one can check that these processes with different ℓ's have to coincide with each other, at least up to indistinguishability — almost surely,

$$A_{t\wedge\ell}^\ell = A_{t\wedge\ell}^{\ell+1}$$

for all $\ell \geq 1$ and $t \geq 0$. Why is that the case? Basically it comes from the uniqueness part of the theorem. Because if I look at — by definition $A^{\ell+1}$ is supposed to make

$$M_{t\wedge(\ell+1)}^2 - A_{t\wedge(\ell+1)}^{\ell+1}$$

a martingale. (Basically we constructed $A^{\ell+1}$ so that this is going to be true.) But if you restrict to the smaller interval $[0, \ell]$, then this thing we wrote is just going to be equal to

$$M_{t\wedge\ell}^2 - A_{t\wedge(\ell+1)}^{\ell+1}.$$

But then we also constructed

$$M_{t\wedge\ell}^2 - A_{t\wedge\ell}^\ell$$

to be a martingale. So we have these two processes such that at least on this common interval, we get a martingale when we subtract the process from M^2. And basically redoing the proof of uniqueness, that's going to tell me the difference of A^ℓ and $A^{\ell+1}$ is both a martingale and a FV process (because they're both FV processes). And that can't happen because their difference is 0.

You want this because you can stitch these things together; so now let's say why.

SO let F be the event that this holds — i.e., that $A_{t\wedge\ell}^\ell = A_{t\wedge\ell}^{\ell+1}$ for all $\ell \geq 1$ and $t \geq 0$. Then we just define

$$\langle M, M\rangle_t = \mathbf{1}_F A_t^{\lceil t\rceil}.$$

(We could have put in any integer greater than or equal to t in the superscript, because on the event F, it's all the same.) Then one checks that this actually defines an increasing process, and that $(M_t^2 - \langle M, M\rangle_t)$ is a martingale. (The fact it's increasing is it coincides with an increasing process on any finite interval; and the fact that it's a martingale is again because if you look at a finite interval $[0, \ell]$ it's a martingale, and ℓ is arbitrary.) Here we again took an indicator on an almost sure event, so this is another place where we need the filtration to be complete (because we want $\langle M, M\rangle$ to be $\mathcal{F}_t$-measurable).

Student Question. *Why did we need to do this partitioning on $[0, K]$ — why couldn't we just do it on the entire real line?*

Answer. When you look at the endpoint, ultimately you want that the process converges, but you usually prove that just by showing the endpoint converges — that

$$\limsup_{n\to\infty} \sup_{m\geq n} \mathbb{E}[(X_K^m - X_K^n)^2] \to 0.$$

And it'd be less clear what's the endpoint otherwise.

So we've basically shown the first part of the theorem, modulo the statement in yellow — at least under the assumption that M is bounded and $M_0 = 0$. Now let's show under those same assumptions why the second part is true. That's basically going to follow from what we already did.

For the second part, note that by construction, we have (almost surely)

$$\langle M, M\rangle_K = A_K^K = M_K^2 - 2Y_K$$

by construction. And this was basically

$$\lim_{j\to\infty} \sum_i (M_{t\wedge t_i^{n_j}} - M_{t\wedge t_{i-1}^{n_j}})^2$$

in L^2. (This is because we showed Y^k itself converges in L^2; and the difference $Y^k - A^k$ doesn't depend on k.)

But L^2 convergence implies convergence in probability. So we get this statement, but actually even more — the convergence is actually in L^2, at least when you have a bounded martingale.

§10.3.2 The general case

So we've proven the theorem under the simplifying assumption that you have a bounded martingale M. Now let's briefly talk about how you extend to the general case of a continuous local martingale not necessarily started at 0. The full details will be on the next homework, because again it's basically that you want to do this stitching thing — by definition a continuous local martingale always has a sequence of stopping times reducing it, so you basically localize so that you have a bounded martingale. Then you apply this result for bounded martingales; and then you take the sequence of stopping times to ∞ and stitch them together. (It's in the book, so you can just read it and write it up if you want, but it'd probably help conceptually to try to work it out.)

In general, youc an write a continuous local martingale M as

$$M_t = M_0 + N_t$$

(where N is a continuous local martingale starting at 0). That means you can expand

$$M_t^2 = M_0^2 + 2M_0N_t + N_t^2.$$

And there's an exercise on this week's homework:

Exercise 10.6. M_0N_t is a continuous local martingale.

So if you can construct the quadratic variation of the N part — your continuous local martingale starting at 0 — then you've constructed the quadratic variation of the whole thing (because the QV is what you have to subtract to get a CLMG, and $M_0^2 + 2M_0N_t$ is already a CLMG).

So with this, we can reduce to the case that $M_0 = 0$ (which we'll assume from now).

And for a general CLMG starting at 0, this will be on the next homework. To start, you basically want to take a sequence of stopping times which not only reduces your local martingale, but makes it bounded — you can for instance take

$$T_n = \inf\{t \geq 0 \mid |M_t| \geq n\}.$$

Then you apply our previous result to the stopped process M^{T_n} — this is always going to be bounded now — and then stitch things together. So that's kind of the summary of what you'd do.

So that kind of — once you do these exercises, that basically completes the proof of this theorem, modulo the hard part (the statement in yellow). Now we'll talk about how you show this claim in yellow.

Student Question. *Why is A finite variation?*

Answer. It's continuous and increasing. And as a result, it's going to be finite variation. Because for instance, in Le Gall there's some statement where if you can write your function as a difference of two continuous increasing functions, then it's finite variation. And here it's just itself.

But the more actual reason why this is true is you basically think of your continuous increasing function as the CDF of a measure — not necessarily a probability measure, but it'll be some positive finite measure. Then you can apply the Caratheodory extension theorem to say why you can go from a CDF to a measure itself — basically a CDF allows you to define a measure on a union of disjoint intervals (by taking the difference of two endpoints), so now you have some sort of pre-measure on unions of disjoint intervals (which is an algebra or something). And by Caratheodory you can extend this to a measure on the real line (or at least on the fixed interval you're starting with).

§10.4 Proving Goal 10.4

Now let's go back to proving this claim. Let's first fix $m \geq n$ — so the partition t^m is finer than t^n. We'll draw t^n in blue, and t^m in yellow on top of it (but you should really think of them as being overlaid).

Before we show this convergence to 0, we just want to compute this second moment; so let's try to compute that. How are we going to compute it? Well, let's start by first computing the cross-term $\mathbb{E}[X_K^m X_K^n]$. (When you expand out the square, you're going to get this cross-term; so we'll have to understand this guy first.)

By the definition of the X^n, we're going to get two sums — one corresponding to each partition — and because we're just looking at the endpoint K, there are no more mins (each t_i is at most K). So we get

$$\mathbb{E}[X_K^m X_K^n] = \sum_{i=1}^{p_n} \sum_{j=1}^{p_m} \mathbb{E}[M_{t_{i-1}^n} \Delta_i^n M_{t_{j-1}^n} \Delta_j^m],$$

where $\Delta_i^m = M_{t_i^n} - M_{t_{i-1}^n}$ is your increment (this will reduce the amount of writing). (Each X_K^n is a sum, so when you take the product you get two sums.)

The first thing to note is that:

Claim 10.7 — If i and j are such that $(t_{i-1}^n, t_i^n]$ and $(t_{j-1}^m, t_j^m]$ are disjoint, then this expectation is 0.

This all comes back to the fact that L^2 martingales are quite special. The point is you want to reduce the amount of terms in your double sum. And actually you note that almost all the cross-terms are zero; there's only certain cross-terms which are nonzero.

Proof. For any i and j satisfying this condition, you can basically reduce to the following scenario: you have four ordered times $u_1 < u_2 < u_3 < u_4$, and you can write this expectation as

$$\mathbb{E}[M_{u_1}(M_{u_2} - M_{u_1})M_{u_3}(M_{u_4} - M_{u_3})].$$

This is because t^m is a finer partition, so either your smaller interval is contained in one of the larger ones, or it's disjoint, in which case you can do this.

And now you use tower — you take a further conditional expectation with respect to the second-to-last endpoint, and this becomes

$$\mathbb{E}[M_{u_1}(M_{u_2} - M_{u_1})M_{u_3}\mathbb{E}[M_{u_4} - M_{u_3} \mid \mathcal{F}_{u_3}]]$$

(everything before is $\mathcal{F}_{u_3}$-measurable). And by the martingale property this inner expectation is 0. □

For some notation, note that for each $1 \le j \le p_m$, there is a unique index i, which we're going to write as $i(j)$, such that

$$(t^m_{j-1}, t^m_j] \subseteq (t^n_{i-1}, t^n_i].$$

And so then by using this claim, we obtain that

$$\mathbb{E}[X^n_K X^m_K] = \sum_{j=1}^{p_m} \mathbb{E}[M_{t^n_{i(j)-1}} \Delta^n_{i(j)} M^m_{t_{j-1}} \Delta^m_j]$$

(we're switching the sum; then if I first sum over j and the inner sum is over i, there's only one i that's going to contribute, and it's precisely $i(j)$).

In particular, for $n = m$, you get that the second moment of X^n_K takes the form

$$\mathbb{E}[(X^n_K)^2] = \sum_{i=1}^{p_m} \mathbb{E}[M^2_{t^n_{i-1}} (\Delta^n_i)^2] \tag{10.1}$$

(the two parts of the product are the same). That's one identity we're going to need.

Now we'll just write out the claim for the formula for $\mathbb{E}[X^n_K X^m_K]$. We've got to this point, and we're going to expand out this second moment. We have these non-cross terms for n, and similarly you're going to have an identity for m like this. And you're also going to have this cross-term. Somehow we need to reorganize these to get an identity we can work with. That final identity is hard to guess, but in the end we get that:

Claim 10.8 — We have

$$\mathbb{E}[(X^n_K - X^m_K)^2] = \sum_{j=1}^{p_m} \mathbb{E}[(M_{t^m_{j-1}} - M_{t^n_{i(j)-1}})^2 (\Delta^m_j)^2].$$

So through some magic reorganization procedure you can get something like this (which we're not going to do today, because we'd run out of time and it'll be on next week's homework). We might indicate a simple case where you have just four points on Monday.

So let's just take this for granted now, and show why this is going to imply Goal 10.4. So this is our new claim, which we'll talk about more on Monday (and part will be next week's homework) — proving this identity starting from the one for $\mathbb{E}[X^n_K X^m_K]$. (You have to do some magic reorganizations adn use the martingale property in the right places.) If you want to get started on this before Monday, try the example with four points (it's in the notes too). You can also try to prove this straight away, but if you work out the 4-point example, you kind of see the right way to reorganize things. Again, if you want to prove things in full generality, it's a good general principle to first try a very simple case.

So now let's assume this and finish off the proof. We want to bound the right-hand side of Claim 10.8. We're going to apply Cauchy–Schwarz. But before we do that, we want to bound each of these differences just by taking a sup — this gives

$$\mathbb{E}[(X^n_K - X^m_K)^2] \le \mathbb{E}\left[\sup_{i,j:i(j)=i} \left|M^m_{t_{j-1}} - M_{t^n_{i-1}}\right|^2 \sum_{j=1}^{p_m} (\Delta^m_j)^2\right]$$

(the point was we replaced the first thing with the max it could be, so we could pull it out of the sum). Now if you apply Cauchy–Schwarz, you can bound this thing by the fourth moments of each thing, and you get

$$\mathbb{E}\left[\sup_{i,j:i(j)=i}\left|M^m_{t^m_{j-1}}-M_{t^n_{i-1}}\right|^4\right]^{1/2}\cdot\mathbb{E}\left[\left(\sum_{j=1}^{p_m}(\Delta^m_j)^2\right)^2\right]^{1/2}.$$

And why does this converge to 0? The first thing is going to converge to 0 because I'm taking the mesh size to 0, so as n gets larger and larger, the distance between these two points t^m_{j-1} and t^m_{i-1} is going to get smaller and smaller, and then you can use uniform continuity.

So let's call these two expectations I_1 and I_2. Then we just showed $I_1 \to 0$ — before you take the expectation this converges to 0 surely, because M is continuous on a bounded interval, so it's uniformly continuous, and so the sup is going to be over smaller and smaller intervals. So the sup is going to converge to 0 almost surely. And M is bounded, so you get this convergence to 0 of the expectation as well.

So then we just need to show that the second term is $O(1)$ — you want to bound it uniformly in m and n. Why is that true? Well, if we expand — let's take the term without the final square root(since that doesn't matter). If we expand the square of this sum, we get

$$\sum_{j=1}^{p_m}(\Delta^m_j)^4+\sum_{j\neq i}(\Delta^m_j)^2(\Delta^m_i)^2$$

(we again get the diagonal terms, and then the off-diagonal terms). And then I want to look at the expectations of these things, which I'll call E_1+E_2.

Let's look at E_1 and E_2 individually. The thing I know about L^2 martingales is that if this were not a 4 but actually a 2, I'd be in business — the increments of L^2 martingales are uncorrelated, so if I sum the squares and take an expectation, that's just the expectation of the square of the endpoint.

But I took a *bounded* martingale, so I can just bound two powers of this by a constant. So let C be the constant with $|M|\le C$. Then I can bound

$$E_1 \le C^2\mathbb{E}\left[\sum_{j=1}^{p_n}(\Delta^m_j)^2\right].$$

And it's a L^2 martingale, so this is precisely

$$C^2\mathbb{E}[M_K^2].$$

(Again in full generality it'd be $(M_K-M_0)^2$ for a general L^2 martingale, but we took $M_0=0$.) And M is bounded, so this is just bounded by C^4.

So the diagonal terms don't cause you a problem. To finish, we just have to say why E_2 is fine. A priori we're afraid because they're cross-terms, so there are way more of them. Again you want to use magic properties of L^2 martingales to combine things — you only want to sum once.

Again, if we fixed j and summed over j' with $j<j'$, I want to be able to bound the inner sum by something good. And this is again going to be the uncorrelated property of L^2 martingales. Basically if you fix j and sum over j', you condition on Δ^m_j or something; and then you use the uncorrelated property of L^2 martingales, even when you take conditional expectations, to prove this.

§11 March 10, 2025

§11.1 Review

Last time we constructed the QV; the main idea was to consider some mesh and take

$$M_t^2 - \sum_i (M_{t_i^n \wedge t} - M_{t_{i-1}^n \wedge t})^2.$$

And you noted that this was a martingale. Then you take a limit as $n \to \infty$, and get an increasing continuous function. The way you show the sum converges to a limit is by showing the *martingale* converges to a limit. (This is all under the assumption that M is bounded; removing that assumption is on the homework.) You want to work with martingales because martingales have nice properties.

Last time, we promised to wrap up some of the things, but we're going to skip that; you can read the notes for the final estimates. We wanted to show some fourth moment was uniformly bounded in n; and this fourth moment is basically the second moment of something like $\sum_i (M_{t_i^n \wedge t} - M_{t_{i-1}^n \wedge t})^2$. And we sshowed that when you expand out the square, the diagonal term is relatively okay, and the cross-terms you can also bound. Again that's a place where you use L^2 martingale theory. And then there's the computation of this second moment identity which was kind of the key to everything; we'll probably skip over the explicit case of when you take a mesh of four points, because you're going to work out the general case for homework. If you're stuck, you can read the special case of four points.

Under the hood, we basically had to construct an instance of a stochastic integral — actually what this martingale Y_t is going to be is a stochastic integral of the sense $Y_t = \int_0^t M_s \, dM_s$. When we construct stochastic integrals and apply Ito's formula — we're working towards constructing something like this, but after doing that we'll see that this martingale Y (which we defined as a limit of the martingales above) is basically going to be this stochastic integral.

So basically, what you're going to get is some identity like

$$M_t^2 = 2 \int_0^t M_t \, dM_t + \langle M, M \rangle_t.$$

This is going to be an instance of Ito's formula. So what we basically did was proving Ito's formula in the case where you're taking the square of a martingale.

Recall the heuristic that if $X(t)$ is a continuously differentiable function in time and we consider $f(X(t))$, then classical calculus tells us that we get the chain rule

$$\frac{d}{dt} f(X(t)) = f'(X(t)) X'(t).$$

Ito's formula is different because it's stochastic calculus. Here if you take $f(x) = x^2$, then if you believed in ordinary calculus you'd (informally) get

$$dM_t^2 = 2 M_t dM_t.$$

But the point here is that X (which is now M_t) is not continuously differentiable — it doesn't even have finite variation (it's a martingale, so locally it varies like $\sqrt{\bullet}$). So you have to add in this quadratic term, getting

$$dM_t^2 = 2 M_t dM_t + d\langle M, M \rangle_t.$$

All this is to say that after we cover stochastic calculus and Ito's formula, when you look back at this you'll see that what we were basically doing was proving this particular instance of Ito's formula.

So basically we've constructed the quadratic variation; and now let's continue with developing the theory of local martingales.

§11.2 Properties of continuous local martingales and quadratic variation

Proposition 11.1

Let M be a continuous local martingale, and let T be a stopping time. Then almost surely, for all $t \geq 0$, we have that

$$\langle M^T, M^T \rangle = \langle M, M \rangle_{T \wedge t}.$$

We could also have written the right-hand side as $\langle M, M \rangle_t^T$ (by definition). So the QV of the stopped martingale M^T is the same as the stopped QV of the original.

What does this mean? If you think about it, the stopped martingale M^T becomes constant after the stopping time T. So after the stopping time, your QV should basically be constant — it's measuring how much your martingale is varying, and if you're constant then there's no variation. So you'd expect that after T, this QV should just become constant. (The QV is an increasing process.) And that's precisely what's satisfied by the stopped QV $\langle M, M \rangle_{T \wedge t}$, because after T this just becomes constant.

That's intuitively how you think about this statement; but the way you prove it is by using uniqueness of the QV process.

Proof. By definition $M_t^2 - \langle M, M \rangle_t$ is a continuous local martingale (CLMG). So when you consider the stopped process

$$M_{T \wedge t}^2 - \langle M, M \rangle_{T \wedge t},$$

you again get a continuous local martingale. But you could've written this as $(M_t^T)^2 - \langle M, M \rangle_{T \wedge t}$. So we're saying this stopped quadratic variation makes the square of the stopped local martingale a local martingale. And by uniqueness, we thus get that $\langle M, M \rangle_{T \wedge t}$ (this stopped QV) has to be the same as $\langle M^T, M^T \rangle_t$ (which is defined as the *unique* increasing process, up to indistinguishability, starting at 0 such that when you subtract it from $(M_t^T)^2$ you get a CLMG). □

Now we basically said that if your QV is constant after some point, then basically your martingale should also be constant — it shouldn't be varying — because the QV is measuring how much your martingale is varying. In particular:

Proposition 11.2

Let M be a continuous local martingale with $M_0 = 0$ almost surely. Then $\langle M, M \rangle = 0$ (for all $t \geq 0$) if and only if $M = 0$.

(Both of these statements are up to indistinguishability.)

Proof. One direction is immediate — if $M = 0$, then the QV has to be 0 (because 0 is already a continuous local martingale, so you can subtract 0 and get a continuous local martingale). So we just have to prove the converse implication.

But for this, we're going to get that M_t^2 is a continuous local martingale — if the QV is just 0, then by definition M^2 is a continuous local martingale. And it's also nonnegative. And thus, because M starts at 0, this means M_t^2 is a supermartingale. (This was a result we proved last week — if you have a nonnegative continuous local martingale started at 0, then it's a supermartingale.) In the notes we wrote out a direct proof of the following fact, but the proof is basically proving this fact (that if you have a nonnegative CLMG started at 0, it's a supermartingale).

Fact 11.3 — We have $\mathbb{E}[M_t^2] \leq \mathbb{E}[M_0^2] = 0$.

But then this basically finishes — it tells you that for any fixed time t, $M_t = 0$ almost surely. And because M is continuous, this means it's indistinguishable from 0. □

So those are two basic properties of the relation between QV and local martingales. Let's now prove another property which allows you to say when your martingale is in L^2, just based on the quadratic variation.

Notation 11.4. If A is an increasing process, we define $A_\infty = \lim_{t\to\infty} A_t \in [0, \infty]$.

The limit could be infinite, but it always exists because A is increasing.

Theorem 11.5

Let M be a continuous local martingale with $M_0 \in L^2$.

(i) The following are equivalent:

(a) M is a (true) martingale bounded in L^2.

(b) $\mathbb{E}[\langle M, M\rangle_\infty] < \infty$.

Furthermore, if these hold, then $M_t^2 - \langle M, M\rangle_t$ is a uniformly integrable martingale, and thus

$$\mathbb{E}[M_\infty^2] = \mathbb{E}[M_0^2] + \mathbb{E}[\langle M, M\rangle_\infty].$$

(ii) The following are equivalent.

(a) M is a (true) martingale with $M_t \in L^2$ for all $t \geq 0$.

(b) $\mathbb{E}[\langle M, M\rangle_t] < \infty$ for all t.

Furthermore, if these hold, then $M_t^2 - \langle M, M\rangle_t$ is a (true) martingale.

For (i), the identity on second moments follows from the uniform integrability (and optional stopping) — whenever you have a uniformly integrable martingale, its expectation at time 0 is the same as its expectation at time ∞. Its expectation at time 0 is just $\mathbb{E}[M_0^2]$, and its expectation at time ∞ is $\mathbb{E}[M_\infty^2] - \mathbb{E}[\langle M, M\rangle_t]$.

Remark 11.6. There's an analogous statement in discrete time — there's some notion of quadratic variation for discrete-time martingales, and you have a similar statement characterizing when you have a bounded L^2 martingale.

This is all consistent with the heuristic we said last time where we think of the QV as the 'variance' of your local martingale. It doesn't have to be integrable, but if your variance is finite in this sense, then you're actually in L^2.

For (ii), sometimes $\mathbb{E}[\langle M, M\rangle_\infty] < \infty$ is too much to require. For instance, if M is a BM, the QV at time t is just t, so the QV at time ∞ is ∞. So even for such a nice martingale, this statement is not true. Here we're not requiring M to be bounded in L^2; it just has to be in L^2 for all times. The integrability of $M_t^2 - \langle M, M\rangle_t$ is just because $M_t^2 \in L^2$ by (a) and $\langle M, M\rangle_t \in L^1$ by (b).

Proof of (i). Assume (a) holds, so M is a (true) martingale bounded in L^2. Let (T_n) reduce $M_t^2 - \langle M, M\rangle_t$. (This is going to be a local martingale, so we can always take a sequence of times for which it reduces.) Then for all times t, we'll have that

$$\mathbb{E}[(M_t^{T_n})^2 - \langle M, M\rangle_{T_n \wedge t}] = \mathbb{E}[M_0^2]$$

(because once we localize using the stopping time T_n, this becomes a UI martingale, and we can use the optional stopping theorem). And by assumption, the right-hand side is finite.

We want to show $\mathbb{E}[\langle M, M\rangle_\infty]$ is finite, so let's rearrange and write

$$\mathbb{E}[\langle M, M\rangle_{T_n \wedge t}] = \mathbb{E}[(M_t^{T_n})^2] - \mathbb{E}[M_0^2].$$

Now what we want to do is take $n \to \infty$ and apply some sort of convergence theorem — either Fatou or monotone convergence or dominated convergence (these are the usual three things you have to work with). That certainly works for the left-hand side — you can either apply Fatou or monotone convergence (because this thing actually is monotone as $n \to \infty$). So on the left-hand side, you just get $\mathbb{E}[\langle M, M\rangle_t]$.

The right-hand side is slightly more problematic — $(M_t^{T_n})^2$ is not dominated by any integrable random variable, so you can't apply DCT. It's also not monotone in n (or there's no reason it has to be). And Fatou's also doesn't really work.

But what you kind of realize is that you can apply the L^2 maximal inequality to this second part. First, we can trivially bound

$$\mathbb{E}[\langle M, M\rangle_{T_n \wedge t}] \le \mathbb{E}[M_0^2] + \mathbb{E}\left[\sup_{s \in [0,t]} M_s^2\right]$$

(because $T_n \wedge t \in [0, t]$). And then the L^2 maximal inequality tells me that this is at most

$$\mathbb{E}[M_0^2] + 4\mathbb{E}[M_t^2].$$

And this is going to be finite.

So now Ihave this inequality

$$\mathbb{E}[\langle M, M\rangle_t] \le \mathbb{E}[M_0^2] + 4\mathbb{E}[M_t^2]$$

for any t. Now if we take $t \to \infty$, we get that

$$\mathbb{E}[\langle M, M\rangle_\infty] \le \mathbb{E}[M_0^2] + 4 \sup_{t \ge 0} \mathbb{E}[M_t^2],$$

and we assumed the right-hand side is finite. So this proves that (a) implies (b).

Now let's prove that (b) implies (a). So now let's start with the identity

$$\mathbb{E}[(M_t^{T_n})^2 - \langle M, M\rangle_{T_n \wedge t}] = \mathbb{E}[M_0^2]$$

and rearrange it so that on the left-hand side we have the thing we want to bound; then we get

$$\mathbb{E}[(M_t^{T_n})^2] = \mathbb{E}[M_0^2] + \mathbb{E}[\langle M, M\rangle_{T_n \wedge t}].$$

Now this is where you can kind of apply Fatou on the left-hand side — if we send $n \to \infty$, we have $M_t^2 = \liminf_{n \to \infty} (M_t^{T_n})^2$, and on the right-hand side you can apply monotone convergence; so we get

$$\mathbb{E}[M_t^2] \le \mathbb{E}[M_0^2] + \mathbb{E}[\langle M, M\rangle_\infty].$$

And we assumed $\mathbb{E}[\langle M, M\rangle_\infty]$ is finite, so we get $\sup_{t \ge 0} \mathbb{E}[M_t^2]$ is finite. That proves (b) implies (a). □

Proof of (ii). For (ii), you just apply (i) to a stopped martingale M^{t_0} (where we fix some time t_0 for which we want to show (ii)). If your martingale is always in L^2, then this stopped martingale is bounded in L^2 (it's bounded by the second moment at the endpoint t_0). So now you have an L^2-bounded martingale, and you get that $\langle M^{t_0}, M^{t_0}\rangle_\infty$ has finite first moment. But by our previous proposition, this is just $\langle M, M\rangle_{t_0 \wedge \infty} = \langle M, M\rangle_{t_0}$. And vice versa (if $\langle M, M\rangle_{t_0}$ is finite, then M^{t_0} will be bounded in L^2; and if you apply this with any t_0, then you get a martingale that's always in L^2). □

The one lesson to take away is any time you're working with L^2 martingales, you always remember the L^2 maximal inequality. If you're trying to prove something and your inequality isn't working out, you might want to use this — it's very strong. (More generally there's an L^p maximal inequality for any $p > 1$.) This is a property specific to martingales, and it's very useful.

§11.3 Bracket of continuous local martingales

Now let's extend the QV to a bilinear form on local martingales. Just like there's variance, there's also covariance; so that's kind of what we're going to define next.

Definition 11.7. Let M and N be continuous local martingales. Then their *bracket*, denoted $\langle M, N\rangle$, is the finite variation process defined as

$$\langle M, N\rangle_t = \frac{1}{2}(\langle M+N, M+N\rangle - \langle M, M\rangle - \langle N, N\rangle).$$

(Sometimes this is also called the *covariation*.)

The whole point of where this definition comes from is, very abstractly, that if you have a notion of L^2 length, then you can basically recover the inner product. In other words, in a Hilbert space, knowing the length of any vector allows you to recover the inner product between any two vectors, by this formula; the reason that's true is because your inner product is bilinear.

So if you want to *define* the inner product between any two martingales, you use this identity — because if you assume bilinearity and imagine expanding out the right-hand side, then formally you exactly get $\langle M, N\rangle$ (if you expand out you get the diagonal terms and two cross-terms, and if the thing is symmetric you get exactly $\langle M, N\rangle$). (But in reality we're *defining* the bracket in this way.)

Proposition 11.8

Let M and N be continuous local martingales.

(i) $\langle M, N\rangle$ is the unique finite variation process (up to indistinguishability) such that $M_tN_t - \langle M, N\rangle_t$ is a continuous local martingale.

(ii) The map $(M, N) \mapsto \langle M, N\rangle$ is bilinear and symmetric.

(iii) If $0 = t_0^n < \cdots < t_{p_n}^n = t$ is an increasing sequence of subdivisions of $[0, t]$ with mesh size tending to 0, then

$$\lim_{n\to\infty} \sum_{i=1}^{p_n} (M_{t_i^n} - M_{t_{i-1}^n})(N_{t_i^n} - N_{t_{i-1}^n}) = \langle M, N\rangle_t$$

in probability as $n \to \infty$.

(iv) For every stopping time T, we have

$$\langle M^T, N^t\rangle = \langle M^T, N\rangle = \langle M, N\rangle^T.$$

(v) If M and N are (true) martingales bounded in L^2, then $M_tN_t - \langle M, N\rangle_t$ is a uniformly integrable martingale. Consequently, $\langle M, N\rangle_\infty$ is well-defined (as the almost sure limit of $\langle M, N\rangle_t$) and in L^1, and

$$\mathbb{E}[M_\infty N_\infty] = \mathbb{E}[M_0N_0] + \mathbb{E}[\langle M, N\rangle_t].$$

For (i), this is completely analogous to the quadratic variation — if $M = N$ then this thing turns out to exactly be the quadratic variation. You'd certainly expect that; the reason it's true is you have to look at $\langle 2M, 2M\rangle$ and say why this is $4\langle M, M\rangle$. And how would you prove this? For instance, you can use the characterization — if you consider $(2M_t)^2 - 4\langle M, M\rangle_t$, this is going to be a local martingale — because you can take out a factor of 4 and you get 4 times a local martingale. And by the uniqueness of QV, we're going to get $\langle 2M, 2M\rangle = 4\langle M, M\rangle$.

For (ii), in fact $\langle -, -\rangle$ is going to actually be a Hilbert space inner product. To say that, you want to say

that if your norm is 0 then your vector is actually 0. And we just proved that — your norm is the QV, and we just proved that if your QV is 0 then the martingale is actually 0.

For (iii), this is a discrete approximation statement; instead of taking a sum of squared increments, we're taking a sum of products of increments. For the proof, you just go back to the definition of the bracket, and use the fact that for QV you have this discrete approximation result. And then you're basically doing this calculation about expanding out a square, and then you're going to get this.

Property (iv) is the most delicate to prove; for the others you can probably see how to prove it, but this one actually takes some argument. The proof that the first and third are equal is just by uniqueness; but the fact that you can remove the stopping on N will take some effort.

Proof sketches. For most of these, we're just going to say verbally what you do; we'll focus on (iv), the hard one. For (i), you prove this the same way you prove uniqueness for QV — if you have a CLMG which is also FV, it has to be 0. For (ii), the fact it's bilinear and symmetric — for instance, the fact that it's symmetric just follows from the definition (if you swap the roles of M and N, you just want to say $M+N=N+M$, which is just true). For bilinearity, the way you'd maybe prove this is by first proving (iii), and then using the fact that you have these discrete approximations; then you can just check bilinearity for these discrete approximations, which is true.

And (iii) is true because you can use the discrete approximations for each of the QV terms and calculate what you'd get, and it's all set up so that you'd get these cross-terms.

For (v), by our previous results, if you have L^2-bounded martingales then M_∞ and N_∞ exist and are in L^2, so their product is in L^1. The fact that the limit of this bracket exists follows from the fact that the limits of $\langle M+N, M+N\rangle$ and so on exist (because these are just QVs). So the limit exists almost surely. And to see that it's in L^1, you again just bound by the L^1 norms of each of those brackets $\langle M+N, M+N\rangle$ and so on — each of these things is individually integrable, so their sum is also integrable. And the final identity should reduce to the one for $\langle M, M\rangle$ (where again you expand out the square and rearrange terms). □

Proof of (iv). Let's first prove the easy part — that the first and third terms are equal. We've already proved this for QV; you can basically use the fact that $M_tN_t - \langle M, N\rangle_t$ is a CLMG, so when you stop it, the result

$$M_t^T N_t^T - \langle M, N\rangle_{t\wedge T}$$

is again a CLMG. And then you use the characterization in (i) — by definition $M_t^T N_t^T - \langle M^T, N^T\rangle_t$ is a CLMG, and $\langle M^T, N^T\rangle$ is the unique process you subtract from $M_t^T N_t^T$ to get this. So that proves the first and third are equal.

Why are the first and second equal? Intuitively, when you stop *both* processes, you're saying that past T my processes are constant, so the QV should also be constant. What happens when I stop just *one* of these processes? Kind of hwat you should think of is after time T, your process basically looks like M_TN_t. And the point is if you write this as $M_T(N_t - N_T) + M_TN_T$ (you think of M_TN_T as some initial value at your stopping time; it's not a process in t), then $M_T(N_t - N_T)$ looks like a local martingale. (If T is a deterministic time, then this is the type of martingale we identified in the proof of the construction of QV — anything like this, when $T \le t$, is going to be a local martingale.) So basically what you're saying is that as a process, M_TN_t is certainly not going to be equal to M_TN_T; because after time T the latter is constant, while the former is a local martingale. But the bracket is precisely the thing you have to subtract to get a local martingale; so with $M_T(N_T - N_t)$ you shouldn't have to subtract *anything*. So that's why you'd believe $\langle M^T, N\rangle$ should be constant after T.

But then how would you prove this? The proof uses this discrete approximation — at the very least if you take T to be fixed, you can use the discrete approximation to prove this.

For any $t \geq 0$, we have that $S_n \to \langle M^T, N\rangle_t$ in probability, where we define

$$S_n = \sum_{i=1}^{p_n} (M_{t_i^n \wedge T} - M_{t_{i-1}^n \wedge T})(N_{t_i^n} - N_{t_{i-1}^n})$$

(we're just writing out what (iii) tells you in this case, where we have some increasing sequence of subdivisions of $[0, t]$). We also have that $S_n' \to \langle M^T, N^T\rangle_t$ in probability, where we analogously define

$$S_n' = \sum_{i=1}^{p_n} (M_{t_i^n \wedge T} - M_{t_{i-1}^n \wedge T})(N_{t_i^n \wedge T} - N_{t_{i-1}^n \wedge T})$$

(we're again just writing out property (iii), taking M^T and N^T as our local martingales).

Now on the event $\{T \geq t\}$, you have that these are exactly the same sums — because when $T \geq t$, the min $t_i^n \wedge T$ is always going to give you t_i^n. So on this event, you have $S_n = S_n'$ for all n.

And by the definition of convergence in probability, you have

$$\mathbb{P}[\left|S_n - \langle M^T, N\rangle_t\right| \geq \varepsilon,\, T \geq t] \to 0$$

as $n \to \infty$ (convergence in probability says that the first probability goes to 0; so certainly if we add in $T \geq t$, it still goes to 0). And smiilarly

$$\mathbb{P}[\left|S_n' - \langle M^T, N^T\rangle_t\right| \geq \varepsilon,\, T \geq t] \to 0.$$

So we then obtain that

$$\mathbb{P}[\left|\langle M^T, N\rangle_t - \langle M^T, N^T\rangle_t\right| \geq 2\varepsilon,\, T \geq t] = 0.$$

And since ε was arbitrary, we have that on the event $\{T \geq t\}$, we have $\langle M^T, N\rangle_t = \langle M^T, N^T\rangle_t$ almost surely.

All this is just saying the intuitive statement that if you're before your stopping time, then the processes $M_t^T N_t$ and $M_t^T N_t^T$ are the exact same (because the extra min with your stopping time isn't doing anything).

Now we have to consider the case where $T < t$. So now let's introduce two times $0 \leq s < t$; we're going to show that

$$\langle M^T, N\rangle_t - \langle M^T, N\rangle_s = 0 \text{ and } \langle M^T, N^T\rangle_t - \langle M^T, N^T\rangle_s = 0$$

almost surely, on the event that $\{T \leq s\}$. The proof of this claim is basically going through the same discrete approximation. You first take your sequence of meshes to always contain your fixed time s (and by definition they also contain your endpoint t). And now you go back to the discrete approximation. When you look at the discrete approximation, you see that

$$\langle M^T, N\rangle_t - \langle M^T, N\rangle_s = \lim \sum_{s \leq t_i^n} (M_{t_i^n \wedge T} - M_{t_{i-1}^n \wedge T})(N_{t_i^n} - N_{t_{i-1}^n})$$

(because I'm taking the discrete approximation for the entire interval, and the part before s gets cancelled out by what we're subtracting). And similarly, if you take the other process, you get exactly the same thing, except now you have a min with T — you get

$$\langle M^T, N^T\rangle_t - \langle M^T, N^T\rangle_s = \lim_{n \to \infty} \sum_{t_{i-1}^n \geq s} (M_{t_i^n \wedge T} - M_{t_{i-1}^n \wedge T})(N_{t_i^n \wedge T} - N_{t_{i-1}^n \wedge T}).$$

But the point is that on the event $\{T \leq s\}$, we have $t_i^n \wedge T = T$ and $t_{i-1}^n \wedge T = T$. Because I'm only summing over these points which are at least s, when I take this min, I just get t; so this second factor is just 0. This means I'm taking a limit in probability of 0; and that gives me that this is equal to 0.

And similarly, in the first case (with $\langle M^T, N\rangle_t$), we get the same thing for the first factor.

So that's why this claim is true, for any fixed $s < t$.

Now that we've proven this, we're basically done; you have to use some sort of density argument, which we'll now describe.

If we combine the first statement for when your stopping time is before your fixed time, and the second statement for when it's after your fixed time, you get that almost surely, for all rationals $q_1, q_2 \in \mathbb{Q}$ with $0 \le q_1 < q_2$ (we're just taking a countable union of these events as s and t vary over rationals), we have

$$\langle M^T, N\rangle_{q_2} = \langle M^T, N^T\rangle_{q_2} \quad \text{if } T \ge q_2$$

(in this case, it's the first statement for $\{T \ge t\}$, taking $t = q_2$). Otherwise, we have

$$\langle M^T, N\rangle_{q_2} = \langle M^T, N\rangle_{q_1} \quad \text{and} \quad \langle M^T, N^T\rangle_{q_2} = \langle M^T, N^T\rangle_{q_1} \quad \text{if } T \le q_1.$$

So we have all these identities on a dense set of times. And now you can just use continuity. If you take q_2 converging to some time less than or equal to T, you actually get that for *all* $t \le T$, you have

$$\langle M^T, N\rangle_t = \langle M^T, N^T\rangle_t$$

(you just replace q_2 by t by taking a limit and using continuity). And for all times $0 \le T \le s < t$, we have that

$$\langle M^T, N\rangle_t = \langle M^T, N\rangle_s \quad \text{and} \quad \langle M^T, N^T\rangle_t = \langle M^T, N^T\rangle_s$$

(i.e., the brackets individually are constant after this time). So now you can take $s = T$. We've proven these two brackets are equal for any time *before* your stopping time. Then for any time after your stopping time, you can use these two identities to get

$$\langle M^T, N\rangle_t = \langle M^T, N\rangle_T$$

(your bracket becomes constant after your stopping time — that's what we just showed). But for any time before your stopping time the two brackets are equal, so this is equal to $\langle M^T, N^T\rangle_T$; and again by the fact they become constant, this is $\langle M^T, N^T\rangle_t$. □

To summarize, you're using discrete approximation to show that at any time before the stopping time, the two brackets are equal; and after your stopping time, the bracket becomes constant (more precisely, for any $s < t$ it's equal at s and t). And once you've shown these using the discrete approximation, you can use this density argument to show that in fact the bracket is constant for all times after your stopping time.

Here's a consequence.

Proposition 11.9

Let B and B' be independent $(\mathcal{F}_t)$-Brownian motions. Then $\langle B, B'\rangle = 0$.

This kind of makes sense — if the BMs are independent, you should think their product is itself going to be a martingale.

Proof. We'll assume both B and B' start at 0 (the bracket doesn't care where you start — in the discrete approximation you're only looking at increments, so the starting time never enters, which means you can ignore it and assume you're starting at 0).

Approach 1 is you can just try to show that BB' is a martingale; then you can just use the uniqueness of the bracket. So as long as you can show BB' is a martingale, their bracket is 0.

What goes into showing this? You want to show that

$$\mathbb{E}[B_t B'_t \mid \mathcal{F}_s] = B_s B'_s.$$

And for this, you'd expand the Brownian motion into increments, and you'd try to use the independent increment property. The left-hand side is going to be

$$\mathbb{E}[(B_s + B_t - B_s)(B'_s + B'_t - B'_s) \mid \mathcal{F}_s].$$

And then when you expand out this product, you're going to get

$$B_s B'_s + \mathbb{E}[(B_t - B_s)(B'_t - B'_s) \mid \mathcal{F}_s]$$

(the cross-terms with B_s or B'_s times the other increment are going to be 0; but then you have this other term with the product of the two increments). So you want to say this last term is 0. But what you want to be careful of is for this, you need to say that $B_t - B_s$ and $B'_t - B'_s$ are independent *given* $\mathcal{F}_s$ — because you want to say this conditional expectation of products is a product of conditional expectations.

But this is not what we assumed. We assumed B and B' are independent, but we didn't assume they're conditionally independent given your filtration. So this is where the approach becomes problematic. It would probably work if your filtration was the canonical filtration of your BMs; but the problem is it doesn't have to be the case that independent BMs are conditionally independent given the filtration. This is why Le Gall does this other approach, because this one is kind of problematic.

So instead, you define this new process

$$X_t = \frac{1}{\sqrt{2}}(B_t + B'_t).$$

Why do you want this $1/\sqrt{2}$? As a distribution $X_t \sim \mathcal{N}(0, t)$ (by assumption B_t and B'_t are independent; and when you sum two independent Gaussians with this factor, you get exactly $\mathcal{N}(0, t)$). Moreover, X_t itself is a Brownian motion. But it's not necessarily a $(\mathcal{F}_t)$-Brownian motion (this is kind of the key point why the above failed) — it's a Brownian motion if you take the natural filtration of B and B', but it's not necessarily one with respect to the given filtration.

But the fact that it is a Brownian motion implies that

$$\langle X, X \rangle_t = t.$$

And now we can use bilinearity — you expand out the definition of X, and you get that

$$\frac{1}{2}\langle B, B \rangle_t + \frac{1}{2}\langle B', B' \rangle_t + \langle B, B' \rangle_t = t$$

(combining the two cross-terms using symmetry). But B and B' are BMs, so their QVs are just t; this implies $\langle B, B' \rangle = 0$. □

§11.4 Orthogonal martingales

We said you can basically think of the bracket as an inner product. In particular:

Definition 11.10. Given two continuous local martingales M and N, we say they're *orthogonal* if $\langle M, N \rangle = 0$, or equivalently if MN is a continuous local martingale.

(This means $\langle M, N\rangle$ is indistinguishable from 0 as a process.)

If M and N are orthogonal CLMGs bounded in L^2, then we have

$$\mathbb{E}[M_t N_t] = \mathbb{E}[M_0 N_0].$$

This is kind of a consequence of what we proved earlier — for any L^2 bounded martingale, you have this identity plus the expectation of the bracket. And here they're orthogonal, so the bracket is just 0; so you get this identity.

Even more, for any stopping time T, you have that the stopping time version of this identity is true — i.e.,

$$\mathbb{E}[M_T N_T] = \mathbb{E}[M_0 N_0].$$

For this you'd use the optional stopping theorem and the fact that $MN - \langle M, N\rangle$ is a UI martingale (when you have L^2 bounded martingales); and the bracket itself is 0, so you just apply the optional stopping theorem to MN.

So those are some remarks about orthogonal martingales. Next time we're going to prove some type of Cauchy–Schwarz inequality (called the Kunita–Watanabe theorem).

§12 March 12, 2025

§12.1 The Kunita–Watanabe theorem

As a quick recap of last time, we mostly defined the bracket of continuous local martingales and showed various properties; it's basically a bilinear extension of the quadratic variation. Continuing with this bracket, today we'll start out by proving the following theorem, which is basically a Cauchy–Schwarz type inequality for the bracket in terms of the QVs.

Proposition 12.1 (Kunita–Watanabe)

Let M and N be continuous local martingales and let H and K be measurable processes. Then almost surely,

$$\int_0^\infty |H_s|\,|K_s|\,|d\langle M, N\rangle_s| \le \left(\int_0^\infty H_s^2\, d\langle M, M\rangle_s\right)^{1/2} \cdot \left(\int_0^\infty K_s^2\, d\langle N, N\rangle_s\right)^{1/2}.$$

Recall the bracket is always a FV process; last week or two weeks ago we discussed integration with respect to FV processes. The assumption is M and N are CLMGs, and H and K are measurable. Before in the section on FV processes, we kind of assumed that the integrand is going to be progressive; you needed this progressive assumption to say the integral itself is an adapted process (that was why you needed it). Here we're not necessarily assuming progressivity, because this is just a statement about integrals — so you can assume just measurability (H is a process, so it's a function of t and ω, and we're asking it to be measurable with respect to the joint σ-algebra; progressivity has the additional assumption that on $[0, t] \times \Omega$, you're measurable with respect to the product σ-algebra involving $\mathcal{F}_t$).

This estimate basically looks like Cauchy–Schwarz, and indeed the proof is basically several applications of Cauchy–Schwarz. The main thing is why you can replace $d\langle M, N\rangle_s$ by the QV, and the reason you can do that is basically Cauchy–Schwarz; you'll see this if you go to the discrete approximation of the bracket, which is what we'll do.

Proof. Only for this proof, we'll use the notation

$$\langle M, N\rangle_s^t = \langle M, N\rangle_t - \langle M, N\rangle_s.$$

By the discrete approximations, we can express

$$\langle M, N\rangle_s^t = \lim_{n\to\infty} \sum_{i=1}^{p_n} (M_{t_i^n} - M_{t_{i-1}^n})(N_{t_i^n} - N_{t_{i-1}^n})$$

(where the limit is in probability, and your subdivisions are of $[s,t]$ instead of $[0,t]$). But for each fixed n, let's call this thing S_n; then you can just apply Cauchy–Schwarz to bound this sum by the corresponding discrete approximations of the QVs — we have

$$S_n \le \left(\sum_{i=1}^{p_n} (M_{t_i^n} - M_{t_{i-1}^n})^2\right)^{1/2} \left(\sum_{i=1}^{p_n} (N_{t_i^n} - N_{t_{i-1}^n}^n)^2\right)^{1/2}.$$

And as $n \to \infty$, th right-hand side converges in probability to $(\langle M, M\rangle_s^t)^{1/2}(\langle N, N\rangle_s^t)^{1/2}$. So from this, you obtain that for all $s < t$, almost surely we have that

$$\langle M, N\rangle_s^t \le (\langle M, M\rangle_s^t)^{1/2}(\langle N, N\rangle_s^t)^{1/2}.$$

Then by taking a countable dense subset of s and t (e.g., the rationals), we can take the 'almost surely' outside and say that almost surely, the above holds for all rational $s < t$.

Assume this event occurs. Basically we're just going to take an outcome ω of our sample space where this is true for all rational $s < t$. The remaining part of the argument is just deterministic — there's no more probability, it's just some real analysis.

First of all, by continuity, this is going to be true for all $s < t$, not just rationals.

Claim 12.2 — For all $s < t$, we have

$$\int_s^t |d\langle M, N\rangle_u| \, du \le (\langle M, M\rangle_s^t)^{1/2}(\langle N, N\rangle_s^t)^{1/2}.$$

When we take this absolute value, we're integrating the total variation measure — recall any FV process decomposes as a difference of two positive measures, and the total variation takes the sum. (We could have written the right-hand side as an integral with the QV instead of the bracket — this is an increasing process, so the total variation is the same as the process itself.)

Proof. Why is this true? We had a lemma one or two weeks ago saying that the total variation is the sup over discrete approximations. So now we can go back to the discrete approximation — again we take some subdivision of points, and we get

$$\sum_{i=1}^{p_n} \left|\langle M, N\rangle_{t_{i-1}}^{t_i}\right|$$

(and the left-hand side is a sup over all possible subdivisions giving you these sums). First, we apply the earlier inequality to replace each of these by the right-hand side, so this is at most

$$\sum_{i=1}^{p_n} (\langle M, M\rangle_{t_{i-1}}^{t_i})^{1/2}(\langle N, N\rangle_{t_{i-1}}^{t_i})^{1/2}.$$

(To be precise, we should have put absolute values on the S_n before, but that inequality is still true.) Then you apply Cauchy–Schwarz and get that this is at most

$$\left(\sum_{i=1}^{p_n} \langle M, M\rangle_{t_{i-1}}^{t_i}\right)^{1/2} \left(\sum_{i=1}^{p_n} \langle N, N\rangle_{t_{i-1}}^{t_i}\right)^{1/2}.$$

And these things are telescoping sums, so this is exactly going to be

$$(\langle M, M\rangle_s^t)^{1/2}(\langle N, N\rangle_s^t)^{1/2}.$$

(Then we take a sup over all meshes and get this claim.) □

We're working towards an approximation argument — you think of H as a general measurable function, and so far we've proven the claim for when H and K are indicator functions on intervals. Now we want to go from indicator functions on intervals to indicator functions on general Borel sets (eventually we'll want to approximate H and K by sums of indicators on general Borel sets — that's usually how you approximate measurable functions). The problem is you can't do this with indicator functions of intervals, so the next step is to go from indicators of intervals to indicators of Borel sets.

Claim 12.3 — For any bounded Borel set $A \subseteq \mathbb{R}_+$, we have

$$\int_A |d\langle M, N\rangle_u| \le \left(\int_A d\langle M, M\rangle_u\right)^{1/2} \left(\int_A d\langle N, N\rangle_u\right)^{1/2}.$$

(This is the same thing, except now we're integrating over A instead of $[s, t]$.)

Proof. For this, you have to do some abstract measure theory stuff. You want to eventually apply some sort of monotone class theorem, for instance. The first step, which we've already proved, is that it holds when $A = [s, t]$ is just an interval (that's precisely what the previous claim says).

The second step is that it also holds when A is a finite disjoint union of intervals. Why is that true? Well, it's another application of Cauchy–Schwarz — if you have a finite disjoint union of intervals, then the left-hand side is the sum over each of these disjoint intervals, and you apply Cauchy–Schwarz (exactly as in the previous step, but using the inequality for a single interval), and you basically get this result. So the argument is basically the same manipulation as in the proof of the previous claim.

And you might recall that the set of all finite disjoint unions of intervals — for example, when you construct the Lebesgue measure on $\mathbb{R}$, you first get a pre-measure by assigning a measure to finite disjoint unions of intervals (by summing the lengths); and then Caratheodory's extension theorem says that you can extend it to the set of Borel sets. The point is that this set of finite disjoint unions of intervals is an 'algebra,' and because of that you'll be able to apply some sort of monotone class theorem. You have to choose the right one — you want to say if I have this identity for a sequence of sets A (which is increasing), then it also holds for their limit (and also if the sets are decreasing). Here's the particular one we need:

Theorem 12.4

Suppose we have a collection of sets $\mathcal{C}$ such that:

(a) For any increasing collection of sets $\{A_n\} \subseteq \mathcal{C}$ with $A_n \uparrow A$, we have $A \in \mathcal{C}$.

(b) For any countable decreasing collection of sets $\{A_n\} \subseteq \mathcal{C}$ with $A_n \downarrow A$, we have $A \in \mathcal{C}$.

(c) $\mathcal{C}$ contains an algebra $\mathcal{A}$ of sets.

Then $\mathcal{C}$ contains $\sigma(\mathcal{A})$.

The first property (a) here is true by monotone convergence (these things are all integrals with respect to positive measures). For (b) you basically want to apply monotone convergence but for decreasing functions; you can apply Fatou's lemma on the left-hand side, and some sort of monotone convergence on the right. (The only way monotone convergence fails for decreasing functions is if you're not integrable, meaning the thing is always ∞; but if the thing is always ∞ then the bound is true. If the right-hand side is always ∞

then you're done; and if it's finite for some A_n then it's finite for all the ensuing ones, and then you can even apply dominated convergence.)

So you have these three assumptions on your collection $\mathcal{C}$, and the theorem says that then it contains the generated σ-algebra of $\mathcal{A}$.

We already said why (a) and (b) are true; and (c) is true taking $\mathcal{A}$ to be the algebra of finite disjoint unions of intervals (it's an algebra because it's closed under finite intersections and unions, and complements if you restrict to a bounded interval).

So if we apply this monotone class argument, w get that $\mathcal{C} \supseteq \mathcal{B}([0,T])$ for any T; and that basically means it contains any bounded Borel set of $\mathbb{R}_+$. That proves the claim. □

Now that we've done that, the next thing is to consider sums of indicator functions.

Claim 12.5 — Suppose that $h = \sum_{i=1}^p \lambda_i \mathbf{1}_{A_i}$ and $k = \sum_{i=1}^p \mu_i \mathbf{1}_{A_i}$ are nonnegative, and the A_i are disjoint Borel sets such that $\bigcup A_i = [0,K]$. Then the statement is true for h and k.

Proof. We can always assume nonnegativity, because the statements are about absolute values of our functions. So we fix some large K and partition this interval into large Borel sets, and look at indicator functions based on this partition. Then we want to prove the theorem for this particular h and k; but this is basically Cauchy–Schwarz again. We can write this integral as

$$\sum_i \lambda_i \mu_i \int_{A_i} |d\langle M, N\rangle_u| ,$$

and then apply the earlier inequality and Cauchy–Schwarz — applying the previous inequality gives

$$\sum_{i=1}^p \lambda_i \mu_i \left(\int_{A_i} d\langle M, M\rangle_u\right)^{1/2} \left(\int_{A_i} d\langle N, N\rangle_u\right)^{1/2}.$$

And then I can apply Cauchy–Schwarz, and I end up with

$$\left(\int h^2 \, d\langle M, M\rangle_u\right)^{1/2} \left(\int k^2 \, d\langle N, N\rangle_u\right)^{1/2}$$

(we get λ_i^2 and μ_i^2, which is why we have h^2 and k^2). □

Finally, we finish by approximating H and K by sums of indicator functions $\{h_n\}$ and $\{k_n\}$. You can definitely approximate a general measurable function H by a sequence of sums of indicators converging up to H, and similarly with K. But why can you ensure they're indicators on the same sets A_i? You can do that by taking a common refinement. So that's the one thing you need to note; but then you can finish by using monotone convergence. □

Somewhere when constructing the stochastic integral, we'll need to apply this.

§12.2 Continuous semimartingales

So in summary we discussed FV processes and CLMGs; now we can finally define what a continuous semimartingale is. It's basically a sum of a FV process and local martingale.

Definition 12.6. A process X is a *continuous semimartingale* (CSMG) if $X_t = M_t + A_t$ where (M_t) is a continuous local martingale and (A_t) is a finite variation process.

Of course this is all relative to a filtration, which we assume is fixed (the notions of martingale and FV processes rely on that — they have to be adapted by definition).

One thing to note is that:

Fact 12.7 — The decomposition $X_t = M_t + A_t$ is unique up to indistinguishability.

This is just because again, you can't have a process be both FV and a CLMG. If you had two different decompositions $M_t^1 + A_t^1 = M_t^2 + A_t^2$, then you could rearrange to get

$$M_t^1 - M_t^2 = A_t^2 - A_t^1.$$

So this difference would be both a CLMG and a FV process, meaning it'd have to be 0 (up to indistinguishability).

You can extend the notion of brackets to these CSMGs.

(We call this the *canonical decomposition* of a CSMG; it's canonical because it's unique.)

Definition 12.8. Let $X = M + A$ and $Y = M' + A'$ be continuous semimartingales. Then the *bracket* of X and Y is defined as

$$\langle X, Y\rangle = \langle M, M'\rangle.$$

In particular, the quadratic variation $\langle X, X\rangle$ is just $\langle M, M\rangle$.

In some sense, this is very intuitive for the heuristic reasons we mentioned last week — recall that intuitively you think of the martingale as like a Brownian motion, so infinitesimally dM_t varies like $\sqrt{dt}$, while the FV part varies like $dA_t \sim dt$. And the whole reason you get the quadratic variation is because you vary like $\sqrt{dt}$ — infinitesimally the QV looks like $dM_t\, dM_t$ (you look at the square of your martingale increment), which varies like dt. Anything on the order of dt, once I integrate I have to care about it. But we actually have $dA_t\, dA_t \sim (dt)^2$, which means if I integrate it I don't have to care about it. Similarly $dM_t\, dA_t \sim (dt)^{3/2}$, so I also don't care about it. So that's why you'd want to define your bracket in this way — if you just assume (which will be true) that it's bilinear, then when you plug in these canonical decompositions and expand out, you get the martingale part plus stuff that look like $dM\, dA$ or $dA\, dA$, which are both basically 0. So you should really just see this martingale part.

And indeed, we're going to show that the bracket of these semimartingales also has a discrete approximation, and it's precisely the one you'd expect.

Proposition 12.9

Let $0 = t_0^n < \cdots < t_{p_n}^n = t$ be an increasing sequence of subdivisions of $[0, t]$ with mesh size tending to 0. Then we have

$$\sum_{i=1}^{p_n} (X_{t_i^n} - X_{t_{i-1}^n})(Y_{t_i^n} - Y_{t_{i-1}^n}) \to \langle X, Y\rangle_t.$$

So it's the same statement as for the bracket of CLMGs. The reason is basically this heuristic. By definition the RHS is the bracket of the martingale parts. And we can expand out each thing on the left-hand side using the decomposition. We'll get the martingale part times the martingale part, which by the decomposition is going to converge to $\langle X, Y\rangle_t$; and the remainder will be lower-order by the above heuristic.

Proof. To simplify notation we'll assume $X = Y$; the proof works the same in the general case (or you should be able to deduce it from just the case $X = Y$, because we defined the bracket of CLMGs based on the $X = Y$ case).

Then when we expand out hte sum, we get

$$\sum_{i=1}^{p_n}(M_{t_i^n} - M_{t_{i-1}^n})^2 + 2\sum_{i=1}^{p_n}(M_{t_i^n} - M_{t_{i-1}^n})(A_{t_i^n} - A_{t_{i-1}^n}) + \sum_{i=1}^{p_n}(A_{t_i^n} - A_{t_{i-1}^n})^2.$$

The first piece converges to the right-hand side (by our discrete approximation result from last time); so we just want to say the last two pieces converge to 0. And that's basically our heuristics from before.

Let's look at the middle term (this should be larger than the last term by our heuristics, and the proof will also show why the last term converges to 0).

We know A is finite variation, so if I ignore the martingale and take absolute values, the sum of that thing is going to be controlled as $n \to \infty$. So the idea is I want to bound

$$|S_n| \le \max_i \left|M_{t_i^n} - M_{t_{i-1}^n}\right| \cdot \sum_{i=1}^{p_n}\left|A_{t_i^n} - A_{t_{i-1}^n}\right|.$$

And the last term is at most $\int_0^t |dA_s|$, which is finite. And the first term converges to 0 in probability — actually it converges to 0 surely, because the mesh size is tending to 0 and M is continuous, so it's uniformly continuous on any compact interval, which means this max of increments goes to 0. So we get $S_n \to 0$ in probability.

And you can say something similar for the last thing. (We didn't actually have to make the earlier heuristic quantitative — we didn't use the fact that M varies like $\sqrt{t}$ on small intervals, even t^ε would have worked.) □

CSMGs are going to be both the integrators and integrands for stochastic integrals — we'll generate CSMGs with respect to CSMGs. But if you want, you should think of BM as your integrator — integrating things with respect to BM is good enough for most applications (but Le Gall develops the theory in general).

§12.3 Stochastic integration

So now we finally come to the whole point of this course, which is stochastic integration. Up to this point, we've been developing skills — kind of like if you play football, you have to spend some time in the weight room or else there's no way you can actually be a professional football player. But you might not care as much about the weight room and just want to play football. Or if you're a musician you have to spend a bunch of time developing your technique, not just playing the pieces you want. We spent the first 4–5 weeks developing the basic foundational material, though that may not be why you wanted to take this class. But that's needed because it'll be used all over the place when we define these stochastic integrals.

But now we're coming to the main part of the course. In some sense, defining the stochastic integral is also some technical thing, and after that we can start playing around with these stochastic integrals and Ito's formula.

We're going to assume throughout in the background that we have a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{P})$, and also that the filtration is complete. So far, the place we needed completeness — or the main place — is when we constructed QV, the fact that you can guarantee it's always continuous (not just almost surely continuous, but it is continuous for every ω) needed completeness, because we wanted to modify it on a set of measure 0 and we wanted that to preserve adaptedness.

As with any integral, you always define it on special cases first, so let's do that.

§12.4 The Hilbert space of L^2 bounded martingales

The frist thing we'll do is define integrals for L^2-bounded martingales. This will take some effort, but then the extension to local martingales is just abstract measure theory arguments — you localize your CLMG so that it's bounded in L^2, and then get the integral, and then you want to remove the localization and stitch these things together. So this is actually the main thing.

For some notation:

Notation 12.10. Let $\mathbb{H}^2$ be the space of L^2-bounded (continuous) martingales with $M_0 = 0$.

(All martingales in this course are continuous. We also identify indistinguishable processes — we treat any two processes as the same if they're indistinguishable.)

Equivalently, a martingale M is in $\mathbb{H}^2$ if and only if M is a continuous local martingale starting at 0 and such that $\mathbb{E}[\langle M, M\rangle_\infty] < \infty$ (this is something we proved last class — recall that the implication from the right to left is just applying the L^2 maximal inequality in the right place).

Why are we defining this? The way you'll construct this integral is basically by functional analysis. You're going to define an isometry on a Hilbert space. And this is going to be our Hilbert space.

We need to check that it's actually a Hilbert space: in particular, it has to have an inner product, and it has to be complete under that inner product. We'll work towards that.

But first, for some more preliminaries, because M is bounded in L^2 it's also UI. This means it's closed, so you can write it as a Doob martingale — we have

$$M_t = \mathbb{E}[M_\infty \mid \mathcal{F}_t],$$

and this converges in L^2 and almost surely to M_∞. (That's another fact about L^2 bounded martingales.)

Also recall that if M and N are L^2-bounded martingales (i.e., $M, N \in \mathbb{H}^2$), then $\langle M, N\rangle_\infty$ is well-defined (it's the limit of brackets at finite times), and it's in L^1 (this is another thing we proved last time).

So we can define a symmetric bilinear form on $\mathbb{H}^2$:

Definition 12.11. We define $\langle M, N\rangle_{\mathbb{H}^2} = \mathbb{E}[\langle M, N\rangle_\infty] = \mathbb{E}[M_\infty N_\infty]$.

We said you think of $\langle -, -\rangle$ as a bilinear form on the space of local martingales; if you restrict to L^2-bounded martingales, it's actually a bilinear form.

Recall the last identity is true because the bracket is such that if you take $M_t N_t - \langle M, N\rangle_t$ you get a UI martingale; and then you use the optional stopping theorem.

So we've defined this bilinear form. In order for it to be an inner product, it needs to first of all be positive definite. In other words, we want to say:

Claim 12.12 — If $\langle M, M\rangle_{\mathbb{H}^2} = 0$, then $M = 0$ (up to indistinguishability).

Why is this true? Well, the condition implies that $\mathbb{E}[\langle M, M\rangle_\infty] = 0$ by definition. And this implies M itself is 0. We showed this last time; the way you see it is if the QV is 0, that means M^2 is a local martingale. But now you have a nonnegative local martingale, so it's a supermartingale; that means $\mathbb{E}[M_t^2] = 0$ for any t, which implies this.

So it satisfies this condition of an inner product. And it's symmetric and bilinear, so all that is fine.

The main thing now to verify — a Hilbert space is an inner product space which is *complete*.

Proposition 12.13

The space $\mathbb{H}^2$ equipped with this inner product is a Hilbert space.

Proof. For the proof, we need to show why it's complete. By definition, completeness means that any Cauchy sequence converges. So let's take a Cauchy sequence $(M^n) \subseteq \mathbb{H}^2$. We need to construct a continuous L^2-bounded martingale such that M^n converges to that L^2-bounded martingale.

Let's start with what a Cauchy sequence means; by definition we get

$$\lim_{N\to\infty} \sup_{\min\{m,n\}\geq N} \langle M^m - M^n, M^m - M^n\rangle_{\mathbb{H}^2} = 0$$

(just writing out the definition of Cauchy in this particular case). By the definition of the norm, this implies that

$$\sup_{m,n\geq N} \mathbb{E}[(M_\infty^n - M_\infty^m)^2] \to 0$$

as $N \to \infty$ (we just wrote the limiting statement slightly differently, and use teh fact that $\mathbb{E}[\langle M, N\rangle_\infty] = \mathbb{E}[M_\infty N_\infty]$).

The first thing we want to note is that this condition here implies that the endpoints (M_∞^n) themselves are Cauchy in $L^2(\Omega, \mathcal{F}_\infty, \mathbb{P})$ — these endpoints are always $\mathcal{F}_\infty$-measurable by assumption, and this is exactly the statement that they're Cauchy in L^2. This means there exists $Z \in L^2(\Omega, \mathcal{F}_\infty, \mathbb{P})$ such that $M_\infty^n \to Z$ in L^2. So at least we extracted a limit for the endpoints.

Now we have to use this endpoint limit to construct a martingale for which the sequence of martingales themselves converge. The natural candidate limit is to take the Doob martingale defined by Z, i.e.,

$$M_t = \mathbb{E}[Z \mid \mathcal{F}_t].$$

This is a martingale, and it's L^2 bounded because $Z \in L^2$. And basically, the faact that the sequence of martingales M^n converges to that martingale under this inner product is true by the fact that $M_\infty^n \to Z$. There's just one slight problem — this martingale is not necessarily continuous. So actually, it's not necessarily in H because it might not have continuous sample paths.

So you want to find a modification of this which is continuous. That's where we're going to take a subsequence — take a subsequence $\{n_k\}$ such that the expectation of consecutive differences along this subsequence is exponentially decaying, i.e.,

$$\mathbb{E}[(M_\infty^{n_{k+1}} - M_\infty^{n_k})^2] \leq 2^{-k}.$$

Now we can use the L^2 maximal inequality, which will imply that

$$\mathbb{E}\left[\sup_t (M_t^{n_{k+1}} - M_t^{n_k})^2\right] \leq 4 \cdot 2^{-k}.$$

And so thus we're going to get that

$$\mathbb{E}\left[\sum_{k=1}^\infty \sup_{t\geq 0} \left|M_t^{n_{k+1}} - M_t^{n_k}\right|\right] < \infty$$

(we lost the square, but you can use Hölder and bound the thing without the square by $\sqrt{4 \cdot 2^{-k}}$, which is still going to be summable). That implies almost surely,

$$\sum_{k=1}^\infty \sup_{t\geq 0} \left|M_t^{n_{k+1}} - M_t^{n_k}\right| < \infty.$$

And that means, at least on this almost sure event, the sequence (M^{n_k}) is Cauchy in the space $\mathcal{C}([0,\infty),\mathbb{R})$ under the sup norm. And then you can use the fact that this space with the sup norm is complete — by the completeness of $\mathcal{C}([0,\infty),\mathbb{R})$, we get that M^{n_k} converges to some limit, which we'll call $\widetilde{M}$, on an almost sure event E (which is the event that our infinite series from before is finite).

Then you can define your martingale M (the one you want to eventually construct) to be $M = \mathbf{1}_E \widetilde{M}$ (so on the event E it's $\widetilde{M}$, and off the event E it's just 0). This ensures it has continuous sample paths. And M is adapted because $(\mathcal{F}_t)$ is complete (E depends on your process at all times, so you need completeness to say $\mathbf{1}_E$ is $\mathcal{F}_0$-measurable).

So we've exhibited this limit M in the sense of the sup norm on continuous functions. To finish, we've obtained M as a subsequential limit of the M^ns. And we obtained Z as a limit of the entire sequence. But subsequential limits have to be the same as sequential limits; that means M_t itself is going to be a martingale, with $M_t = \mathbb{E}[Z \mid \mathcal{F}_t]$ almost surely. For example, this is because M is an almost sure limit of the M^{n_k}s. And those are martingales, so you can write

$$M_t = \lim_k M_t^{n_k} = \lim_k \mathbb{E}[M_\infty^{n_k} \mid \mathcal{F}_t].$$

And if these endpoints converge to Z in L^2, then the conditional expectations also converge in L^2, and thus in probability.

So this verifies M is a martingale; and it's continuous and bounded in L^2, so it's an element of $\mathbb{H}^2$.

Finally, we just need to say why our original sequence M^n converges to M. But this is just because

$$\langle M^n - M, M^n - M\rangle_{\mathbb{H}^2} = \mathbb{E}[(M_\infty^n - Z)^2],$$

because $M_\infty = Z$ almost surely. But then by assumption on Z, we precisely defined Z as the L^2 limit of the M_∞^n, so this goes to 0. □

The main points here are again using the L^2 maximal inequality to upgrade the convergence of the endpoint to convergence of the sup of all your processes. Then completeness basically comes from the completeness of L^2 and $\mathcal{C}^0$ (the space of continuous functions); and you do the usual tricks with subsequences (any time you want a continuous process, you usually construct it as a uniform limit of continuous processes).

Now let's set some more notation.

Notation 12.14. Let $\mathcal{P}$ be the progressive σ-field on $\Omega \times \mathbb{R}_+$.

Recall that one way of characterizing the progressive σ-field is the smallest σ-algebra with respect to which all progressive processes are measurable. It's also the collection

$$\mathcal{P} = \{A \subseteq \Omega \times \mathbb{R}_+ \mid (\omega,t) = \mathbf{1}_A(\omega,t) \text{ is progressive}\}.$$

There was some homework problem about proving equivalent ways of thinking about this σ-algebra.

Definition 12.15. For $M \in \mathbb{H}^2$, let $L^2(M)$ be the set f progressive processes H such that

$$\mathbb{E}\left[\int_0^\infty H_s^2 \, d\langle M, M\rangle_s\right] < 0.$$

This is basically the L^2 norm of H with respect to the quadratic variation of M.

We're going to identify processes H and H' which are such that $H_s = H'_s$ almost everywhere with respect to $d\langle M, M\rangle_s$ almost surely — H is a function of two variables ω and s. An easier way to think about this is this means

$$\mathbb{E}\left[\int_0^\infty (H_s - H'_s)^2 \, d\langle M, M\rangle_s\right] = 0.$$

The reason you want to identify such processes is that then you can view $L^2(M)$ as an actual L^2 space — it's a Hilbert space

$$L^2(\Omega \times \mathbb{R}_+, \mathcal{P}, d\mathbb{P}\, d\langle M, M\rangle).$$

What we're saying is on $\Omega \times \mathbb{R}_+$, we have a σ-algebra $\mathcal{P}$. And we can define a measure on this measure space which maps a given $A \in \mathcal{P}$ to

$$\mathbb{E}\left[\int_0^\infty \mathbf{1}_A(\omega, s)\, d\langle M, M\rangle_s\right].$$

One verifies that this actually defines a measure (that's just monotone convergence — you have to say why it's countably additive). With this definition, if you integrate H^2, it's precisely going to give you the thing $\mathbb{E}[\int_0^\infty H_s^2\, d\langle M, M\rangle_s]$. (That's an exercise by the standard machine thing where if this is your measure, then when you integrate functions you get precisely something like $\mathbb{E}[\int_0^\infty H_s^2\, d\langle M, M\rangle_s]$.)

And the inner product on this space is just the inner product of a L^2 space — we deifne

$$\langle H, K\rangle_{L^2(M)} = \mathbb{E}\left[\int_0^\infty H_s K_s\, d\langle M, M\rangle_s\right]$$

(as the integral of HK against your measure).

We're first going to construct a stochastic integral where the integrand is H, and you're going to integrate H with respect to your L^2-bounded martingale M. We're basically going to define a map from this Hilbert space $L^2(M)$ to $\mathbb{H}^2$, the space of L^2-bounded continuous martingales. This map, by the way you define it, is going to have a natural interpretation as a stochastic integral.

§12.5 Integrals for simple processes

We're going to define this integral as an isometry, but the way you get started is by defining it on simple processes (as with any integral). What are the classes of simple processes we want to work with?

Definition 12.16. An *elementary process* is a progressive process H of the form

$$H_s(\omega) = \sum_{i=1}^{p-1} H_i(\omega)\mathbf{1}_{(t_i, t_{i+1}]}(s)$$

where $H_i(\omega)$ is bounded and $\mathcal{F}_{t_i}$-measurable for all i, and $0 = t_0 < \cdots < t_-$.

So you have indicators of intervals times some random variables; and this random variable is measurable with respect to the *left* endpoint. The fact you take the left endpoint is needed to make the theory work. It's somehow related to how when we constructed the QV, we found this martingale, and it was only a martingale because it looked like a difference of your martingale times some function of the left endpoint. That's intimately related to why you want to take the left endpoint here.

Notation 12.17. We write $\mathcal{E}$ for the set of all elementary processes.

First of all, $\mathcal{E} \subseteq L^2(M)$ for all $M \in \mathbb{H}^2$. Why? Well, the H_is are bounded by some deterministic constant, so the finiteness condition in the definition of $L^2(M)$ is always true (also because you're taking a finite set of points).

Basically what we're going to do is for such processes, there's kind of a natural way to define this integral. So let's talk about that in the last five minutes.

How would you define some formal symbol like $\int H_s\,dM_s$ for a process of the above form? Well, you just write it as a finite sum of increments

$$\int H_s\,dM_s = \sum H_i(M_{t_{i+1}} - M_{t_i}).$$

this is maybe the integral over the whole real line $[0,\infty)$. If you want to think about this as a process in an additional time parameter t, you'd simply take a min with t in each of these times — we define

$$\int_0^t H_s\,dM_s = \sum H_i(M_{t_{i+1}\wedge t} - M_{t_i\wedge t}).$$

This looks very much like the proof of the QV; in that proof H itself was $M_{t_i\wedge t}$. But as long as it's $\mathcal{F}_{t_i\wedge t}$-measurable, this thing is going to be a martingale.

So this defines $H\cdot M = ((H\cdot M)_t)$. And we're going to check that this itself is in $\mathbb{H}^2$. If you believe what we just said, it's true — it's a martingale, the H's themselves are bounded, and M is in $\mathbb{H}^2$ — so this is believable.

So you're going to have a map $\mathcal{E}\to\mathbb{H}^2$, where you think of $\mathcal{E}$ as contained in $L^2(M)$. So you're fixing your M, and you define this map $H\mapsto H\cdot M$.

The thing you need to check is:

Proposition 12.18

The map $H\mapsto H\cdot M$ is an isometry from $\mathcal{E}$ to its image, i.e., it preserves norms.

Second, it's natural what the approach is — you want to extend your map from the subspace $\mathcal{E}$ to the whole space $L^2(M)$. You can do that if your subspace is actually dense. So you want to show:

Proposition 12.19

The space $\mathcal{E}$ is dense in $L^2(M)$.

Once you show that, you're basically done — you have a densely defined isometry from a subspace of one Hilbert space into another, and you can extend it using abstract Hilbert space arguments. That's what the stochastic integral will be.

There are some things to check. First, in principal your elementary process could have different representations. This is true when you construct Lebesgue integrals as well — you have to check your representation is independent of the particular sum you represented it as. Also you have to check that this is a linear map. But this is kind of the roadmap for constructing the stochastic integral.

§13 March 17, 2025

§13.1 Elementary processes

To review what we talked about at the end of last time, we're getting started towards defining the stochastic integral. The first thing you do is define it for indicator functions, more or less. In this setting, you want them to be of the following form:

Definition 13.1. An *elementary process* H is a progressive process of the form

$$H = \sum_{i=0}^{p-1} H_{(i)} \mathbf{1}_{(t_i, t_{i+1}]},$$

where $0 = t_0 < \cdots < t_p$ and $H_{(i)}$ is bounded and $\mathcal{F}_{t_i}$-measurable. We use $\mathcal{E}$ to denote the set of elementary processes.

So here you have an indicator of an interval, and a random variable which is measurable with respect to the *left* endpoint. This is kind of a very special thing that allows you to define the stochastic integral. It's very important that $H_{(i)}$ is $\mathcal{F}_{t_i}$-measurable and not just $\mathcal{F}_{t_{i+1}}$-measurable — this is going to give you a L^2 martingale, which is exactly what'll give you the isometry property that we'll prove today (that's how you define the integral, by an isometry).

Remark 13.2. For any $M \in \mathbb{H}^2$ (recall that $\mathbb{H}^2$ is the set of L^2-bounded continuous martingales), $\mathcal{E}$ is a linear subspace of $L^2(M)$ (the set of continuous processes which are square-integrable with respect to $\langle M, M \rangle$).

§13.2 Density of elementary functions

We had an outline last time of how we'll define the stochastic integral. The first thing we'll do is prove that $\mathcal{E}$ is dense (as you usually expect of indicator functions — usually you expect to be able to approximate any measurable function by indicators, and here we can approximate any function in $L^2(M)$ by an elementary process).

Proposition 13.3

For all $M \in \mathbb{H}^2$, the set $\mathcal{E}$ is dense in $L^2(M)$.

Proof. To prove a set is dense (this is just abstract Hilbert space theory), it suffices to show that if you have $K \in L^2(M)$ which is orthogonal to every element in $\mathcal{E}$ (i.e., every elementary process) — with respect to the inner product on $L^2(M)$ that we defined last time — then $K = 0$ in $L^2(M)$.

Remark 13.4. For a set to be dense in a Hilbert space, it suffices to have this property. Intuitively, the reason this property characterizes denseness is that it's a vector space, so if the set is not dense, then it basically has to look like a hyperplane (when you take the closure). Then you can take a vector orthogonal to everything in your set (which is not 0).

So if the set isn't dense, this property shouldn't be true; contrapositively, if this property is true, then the set has to be dense.

So assume this orthogonality. We'll define a process

$$X_t = \int_0^t K_s \, d\langle M, M \rangle_s$$

(so X is the integral of K against the QV of M). Our goal is to show that X is both finite variation and a martingale. Then we can use the result that if you're both, you have to be 0 (in the sense that almost surely, $X_t = 0$ for all t). And then this will imply that $K = 0$ (this is somewhat believable, though we might say a couple more words at the end — that if you integrate K and get something that's zero, then K had to be zero, at least almost everywhere with respect to this measure).

The fact that X is a FV process just comes from the fact that $K \in L^2(M)$. The fact that it's actually a martingale is the key thing, and that comes from the orthogonality assumption. You're assuming K is orthogonal to every elementary process, and the key observation is that implies you have a martingale; so we'll see that.

But first, let's show why X is a finite variation process. First, why is this integral even well-defined? You think of $d\langle M, M\rangle_s$ as a (positive) measure; but why is this integral well-defined? Let's bound

$$\mathbb{E}\left[\int_0^t |K_s| \, d\langle M, M\rangle_s\right].$$

We'll apply Cauchy–Schwarz — you want to do this because you know $K \in L^2$, so you want K^2. If we do that (applying Cauchy–Schwarz to the integral, before taking the expectation), we get that this is at most

$$\mathbb{E}\left[\left(\int_0^t K_s^2 \, d\langle M, M\rangle_s\right)\left(\int_0^t \langle M, M\rangle_s\right)^{1/2}\right].$$

And then you can apply Cauchy–Schwarz again, but this time to the expectation, to separate them; and you get

$$\mathbb{E}\left[\int_0^t K_s^2 \, d\langle M, M\rangle_s\right]^{1/2} \cdot \mathbb{E}\left[\langle M, M\rangle_t\right]^{1/2}.$$

The first term is finite because $K \in L^2(M)$, and the second is finite because M is a L^2-bounded martingale. So almost surely, this thing is finite for all t; and we had a result a few weeks ago that when you have this condition that the integral is finite, that defines for you a finite variation process. (This is where you have to use completeness — the expectation being finite just means that this process is finite *almost surely*, so you might have to modify it on measure-0 sets.)

So that shows X is going to be finite variation; now it remains to show X is a martingale. This is the key step in the proof — that the assumption of orthogonality implies X is going to be a martingale.

Before this, another thing is that for X to be a martingale, we need $X_t \in L^1$ for all t. But that's just a consequence of the previous discussion; so the integrability condition is satisfied.

Now to verify the martingale property, we want to show

$$\mathbb{E}[X_t \mid \mathcal{F}_s] = X_s.$$

Another way you can put this is that for all bounded $\mathcal{F}_s$-measurable random variables F, we should have

$$\mathbb{E}[F(X_t - X_s)] = 0.$$

If you had this relation, that would imply $\mathbb{E}[X_t \mid \mathcal{F}_s]$ is exactly given by X_s (because X_s is $\mathcal{F}_s$-measurable, and if you take F to be an indicator function on an event in $\mathcal{F}_s$, that's precisely the definition of the conditional expectation).

So how do we show this? We simply take H to be an elementary process of the form

$$H = F\mathbf{1}_{(s,t]}.$$

(If you've reduced to showing this, you might guess that you should take this type of elementary process.) And now we just compute the inner product between K and H. By assumption, this is always 0 (because we assumed K is orthogonal to all elementary processes). But if you recall how we defined this inner product, that says

$$0 = \langle K, H\rangle_{L^2(M)} = \mathbb{E}\left[\int_0^t K_r H_r \, d\langle M, M\rangle_r\right].$$

Now we plug in our definition for H — H is just an indicator function times a random variable. So you can basically treat F as a constant in the integration variable r (F is just a random variable, it doesn't depend on r); this means you can pull it out of your integral, and you get

$$\mathbb{E}\left[F\int_s^t K_r\, d\langle M,M\rangle_r\right]$$

(now the fact that you're just integrating from s to t comes from the fact that you have $\mathbf{1}_{(s,t]}$). But now you see that this integral is precisely $X_t - X_s$. So we get that this is $\mathbb{E}[F(X_t - X_s)]$; this means we've shown

$$0 = \mathbb{E}[F(X_t - X_s)],$$

and therefore that X is actually a martingale.

So to wrap up, we now have a finite variation process which is also a continuous martingale; that means almost surely, $X_t = 0$ for all t. (In fact, we maybe don't even need 'almost surely.')

We now want to say why this implies $K = 0$ almost everywhere with respect to $d\langle M,M\rangle$ (you think of the QV as a measure on $[0,\infty)$). Why is this true? If you think of the measure μ defined as $d\mu = K\, d\langle M,M\rangle$ (so μ has density K against the QV), then the statement that $X_t = 0$ for all t implies that $\mu([0,t]) = 0$ for all t. (Here μ may be a signed measure, because K could be positive or negative.) So we have a signed measure with the property that it assigns every interval $[0,t]$ to be 0. This is going to imply (by e.g. Caratheodory extension) that μ is actually just 0 as a measure — it actually assigns every Borel set 0. (Probably you don't even need this extension theorem. You just want to say why a measure is uniquely determined by its values on these intervals $[0,t]$; and you can use the fact that they generate the Borel σ-algebra and maybe the π-λ theorem or something like that.) So you get that this measure is 0. And that implies the density K has to be 0 — for example, you can look at $\mu(\{r \mid K_r > 0\})$, and you're going to get that this is 0 (since this measure is going to be 0, you can apply it to an event of this form, and that's going to have measure 0). Similarly you also get $\mu(\{r \mid K_r < 0\}) = 0$; and that means $K = 0$ almost everywhere with respect to this measure.

And once you have this, that means $K = 0$ in $L^2(M)$ — because what it means to be 0 in a hilbert space is that its norm is 0. And we have

$$\langle K,K\rangle_{L^2(M)} = \mathbb{E}\left[\int_0^\infty K_r^2\, d\langle M,M\rangle_r\right].$$

But if you know K is 0 almost everywhere with respect to the measure you're integrating against, then even the integral (before taking an expectation) is 0. □

Student Question. *Why does the expectation being finite imply that X is FV?*

Answer. It means that almost surely $\int_0^t |K_r| \cdots < \infty$.

§13.3 Stochastic integrals with respect to L^2-bounded martingales

Now that we know $\mathcal{E}$ is dense, we can define the first version of stochastic integrals.

First, some notation:

Notation 13.5. We use X^T to denote the stopped process $X_t^T = X_{t\wedge T}$.

There's also this property (which we proved last week) that

$$\langle M^T, M^T\rangle_\infty = \langle M, M\rangle_T$$

(intuitively the QV measures how much your process is varying, so if your process becomes constant after time T, so should the QV). As a consequence of this, if $M \in \mathbb{H}^2$, then we also have $M^T \in \mathbb{H}^2$ — because an equivalent characterization of what it means to be in $\mathbb{H}^2$ is that the expectation of the QV at $t = \infty$ is finite. And if this is true for M, then it's also true for M^T, because

$$\langle M^T, M^T\rangle_\infty = \langle M, M\rangle_T \le \langle M, M\rangle_\infty.$$

Also, if you have a progressive process $H \in L^2(M)$ and T is a stopping time, then we also have

$$\mathbf{1}_{[0,T]}H \in L^2(M).$$

This is intuitive because the L^2 norm of this thing can be no bigger than the L^2 norm of H. And the fact that this is going to be a progressive process goes back to the measure theory stuff we spent the first weeks of class discussing.

Now we can finally define stochastic integrals, with respect to L^2-bounded martingales. This is the main step — to extend to continuous local martingales, you can just localize them to get L^2-bounded martingales, apply this theorem, and then stitch things together.

Theorem 13.6

Let $M \in \mathbb{H}^2$. Then for every elementary process $H \in \mathcal{E}$ of the form $H = \sum_{i=0}^p H_{(i)}\mathbf{1}_{(t_i,t_{i+1}]}$, we can define

$$(H \cdot M)_t = \sum_{i=0}^{p-1} H_{(i)}(M_{t_{i+1}\wedge t} - M_{t_i\wedge t}).$$

Then $H \cdot M \in \mathbb{H}^2$ (i.e., it's a L^2-bounded martingale). Moreover, the mapping $H \mapsto H \cdot M$ extends to a linear isometry $L^2(M) \to \mathbb{H}^2$.

You should think of $H \cdot M$ as the stochastic integral of H with respect to your martingale M.

Where does this formula come from? You're trying to define something like $\int_0^t H\,dM_s$ (intuitively). The whole problem is that M is a martingale, so it's not FV — it has finite quadratic variation, but that means the derivative of M in a classical sense doesn't exist. (For example, BM is 1/2-Hölder, so taking a derivative of it isn't well-defined — if it were, then you could define this integral using classical calculus, but it's not.) The point is if you have an indicator function, it's very easy to define this integral — you just say formally that because H is constant in time I can pull it out, and then I've reduced to integrating indicator functions on intervals — I want something like $\sum_{i=0}^{p-1} H_{(i)} \int_{t_i\wedge t}^{t_{i+1}\wedge t} dM_s$. And in any reasonable definition of an integral, when you integrate against the indicator on an interval, you should just get the difference of M at your two endpoints. So that's where this definition comes from.

Definition 13.7. This process $H \cdot M$ is often denoted as $(H \cdot M)_t = \int_0^t H_s\,dM_s$, and it's called the *stochastic integral* of H with respect to M.

Remark 13.8. Le Gall calls it the *stochastic integral*, but often it's called the Itô integral.

Now let's prove this. One part of the proof we'll defer until later is that this thing $H \cdot M$ is actually well-defined. The same elementary process could have multiple ways of writing it as a sum of indicators. So one thing you have to check is that when you have multiple ways of writing H, the formulas you get are consistent. But we'll do that later.

Proof. The first thing we'll observe is that $H \cdot M$ is manifestly a martingale — for each i, if we define

$$M_t^{(i)} = H_{(i)}(M_{t_{i+1}\wedge t_i} - M_{t_i \wedge t})$$

(the ith thing in the sum), then it's in $\mathbb{H}^2$. Why? First, it's L^2-bounded — H is a bounded random variable, and the second term is L^2-bounded. And it's a martingale by the same computation from when we constructed the QV — any time you have something of this form (where $H_{(i)}$ is $\mathcal{F}_{t_i}$-measurable — this is the important thing, and why we defined elementary processes the way you did), you get a martingale. And if you sum a finite number of elements in $\mathbb{H}^2$, you still get an element of $\mathbb{H}^2$; so $H \cdot M = \sum_{i=0}^{p-1} M^{(i)} \in \mathbb{H}^2$. So that's the first thing.

Next, let's say why this is going to be an isometry. This was maybe Itô's key observation — the fact that if you compute the norm of $H \cdot M$ as an element in $\mathbb{H}^2$, it's actually going to be exactly the same as the norm of H as an element of $L^2(M)$.

Claim 13.9 — We have $\langle H \cdot M, H \cdot M\rangle_{\mathbb{H}^2} = \langle H, H\rangle_{L^2(M)}$.

This is what it means to be an isometry — we're mapping $H \in L^2(M)$, and we're defining this martingale $H \cdot M$ (so we're mapping from a subspace of $L^2(M)$ into $\mathbb{H}$), and we're saying the norms are preserved.

The fact you'll get this again comes from the fact that you have this martingale structure where $H_{(i)}$ is $\mathcal{F}_{t_i}$-measurable.

Proof. Recalling the definitions of the inner product on $\mathbb{H}^2$, we have

$$\langle H \cdot M, H \cdot M\rangle = \mathbb{E}[\langle H \cdot M, H \cdot M\rangle_\infty].$$

So to calculate this, we need to calculate the QV of the process we just defined; so let's do that. The QV of $H \cdot M$, at least at a fixed time t — well, we wrote $H \cdot M$ as a linear combination of martingales, so by the bilinearity of the QV you can expand out and get

$$\langle H \cdot M, H \cdot M\rangle = \sum_{i,j} \langle M^{(i)}, M^{(j)}\rangle_t.$$

And then the thing you realize (very reminiscent, and maybe a generalization, of the fact that the increments of a L^2 martingale are uncorrelated):

Claim 13.10 — If $i \neq j$, then we have

$$\langle M^{(i)}, M^{(j)}\rangle = 0$$

(for all t).

Proof. Why is this true? If we visualize the points t_i, t_{i+1}, t_j, and t_{j+1}, then the assumption that $i \neq j$ means the interiors of these intervals are disjoint. And $M^{(i)}$ only varies on the first, while $M^{(j)}$ is constant on this interval and only varies on the second. And this is just the bracket of two martingales; if you have a bracket where it's always the case that one of the two things in the bracket is constant, then the bracket has to be 0. (For instance, you can use the discrete approximation of $\langle \bullet, \bullet\rangle$ that we showed last week.) □

So when you expand out the QV, you only get the diagonal terms; this is very similar to the fact that the increments of a L^2 martingale are uncorrelated.

Now let's see what the QV is when $i = j$ — we can compute

$$\langle M^{(i)}, M^{(i)}\rangle_t = H_{(i)}^2(\langle M, M\rangle_{t_{i+1}\wedge t} - \langle M, M\rangle_{t_i \wedge t})$$

(We had $M_t^{(i)} = H_{(i)}(M_{t_{i+1}\wedge t} - M_{t_i \wedge t})$, and you think of $H_{(i)}$ as constant in time; this means you can pull out two powers of it.)

So thus, you get that

$$\langle H \cdot M, H \cdot M \rangle_t = \sum_{i=0}^{p-1} \langle M^{(i)}, M^{(i)} \rangle_t = \sum_{i=0}^{p-1} H_{(i)}^2 (\langle M, M \rangle_{t_{i+1}\wedge t} - \langle M, M \rangle_{t_i \wedge t}).$$

But this looks almost like an integral of indicator functions — it's precisely going to be

$$\int_0^t H^2 \, d\langle M, M \rangle_s.$$

So we get this identity, and that precisely shows the isometry property we wanted — if we now apply this at time ∞, we get

$$\mathbb{E}[\langle H \cdot M, H \cdot M \rangle_\infty] = \mathbb{E}\left[\int_0^\infty H_r^2 \, d\langle M, M \rangle_r \right] = \langle H, H \rangle_{L^2(M)}$$

(by definition). So we've succeeded in showing this claim (that the map is an isometry). □

To emphasize, we've shown this is an isometry, but we haven't shown it's linear yet. (Usually when you say 'isometry,' you mean a linear isometry. We've shown that the norms are preserved, but we'll see why it's linear soon.)

But basically we're almost done proving this; we just have to say why this thing is well-defined and why the thing is linear. The proof of those two things turn out to be quite similar.

First let's say why $H \cdot M$ is well-defined.What do we need to do to check that this definition is well-defined? Well, we suppose that H can be represented in two different ways, as

$$H = \sum_{i=0}^{p-1} H_{(i)} \mathbf{1}_{(t_i, t_{i+1}]} = \sum_{j=0}^{q-1} K_{(j)} \mathbf{1}_{(s_j, s_{j+1}]}$$

(Le Gall skips over this in the book, but you have to think a bit to fill in the details, which is why we're doing it) — where this equality is in $L^2(M)$. And then we want to say the corresponding two processes are the same — we want to show that

$$\sum_{i=0}^{p-1} H_{(i)}(M_{t_{i+1}\wedge t} - M_{t_i \wedge t}) = \sum_{j=0}^{q-1} K_{(j)}(M_{s_{j+1}\wedge t} - M_{s_j \wedge t}).$$

(This is what we mean when we say the definition is well-defined.)

You have this same problem with defining the Lebesgue integral — there you first define it for simple functions (sums of indicators), and you have to think there about why it's well-defined. And you use the same trick there as here. What you do is you use this refinement trick. First, forget about the second representation, and let's focus on H being equal to the first sum of indicators. Then for any refinement of our subdivision $\{u_\ell\} \supseteq \{t_i\}$ — we had a subdivision $\{t_i\}$ which we used to represent H; we can always take a refinement, which is a subdivision containing this one (you visualize this by taking your first subdivision, and then adding some more points) — for any such refinement $\{u_\ell\}$, we can write

$$\sum_{i=0}^{p-1} H_{(i)} \mathbf{1}_{(t_i, t_{i+1}]} = \sum_{t=0}^{r-1} J_{(t)} \mathbf{1}_{(u_\ell, u_{\ell+1}]}$$

(the point is that you can basically take $\mathbf{1}_{(t_i, t_{i+1}]}$ and split it into a finer sum of indicators). So now we have a sum over these finer indicators, where the J's are properly chosen H's — more precisely,

$$J_{(\ell)} = H_{(i_\ell)} \quad \text{where } i_\ell \text{ is such that } (u_\ell, u_{\ell+1}] \subseteq (t_{i_\ell}, t_{i_\ell+1}].$$

Basically you're just inserting a telescoping sum for each $(t_i, t_{i+1}]$.

So we can always switch to a finer sum. And by explicit computation, you can check that $H \cdot M$ using this subdivision is going to be the exact same as the original process you defined, i.e.,

$$\sum_{i=0}^{p-1} H_{(i)}(M_{t_{i+1}\wedge t} - M_{t_i \wedge t}) = \sum_{\ell=0}^{r-1} J_{(\ell)}(M_{u_{\ell+1}\wedge t} - M_{u_\ell \wedge t}).$$

Why is this going to be true? If you for example consider the interval $(t_0, t_1]$, you're decomposing it into a disjoint union of intervals $(u_0, u_1] \cup \cdots \cup (u_{r-1}, u_r]$. And you can check that

$$M_{t_i \wedge t} - M_{t_0 \wedge t} = (M_{u_1 \wedge t} - M_{u_0 \wedge t}) + \cdots + (M_{u_r \wedge t} - M_{u_{r-1} \wedge t}).$$

The point is that this is just a telescoping sum.

In summary, what we've said is that at least if we take a refinement of H, *that's* well-defined — your stochastic integral upon taking refinements of the partitions stays the same.

But now given two representations of H, we can take a common refinement — given the two subdivisions $\{t_i\}$ and $\{s_j\}$, you can find a subdivision containing both of these, which will be a common refinement.

So we had this slight detour using refinements, but now let's come back to the situation where we had two representations of the same elementary process (with $H_{(i)}$ and $K_{(j)}$). By taking a common refinement $\{u_\ell\}$ of $\{t_i\}$ and $\{s_j\}$, we can write both sides by the same mesh $\{u_\ell\}$ — we basically get

$$\sum_{\ell=0}^{r-1} \widetilde{H}_{(\ell)} \mathbf{1}_{(u_\ell, u_{\ell+1}]} = \sum_{\ell=0}^{r-1} \widetilde{K}_{(\ell)} \mathbf{1}_{(u_\ell, u_{\ell+1}]}$$

(where $\widetilde{H}$ and $\widetilde{K}$ correspond to J from before), at least in $L^2(M)$. And now that you have both things on a common mesh (or a common subdivision), it suffices to show this identity with tildes, and with the sum running from 0 to $r-1$ — so it suffices to show that

$$\sum_{\ell=0}^{r-1} \widetilde{H}_{\ell}(M_{u_{\ell+1}\wedge t} - M_{u_\ell \wedge t}) = \sum_{\ell=0}^{r-1} \widetilde{K}_{\ell}(M_{u_{\ell+1}\wedge t} - M_{u_\ell \wedge t}).$$

Why does this suffice? We're saying that the left-hand side, by our refinement trick, is going to be equal to $\sum_{i=0}^{p-1} H_{(i)}(M_{t_{i+1}\wedge t_i} - M_{t_i \wedge t})$; and similarly the right-hand side will be at most the right-hand side of what we wanted.

So it remains to show this identity. And for that, we'll have to use our assumption that these are equal in $L^2(M)$; bringing the right-hand side over, we get that the elementary process

$$\sum_{\ell=0}^{r-1} (\widetilde{H}_{\ell} - \widetilde{K}_{(\ell)}) \mathbf{1}_{(u_\ell, u_{\ell+1}]} = 0$$

in $L^2(M)$. And by the isometry property we proved earlier, the process we get by integrating this with respect to M is just going to be 0 — i.e., the process

$$t \mapsto \sum_{\ell=0}^{t-1} (\widetilde{H}_{(\ell)} - \widetilde{K}_{(\ell)})(M_{u_{\ell+1}\wedge t} - M_{u_\ell \wedge t}) = 0$$

in $\mathbb{H}^2$ (because of the isometry property we proved earlier, applied to the particular elementary process $\widetilde{H}_{(\ell)} - \widetilde{K}_{(\ell)}$, which is 0 in $L^2(M)$). And if this is 0, then we get the desired identity.

Remark 13.11. This is something Le Gall kind of handwaves, but if you want to know the details, you kind of have to go through it — it's not that complicated, but you do have to be a bit careful and think through the details.

So now we've shown this thing is well-defined; the last thing we need to show is why it's a linear map. What do we mean by linearity? We want to show that for $H_1, H_2 \in \mathcal{E}$, we have

$$(H_1 + H_2) \cdot M = (H_1 \cdot M) + (H_2 \cdot M).$$

Again, this is basically the same refinement trick. *A priori* the problem you might have is that H_1 and H_2 might be defined using different subdivisions. But you can use the refinement trick to write them using a common subdivision, at which point this identity is elementary (it's basically the same argument as how we proved the thing is well-defined). □

So in summary, we get this linear isometry which is densely defined. And by functional analysis, that allows you to extend the isometry to all of $L^2(M)$; that's how we get an isometry on this entire space. So that constructs the stochastic integral with respect to L^2-bounded martingales.

Remark 13.12. In the last chapter we discussed semimartingales; so far we've just focused on martingales. But we basically defined integrals with respect to FV processes. A CLMG is just a MG plus a FV process, and we know how to integrate against FV processes; so the point is that as long as we know how to define the integral against martingales, you can also define integrals with respect to semimartingales. But the main point is why you can integrate against martingales.

§13.4 Properties of the stochastic integral

Now we'll see some properties. The first is a characterization.

Proposition 13.13

The martingale $H \cdot M$ is the unique martingale in $\mathbb{H}^2$ such that $\langle H \cdot M, N \rangle = H \langle M, N \rangle$ for all $N \in \mathbb{H}^2$.

So that's a characterization claim. The next, which is natural if you think of this as an actual integral:

Proposition 13.14

If T is a stopping time, then

$$(\mathbf{1}_{[0,T]}]H \cdot M = (H \cdot M)^T = H \cdot M^T.$$

We'll see the second claim follows from this characterization, so the main point is proving the characterization. If we think about why this is true, the intuition is we can kind of write the left-hand side as

$$\int_0^t \mathbf{1}_{[0,T]} \, H_s \, dM_s.$$

This means as a process in T, the integrand should become 0 after time T; so that's why you'd expect the first identity to be true. And why should we expect the second to be true? Now you've got $\int_0^t H_s \, dM_s^T$, so you're integrating against something that becomes constant after time T; so you'd intuitively expect the integral to eventually become constant after T.

If you try to prove this directly, even though it's intuitive, it actually gets messy. So it's kind of nice that it actually follows from the characterization.

One some level, if you think about Hilbert spaces, any vector on a Hilbert space is characterized by the linear functional it induces ($v \mapsto \langle u, v\rangle$); if you know your linear functional, then you know the vector. This is saying something simple — $H \cdot M$ is an element of your Hilbert space, and $\langle H \cdot M, N\rangle = H\langle M, N\rangle$ kind of describes the linear functional it induces (though this characterization is slightly more general).

Proof of Proposition 13.13. First, uniqueness is pretty immediate — if we had $X, Y \in \mathbb{H}^2$ both satisfying this property, then you could take $N = X - Y$. And then if you unwrap what the property is saying, it's saying that

$$\langle X, N\rangle = H\langle M, N\rangle = \langle Y, N\rangle,$$

which implies that

$$\langle X - Y, X - Y\rangle = 0$$

(using bilinearity to move the other bracket to the left-hand side, and using the fact that N itself is $X - Y$). This implies $X = Y$.

Now we actually have to show that this property is satisfied. First, why is $H\langle M, N\rangle$ well-defined? I'm integrating some element of $L^2(M)$ against a bracket, which is a signed measure.

This is where the Cauchy–Schwarz inequality from before (the Kunita–Watanabe theorem) comes into play. For this to be well-defined, the thing has to be integrable. So let's try to bound

$$\mathbb{E}\left[\int_0^\infty |H_s|\ |d\langle M, N\rangle_s|\right].$$

So we've fixed some $N \in \mathbb{H}^2$ and we want to estimate this thing. We need this to be finite, and it's a direct application of Kunita–Watanabe — you think of H_s as $H \cdot 1$, and you get that we can bound this by

$$\|H\|_{L^2(M)} \cdot \|N\|_{\mathbb{H}^2}$$

(since $\|N\|_{\mathbb{H}^2} = \|1\|_{L^2(M)}$ — the latter is just going to be $\mathbb{E}[\int_0^\infty 1^2\, d\langle N, N\rangle_s]$). And by assumption, both are finite — we have $H \in L^2(M)$ and $N \in \mathbb{H}^2$. (We didn't write it down in the proposition, but it's supposed to be part of the statement.)

So this random variable $\int_0^\infty H_s\, d\langle M, N\rangle_s$ — this notation means $(H \cdot \langle M, N\rangle)_\infty$. This is well-defined and integrable — i.e., it's in $L^1(\Omega, \mathcal{F}, \mathbb{P})$.

Now we need to actually check the property. The way you usually check these things (as in all areas of probability and analysis) is you check it in the simplest possible case, and extend by density. The simplest possible case is when we have an elementary function, because that's the only case where we have an explicit formula for what $H \cdot M$ is (otherwise it's just defined as the limit of something coming from extending your isometry).

So let's take $H \in \mathcal{E}$, and say $H = \sum_{i=0}^{p-1} H_{(i)} \mathbf{1}_{(t_i, t_{i+1}]}$. Recall the notation $M_t^{(i)} = H_{(i)}(M_{t_{i+1}\wedge t} - M_{t_i \wedge t})$ (this is the ith martingale you get when you intreate agains tM). Then we have

$$H \cdot M = \sum_{i=0}^{p-1} M^{(i)}.$$

Now let's evaluate the bracket of this integral with respect to another martingale N — by linearity of the bracket, we cna write

$$\langle H \cdot M, N\rangle = \sum_{i=0}^{p-1} \langle M^{(i)}, N\rangle.$$

And then we claim that this is going to be equal to

$$\sum_{i=0}^{p-1} H_{(i)}(\langle M,N\rangle_{t_{i+1}\wedge t} - \langle M,N\rangle_{t_i\wedge t}).$$

Intuitively,t he reason is your martingale only varies on $[t_i, t_{i+1})$. The way you'd fully verify this is ofr instance using the fact that

$$\langle M^{t_i}, N\rangle = \langle M,N\rangle^{t_i}$$

(we showed this last class), and you also do this for t_{i+1}. But now if you evaluate the right-hand sides at times t and take their difference, that's precisely what this thing in the parentheses is. And if you take the difference on the left, it's precisely this increment $M_{t_{i+1}\wedge t} - M_{t_i\wedge t}$.

So now we've done that, and we get that

$$\langle H\cdot M, N\rangle = \int_0^t H\, d\langle M,N\rangle$$

(H is an indicator function, and any time I have an indicator function, I'm supposed to just get a sum like this.) So the property is true for $H\in\mathcal{E}$.

Finally, you just argue by density. For a general progressive process $H\in L^2(M)$, by the density of $\mathcal{E}$ we can take a sequence $\{H^n\}\supseteq L^2(M)$ with $H^n\to H$. Then by the isometry we proved earlier (the fact that our map $L^2(M)\to\mathbb{H}^2$ is an isometry), that means

$$H^n\cdot tM \to H\cdot M$$

in $\mathbb{H}^2$ (by the isometry of the stochastic integral).

So to wrap up, if we now have a general L^2 bounded martingale $N\in\mathbb{H}^2$, we can say

$$\langle H\cdot M, N\rangle_\infty = \lim_{n\to\infty}\langle H^n\cdot M, N\rangle_\infty$$

(all the convergence is in L^1 — one can verify using Kunita–Watanabe). And now we can use the fact that the processes is true for elementary processes to get that this is equal to

$$\lim_{n\to\infty}(H^n\cdot\langle M,N\rangle)_\infty.$$

To finish, we just need to say that this is $(H\cdot\langle M,N\rangle_\infty)$. We're out of time, so we'll finish it next time. □

§14 March 19, 2025

Some fundamental properties of the Itô integral (which we defined for L^2 bounded martingales):

Proposition 14.1
$\int_0^t H_s\, dM_s$ is a martingale.

This is nice because martingales have mean 0 — often in probability you want to compute the expectations of things, and martingales make this easy (their expectation is 0).

The second thing is the isometry property we used to define the integral.

Proposition 14.2 (Ito isometry)
We have $\mathbb{E}[(\int_0^t H_s\,dM_s)^2] = \mathbb{E}[(\int_0^t H_s^2\,d\langle M,M\rangle_s)^2]$. More generally,

$$\mathbb{E}[\int_0^t H_s\,dM_s \cdot \int_0^t K_s\,dM_s] = \mathbb{E}[\int_0^t H_s K_s\,d\langle M,M\rangle_s].$$

This isometry property is precisely the reason we're able to define the integration — we first define it for simple processes and observe that for those you have this property, and then we extend by density.

At the discrete level, it's good to keep in mind that the isometry of the discrete level says that if you have a sum of increments $\sum_i H_i(M_{t_{i+1}} - M_{t_i})$ (where H_i is measurable with respect to the left endpoint of the interval), then

$$\mathbb{E}\left[\left(\sum_i H_i(M_{t_{i+1}} - M_{t_i})^2\right)\right] = \mathbb{E}\left[\sum_i H_i^2(M_{t_{i+1}} - M_{t_i})^2\right].$$

The whole point is this sort of encodes the cancellations — in principle you have a square of a sum, so you could have n^2 terms, but actually the off-diagonal terms all cancel out and only the diagonal terms matter (it's a fundamental fact about L^2 martingales that your increments are uncorrelated).

§14.1 A characterization property

Now let's continue proving the proposition from last time (the characterization property).

Proposition 14.3
For a L^2 bounded integral M, we have that $H \cdot M$ is the unique element of $\mathbb{H}^2$ such that for all N, we have $\langle H \cdot M, N\rangle = H\langle M, N\rangle$.

Moreover, if T is a stopping time, then

$$(\mathbf{1}_{[0,T]}H) \cdot M = (H \cdot M)^T = H \cdot M^T.$$

Recall that the right-hand side means you integrate your integrand H against the measure given by $\langle M, N\rangle$.

The second property will be needed when we extend Ito integrals to local martingales (doing a localization with stopping times). It says that you have the processes you'd expect of a stopping time — truncating your integrand at time T is the same as stopping the integral process at time T (making it constant), which is the same as stopping the integrator martingale.

Last time, we got part of the way through. Uniqueness was relatively quick. For existence, the strategy was to first verify the identity for elementary processes H; we did that completely last time. For that, any time you have an elementary process, you can just do the calculation, so it's possible to explicitly verify this. And then you take limits; that's what we were in the middle of last time. So let's complete the proof.

Given a general integrand $H \in L^2(M)$, take a sequence of elementary processes $\{H^n\} \subseteq \mathcal{E}$ with $H^n \to H$ (in your Hilbert space $L^2(M)$). Now we want to justify why you can kind of interchange the limits.

Let's first try to verify this identity at time ∞. (Both sides are processes, and their value actually exists at time ∞ — that's a consequence of the assumption that we're working with L^2 bounded martingales.) You can show that

$$\langle H \cdot M, N\rangle_\infty = \lim_{n\to\infty} \langle H^n \cdot M, N\rangle_\infty$$

(where the limit is in L^1, or even in probability). (This is because $H^n \to H$, so $H^n \cdot M \to H \cdot M$ in L^2, and this lets you verify the limit.)

Then we can use the fact that we verified the identity for elementary processes, so for every n we have $\langle H^n \cdot M, N\rangle = (H^n \cdot \langle M, N\rangle)_\infty$. So to finish, we just want to say why

$$\lim_{n\to\infty} (H^n \cdot \langle M, N\rangle)_\infty = H \cdot \langle M, N\rangle_\infty$$

(we'll show this limit holds in L^1, so it also holds in probability). To justify this limit, we can look at the L^1 norm of the difference and use the Kunita–Watanabe theorem to bound this by

$$\mathbb{E}[|((H - H^n) \cdot \langle M, N\rangle)_\infty|] \le \|H^n - H\|_{L^2(M)} \|N\|_{\mathbb{H}^2} \to 0.$$

So this shows why the property is true for a general integrator H.

So we've verified the identity at time ∞. But then to verify the identity at any *finite* time, you can replace N with N^t (the stopped version of N at your finite time t). If you apply this identity, you then get

$$\langle H \cdot M, N^t\rangle_\infty = \langle H \cdot M, N\rangle_\infty^t = \langle H \cdot M, N\rangle_t$$

(using properties of how the bracket behaves when you stop one of the processes inside it). And by the identity, the thing on the left is equal to

$$(H \cdot \langle M, N^t\rangle)_\infty = (H \cdot \langle M, N\rangle)_\infty^t = (H \cdot \langle M, N\rangle)_t$$

(maybe we didn't need the intermediate step). The point is that $\langle M, N^t\rangle_\infty$ becomes a measure which is 0 after time t, so if you want to integrate a general function against that measure, you just need to look at the integral up to time t.

So that verifies that as processes, the two things are equal. Now we can get to the second part of the proposition. Again, this is basically just playing around with how the bracket behaves when you consider stopping a process. Let's look at

$$\langle (H \cdot M)^T, N\rangle_t.$$

Because we're looking at a stopped process, we can move the stopping time into our evaluation time as

$$\langle H \cdot M, N\rangle_{T\wedge t}.$$

But then we can use the identity we just proved to say this is

$$(H \cdot \langle M, N\rangle)_{T\wedge t}.$$

And this we could have written as

$$(H\mathbf{1}_{[0,T]} \cdot \langle M, N\rangle)_t$$

(because this is a statement about integrating a function which becomes 0 after your time T).

By uniqueness in the first part, this is going to imply that

$$(H \cdot M)^T = (H\mathbf{1}_{[0,T]} \cdot M).$$

This is one of the identities in the chain (the first one). And the second one is similar — for the second one, we look at

$$\langle H \cdot M^T, N\rangle_t.$$

This is an integral of H againsta stopped martingale. So now we first use the identity in the property to write this as

$$(H \cdot \langle M^T, N\rangle)_t.$$

And now we can move T into the evaluation time to get that this is

$$(H \cdot \langle M, N\rangle)_{T\wedge t}.$$

And now this is the same as one of the steps in our previous chain, so this implies $H \cdot M^T = (H \cdot M)^T$ (again by uniqueness), which completes the second identity.

(This part is just algebraic manipulations with the time, but you can kind of guess it by thinking about what happens if something becomes constant past your stopping time T.)

As one quick application: suppose we look at the quadratic variation of the integral $H \cdot M$. Applying the property twice, we get

$$\langle H \cdot M, H \cdot M \rangle = H \cdot (\langle M, H \cdot M \rangle) = H \cdot (H \cdot \langle M, M \rangle) = H^2 \langle M, M \rangle$$

(the last step is the associativity of usual integrals). In some sense, this has to be true — it's a more general version of the Ito isometry, where $\mathbb{E}[(H \cdot M)_t^2] = \mathbb{E}[\langle H \cdot M, H \cdot M \rangle_t]$, and by this identity this is supposed to be $\mathbb{E}[\int_0^t H_s^2 \, d\langle M, M \rangle_s]$. So this is consistent with the Ito isometry. In fact, it's more general, because it's not just about expectations but about the actual paths.

Another way you can kind of remember this, or more generally (if we take different integrators and integrands, the same thing applies):

Fact 14.4 — We have $\langle H \cdot M, K \cdot N \rangle_t = (HK \langle M, N \rangle)_t$.

§14.2 SDE notation

Usually, people write these stochastic integrals using the following notation (rather than $H \cdot M$). If you say

$$dX_t = H_t \, dM_t,$$

this is kind of a SDE where you're specifying the time-evoluation of X. We define this to mean that

$$X_t = X_0 + \int_0^t H_s \, dM_s \quad \text{for all } t.$$

So you're just interpreting a SDE in integral form. (*A priori* $dX_t = H_t \, dM_t$ doesn't have classical meaning — if you think of M as a Brownian motion, its derivative is of negative regularity. So you'd have to talk about ODEs where your thing isn't even differentiable. But whenever we write a SDE like this, we just mean that this integral equation is true.)

Intuitively, what does it mean for you to satisfy this SDE? If you recall how we defined the Ito integral, that means

$$X_{t_{i+1}} - X_{t_i} \approx H_{t_i}(M_{t_i+1} - M_{t_i})$$

(at least, on a discrete level). Here it's very important, as usual, that H_{t_i} is measurable with respect to the *left* endpoint.

Now when we look at this bracket, we're basically saying that if $dX_t = H_t \, dM_t$ and $dY_t = K_t \, dM_t$, then

$$d\langle X, Y \rangle_t = H_t K_t \, d\langle M, N \rangle_t.$$

This isn't really even a stochastic differential equation — it's an actual differential equation (because you're saying the time-derivative of some increasing function is some continuous function). If you think on the discrete level, this is perfectly natural — because on a discrete level, you have

$$d\langle X, Y \rangle_t \approx (X_{t_{i+1}} - X_{t_i})(Y_{t_{i+1}} - Y_{t_i}).$$

And if we insert each of the above approximations, you get

$$H_{t_i} K_{t_i} (M_{t_{i+1}} - M_{t_i})^2,$$

which is basically the right-hand side $H_t K_t \, d\langle M, M \rangle_t$.

So these are all heuristic remarks about how you think about these stochastic differentials.

§14.3 Associativity of stochastic integrals

We want to get to defining the integral in full generality, but first we have one more preliminary proposition.

Proposition 14.5

Let $H \in L^2(M)$, and let K be a progressive process. Then $KH \in L^2(M)$ if and only if $K \in L^2(H \cdot M)$. Furthermore, if these properties hold, then

$$K \cdot (H \cdot M) = (KH) \cdot M.$$

Here $H \cdot M$ is a L^2-bounded martingale, so you can consider its space of progressive processes $L^2(H \cdot M)$. When we look at $K \cdot (H \cdot M)$, this is an integral with respect to some L^2-bounded martingale.

This is a natural property that you'd expect to be true, but you kind of have to prove it.

Remark 14.6. Often you'd write this in SDE notation; there you're just saying that if $dX_t = H_t\, dM_t$ and $dY_t = K_t\, dX_t$, then you should be able to 'substitute' dX_t in the first identity and conclude that $dY_t = K_t H_t\, dM_t$. So this is a very natural identity that you'd certainly expect to be true when written this way — you should just be able to substitute the evolution of X into here.

Proof. Let's first talk about the if and only if. We had this isometryu property from before, so we know

$$\langle H \cdot M, H \cdot M \rangle_t = \int_0^t H_s^2\, d\langle M, M \rangle_s.$$

Maybe it's better to write this in differential notation, where it says

$$d\langle H \cdot M, H \cdot M \rangle_t = H_t^2\, d\langle M, M \rangle_t.$$

What does it mean for K to be in L^2 of this space? This means we have to consider

$$\int_0^\infty K_t^2\, d\langle H \cdot M, H \cdot M \rangle_t$$

(the expectation of this thing would be finite if and only if $K \in L^2(H \cdot M)$). But now you can express this using the above formula — you're integrating K against this measure, but this measure has a density with respect to the other measure, so we can rewrite this as

$$\int_0^\infty K_t^2 H_t^2\, d\langle M, M \rangle_t.$$

This is even an identity of random variables; so if you take expectations, the first is finite if and only if the second is.

So that proves the 'if and only if'; now we need to prove the associativity property. This is where the characterization of the stochastic integral comes in handy. To prove the second part, let's consider the bracket

$$\langle K \cdot (H \cdot M), N \rangle$$

(where N is some other L^2 bounded martingale). We want to show that this is the same as what we'd get if instead of $K \cdot (H \cdot M)$, we input $(KH) \cdot M$. But this is just a sequence of algebraic steps. First we can move K outside to rewrite this as

$$K \cdot \langle H \cdot M, N \rangle.$$

Then we can move H outside and get

$$K \cdot (H \cdot \langle M, N \rangle).$$

And here you just use associativity of regular itnegrals — you think of this whole thing as the integral of some measurable function against some measure, and that measure has a density with respect to $\langle M, N \rangle$, so you can write this as $(KH) \cdot \langle M, N \rangle$. ANd then you can put KH into the bracket, so this becomes

$$\langle (KH) \cdot M, N \rangle,$$

which proves the characterization identity. □

§14.4 Integration against general CLMGs

Now we'll define the processes that we're going to integrate against CLMGs. The point is that maybe a process belonging to $L^2(M)$ is too restrictive; you might want to integrate more general processes. So we'll somewhat relax that restriction.

Definition 14.7. If M is a CLMG, write $L^2_{\text{loc}}(M)$ for the set of progressive processes H such that almost surely,

$$\int_0^t H^2 \, d\langle M, M \rangle_s < \infty \quad \text{for all } t.$$

('Loc' stands for 'local.') So we're not making any assumptions on the expectation — that could be infinite, just the random variable has to be finite.

Definition 14.8. We define $L^2(M)$ as the set of progressive processes H such that

$$\mathbb{E}[\int_0^\infty H_s^2 \, d\langle M, M \rangle_s] < \infty.$$

This condition is much stronger than the first one.

Remark 14.9. Again, you can view $L^2(M)$ as an actual L^2 space, where your measure is basically $d\mathbb{P}$ where $\mathbb{P}$ is the probability measure on your sample space, and then $d\langle M, M \rangle$.

Now let's construct Ito integrals with respect to CLMGs.

Theorem 14.10 (Itô integrals with respect to CLMGs)

Let M be a CLMG. Then for every $H \in L^2_{\text{loc}}(M)$, there exists a unique CLMG starting at 0, which we denote $H \cdot M$, such that for any CLMG N, we have

$$\langle H \cdot M, N \rangle = H \cdot \langle M, N \rangle.$$

('Unique' is always up to indistinguishability.)

This is kind of abstract, but the point is, how should you define a stochastic integral in this general setting? Well, we had thsi charcterization property before. And we're saying we'll define this thing to be whatever continuous local martingale satisfies this characterization property (in the general setting where you have local martingales); this characterization property characterizes a thing, so there's a unique process.

Theorem 14.11

If T is a stopping time, then

$$(\mathbf{1}_{[0,T]}H) \cdot M = (H \cdot M)^T = H \cdot M^T.$$

This is the same thing as before — that you can put the stopping time in various places.

Theorem 14.12

If $H \in L^2_{\text{loc}}(M)$ and K is a progressive process, then $HK \in L^2_{\text{loc}}(M)$ if and only if $K \in L^2_{\text{loc}}(H \cdot M)$, and in this case we have

$$K \cdot (H \cdot M) = (KH) \cdot M.$$

This is basically the same associativity property as before, but in this more general setting.

The final claim is that this definition actually extends the one we had for L^2-bounded martingales (i.e., it's the same as what we constructed before).

Theorem 14.13

If $M \in \mathbb{H}^2$ and $H \in L^2(M)$, then $H \cdot M$ is the same as before.

So that's the statement of the theorem. The main thing is that integrals against local martingales are still going to e local martingales (though not necessarily bounded).

Before we get into this proof, the statement is kind of nonconstructive — what is a formula for $H \cdot M$? The intuition is that you still think of it as given by a discrete approximation

$$H \cdot M \approx H_i(M_{t_{i+1}} - M_{t_i}).$$

The point (which we'll prove later) is that if you take a mesh, form this linear combination, and take mesh size to 0, then you converge to the stochastic integral. So this is how you think of the martingale.

Again, it's very important that H is measurable with repsectt to the left endpoint.

Proof. Without loss of generalith let's take $M_0 = 0$. (Otherwise you can write $M = M_0 + M'$ and set $H \cdot M = H \cdot M'$. The point is if you go back to the discrete approximation, you only care about the increments of M; you don't care about where it started. So we may as well assume M starts at 0.)

We can also assume that for all $\omega \in \Omega$, we have

$$\int_0^t H_s^2 \, d\langle M, M\rangle_s < \infty \quad \text{for all } t.$$

The assumption is that H is in L^2_{loc}, so this is true almost surely. But at the beginning of this chapter we said we assume the filtration is complete, so we can modify everything on the almost sure event where this holds, and we still have all the relevant properties (adaptiveness, progressiveness, and so on). So by doing this modification, we can just assume that this is true for *all* ω (not just almost all).

Then, as with most proofs involving local martingales, the idea is to define a sequence of localization times — stopping times T_n which reduce your local martingale.

So for $n \geq 1$, let's define

$$T_n = \inf\left\{t \geq 0 \mid \int_0^t (1 + H_s^2) d\langle M, M\rangle_s \geq n\right\}.$$

It's basically the integral of H_s^2 (the reason we put a $+1$ is just because we also want to say that $\langle M, M\rangle_s \leq n$ up to your stopping time). Then $T_n \uparrow \infty$ (this is just coming from the fact that this integral is going to be finite for all t).

And like we just said, because we have this extra 1, we have

$$\langle M^{T_n}, M^{T_n}\rangle_t = \langle M, M\rangle_{T_n \wedge t} \leq n$$

by definition. This actually means the stopped local martingale M^{T_n} is actually L^2 bounded (because an equivalent condition to be a L^2 bounded martingale is to say the expectation of your QV at time ∞ is finite; and here your QV is just bounded, which is even stronger). So $M^{T_n} \in \mathbb{H}^2$.

So here's for instance maybe where we assume M starts at 0 (if it doesn't, then if M_0 was not integrable M^{T_n} would not be a martingale — because this condition is only about the increments of M).

Another thing to note is — we now have a L^2 bounded martingale, and we know how to integrate against such things. But we need our integrand to be in L^2 of our L^2-bounded martingale. Why is that going to be true? We claim that $H \in L^2(M^{T_n})$. And that's going to be true because

$$\int_0^\infty H_s^2 \, d\langle M^{T_n}, M^{T_n}\rangle_s = \int_0^{T_n} H_s^2 \, d\langle M, M\rangle_s$$

(because after time T_n, M is just constant and this thing becomes 0). And by the definition of the stopping time, this is just finite — it's bounded by n. So this actually means $H \in L^2(M^{T_n})$.

Now that we've verified that, you want to define your integral — at least for times before T^n — as the integral of H against the stopped martingale (that's the ntural guess). One thing you have to be careful of is you want to be able to stitch together these processes. So we need to check that if I have a time before both T_n and T_m, then the corresponding integrals at that time are the same. For $m \geq n$, we have that

$$(H \cdot M^{T_m})^{T_n} = H \cdot (M^{T_m})^{T_n} = H \cdot M^{T_n}$$

(the first equality is because of how integrals behave with respect to stopping times, and the second because $T_m \geq T_n$). What this means is we can define

$$(H \cdot M)_t = \lim_{n \to \infty} (H \cdot M^{T_n})_t.$$

The point is that eventually this thing stabilizes — once n is large enough that $T_n \geq t$, this thing is actually just constant in n by the above identity. So the RHS stabilizes, which means the limit actually exists.

So now we have our candidate process. It'll definitely be adapted and have continuous sample paths because of the way we defined it. But the next thing is, why is this a local martingale? So we need to exhibit a sequence of stopping times that reduces this thing. And we claim the sequence (T_n) actually just reduce this process. Why is that? If we look at the stopped process $(H \cdot M)^{T_n}$, we want to say why this stopped process is going to be a UI martingale (that's what it means for a sequence of stopping times to reduce a local martingale). Why is this going to be true? Well, the reason is that if I believe that the second property $(H \cdot M)^T = H \cdot M^T$ is going to hold (we'll prove it later, or maybe we'll skip the proof because it's the same thing from before), then you should believe that this is going to precisely be $H \cdot (M^{T_n})$. And M^{T_n} is a L^2 martingale.

This identity is not just trivial — we chose to define $H \cdot M$ in this way, so we need to show that with this definition, the above identity $(H \cdot M)^{T_m} = H \cdot (M^{T_n})$ is true.

To show this, let's insert the definition of $H \cdot M$; then we get that

$$(H \cdot M)^{T_n} = \lim_{m \to \infty} (H \cdot M^{T_m})_t^{T_n}.$$

Now in the pre-limit, we can put the T_n onto this martingale — this is going to be exactly equal to

$$\lim_{m\to\infty}(H\cdot M^{T_n\wedge T_m})_t.$$

Why? The point is that now I'm back in the L^2 bounded case, so the identity where I put the T_n into the martingale is actually true (because we proved this in the L^2 bounded case). And now if I send $m\to\infty$, I just get $(H\cdot M^{T_n})_t$.

So we've shown this identity; and $H\cdot(M^{T_n})$ is a L^2-bounded martingale, so (T_n) does reduce our process. This means $H\cdot M$ really is a local martingale.

And then you can probably check that if M was a L^2 bounded martingale to start with and $H\in L^2(M)$, you kind of recover the original definition (that $H\cdot M$ is the same as before). (We won't prove this; if you want, you can check for yourself.)

The next step is we want to show the characterization identity. We've succeeded in constructing a CLMG which starts at 0, and now we want to show why it satisfies $\langle H\cdot M,N\rangle = H\cdot\langle M,N\rangle$.

Again, the idea is that we've verified it before taking the limit (because you're in the L^2 bounded case). So we just have to explain why you can commute the limit.

That requires slightly additional things. Let N be a CLMG. Now we also need to reduce N (before, the identity was true when both M and N were L^2 bounded martingales, so we want to get a L^2 bounded martingale out of N). So we let

$$T_n' = \inf\{t\geq 0 \mid |N_t|\geq n\},$$

so that (T_n') reduces N, and when you stop N at these stopping times, it becomes bounded (in particular, it's a L^2 bounded martingale).

Before we write that, we'll define a stopping time $S_n = T_n\wedge T_n'$. Recall that for instance, when we observed that the sum of CLMGs is a CLMG, the way you prove that is that if you have one sequence of stopping times reducing your first CLMG and another sequence reducing the other, then when you take their min that reduces both. So we'll take this min of the two stopping times. One consequence of this is that N^{S_n} is a L^2 bounded martingale.

So we should have that

$$\langle H\cdot M,N\rangle_t = \lim_{n\to\infty}\langle H\cdot M,N\rangle_t^{S_n}$$

(because $S_n\uparrow\infty$). Now I just want to put S_n on each of these terms to get L^2-bounded martingales — I want to say this limit is the same as

$$\lim_{n\to\infty}\langle (H\cdot M)^{S_n},N^{S_n}\rangle$$

(we used the fact that you can put the stopping time on either of the two terms, at least in the L^2 bounded case). (There might be some additional justification needed, but supposing this is true...) We should now further get that this is

$$\lim_{n\to\infty}\langle H\cdot(M^{S_n}),N^{S_n}\rangle$$

(where we do one more step and put the stopping time on M). Now the point is I have L^2 bounded martingales, and using the characterization property, I get that this is

$$\lim_{n\to\infty} H\cdot\langle M^{S_n},N^{S_n}\rangle_t.$$

Now this is

$$\lim_{n\to\infty} H\cdot\langle M,N\rangle_t^{S_n}$$

(this is a property we proved for the bracket of CLMGs — when you have a bracket, you can put your stopping time on either or both terms). Now we can say this is

$$\lim_{n\to\infty}(H\cdot\langle M,N\rangle)_t^{S_n} = (H\cdot\langle M,N\rangle)_t. \qquad \square$$

So we proved the characterization identity. We're not going to show uniqueness; it's basically the argument. And the fact about stopping times (with T) is also basically the same argument, manipulating these arguments and stopping times, so we'll skip it as well.

Remark 14.14. One remark: We continue to use SDE notation. So now more generally if M is a CLMG and $H \in L^2_{\text{loc}}(M)$, then $dX_t = H_t\,dM_t$ just means that $X_t = X_0 + \int_0^t H_s\,dM_s$. When we talk about Ito's formula, we'll probably write things in SDE notation. In practice that's usually how we write things, so it's good to get comfortable with it.

§14.5 Integrals with respect to continuous semimartingales

In summary, we defined integrals with respect to CLMGs. And when we talked about FV processes, we also defined integrals with respect to FV processes (this is just analysis, where you're integrating against a measure — that measure is a random thing now, but that's fine). Now we can combine these to define integrals against semimartingales. There's basically nothing to prove here; we just make the very natural definitions.

What are our integrands going to be? Well, for integrals with respect to local martingales, the natural integrands are these things in L^2_{loc}. And with FV processes, they're basically things that are integrable with respect to the corresponding measure. So you want your condition to be that both of those things are true.

One way to ensure both things are satisfied at the same time is:

Definition 14.15. We say a progressive process is *locally bounded* if almost surely, we have

$$\sup_{0\le s\le t} |H_s| < \infty \quad \text{for all } t \ge 0.$$

In other words, this says it's (almost surely) bounded on any finite interval.

Example 14.16

Any adapted process with continuous sample paths is locally bounded. (One of the general measurability results we showed earlier on is any adapted process with right-continuous sample paths is progressive. And because it's continuous, this locally bounded condition is definitely true.)

The point is that any such process is in L^2_{loc} of a local martingale, and also you can integrate any such process with respect to a FV process — if H is progressive and locally bounded, then for any FV process V, we have that almost surely,

$$\int_0^t |H_s|\,|dV_s| < \infty \quad \text{for all } t \ge 0,$$

because you can use the fact that H is bounded on all these intervals $[0,t]$ and pull the bounding constant out; and then you're just saying that $\int |dV_s|$ is bounded on every finite interval $[0,t]$, which is basically just the definition of a FV process. And similarly $H \in L^2_{\text{loc}}(M)$ for all CLMGs M. Why? Recall that being in L^2_{loc} means that $\int_0^t H_s^2 d\langle M,M\rangle_s < \infty$ for all t. But this is the same situation as before — $\langle M,M\rangle$ is a FV process and H is bounded.

So now we can integrate H against FV processes and CLMGs, which means we can integrate it against semimartingales.

Definition 14.17. Let X be a CSMG, and let its canonical decomposition be $X = M + V$. If H is a locally bounded progressive process, then we define

$$H \cdot X = H \cdot M + H \cdot V = .$$

Recall that a CSMG is one you can write as a sum of a local martingale M and FV process V; this is unique up to indistinguishability (because you can't be both a CLMG and FV process unless you're just 0).

More explicitly, this says that $(H \cdot X)_t = \int_0^t H_s \, dM_s + \int_0^t H_s \, dV_s$.

Definition 14.18. We similarly write $(H \cdot X)_t = \int_0^t H_s \, dX_s$.

To recap the stochastic integral, when we talked about FV processes, we mentioned that in $H \cdot V$ you're just using the fact that you can integrate against measures; this is just real analysis. The thing that's fundamentally new in stochastic calculus is defining $\int H_s \, dM_s$, and this uses hte isometry property and special properties of L^2 bounded martingales. Then you can put these together and define integrals with respect to CSMGs.

Remark 14.19. Again, in SDE notation, you could write your semimartingale as

$$dX_t = dM_t + dV_t.$$

But what does this mean? Usualy we take this to mean that $X_t = X_0 + \int_0^t dM_s + \int_0^t dV_s$ (at least, that's what we said for stochastic integrals). We have $\int_0^t dV_s = V_t$, kind of just by definition — that's what this means. So this part is consistent — if we want to say that this implies $X = M + V$, then we have the V part. But why is the first part equal to M_t? Certainly this is a very natural property expect — the fact that if you interpret this SDE notation in the integral form or as $X = M + V$, the two interpretations are consistent.

Is this obvious from the construction of the stochastic integral? We can probably use the characterization property. If we define H to be the process which is always 1 (certainly that's a progressive process in L^2_{loc}), the claim is that if I integrate 1 against a local martingale, I should just get the local martingale back — i.e., $H \cdot M = M$. Hopefully this shoudl be true (or else this would be a really bad notion of integral), but how do we prove it? We just need to check that

$$\langle H \cdot M, N \rangle = \langle M, N \rangle.$$

Why is that true? We can use the earlier identity to pull the H out, and write $\langle H \cdot M, N \rangle = H \cdot \langle M, N \rangle$. And now this is a usual integral in analysis — I'm just integrating 1 against a measure, so I get the measure back, and this is $\langle M, N \rangle$.

Remark 14.20. Another way to see this is that eventually we're going to prove that the discrete approximations $\sum_i H_i (M_{t_{i+1}} - M_{t_i})$ converge to the integral (we haven't done this yet). But if H is 1, then this telescopes and it's just M_t.

Remark 14.21. As is often the case when you define integrals so abstractly, when actually computing them you have to think. We were able to compute this one (where we integrate 1 against M). But another common integral to compute is $\int_0^t M_s\,dM_s$. This is always well-defined because if M is a local martingale, it has continuous sample paths, so it's a progressive locally bounded process. So we can write down this integral; we've certainly constructed it. But what is it — can you actually work with it?

The point is we should have

$$\int_0^t M_s\,dM_s = \frac{1}{2}(M_t^2 - \langle M, M\rangle_t).$$

We're going to see this from Itô's formula. We may have mentioned this in the QV part of the course. Recall that the QV is defined as the unique increasing process such that when you subtract it from M^2, you get a local martingale; and in the proof of constructing the QV, we were working with discrete approximations to precisely this thing $\int_0^t M_s\,dM_s$. We basically showed that as a sequence of local martingales, when you take mesh size to 0, we have $\sum_i M_{t_i}(M_{t_{i+1}} - M_{t_i})$ converging to a local martingale, which is this one.

(This discussion is circular because we need the QV to construct these stochastic integrals; but it's at least a consistency check.)

And we'll see more examples later.

§14.6 Properties of the integral with respect to CSMGs

Let's finish by discussing some properties of the integral with respect to semimartingales.

Proposition 14.22

The map $(H, X) \mapsto H \cdot X$ (where H is a locally bounded progressive process and X is a semimartingale) is bilinear.

This is kind of just by definition. It's maybe a good thing to check tha tthe stochastic integral is linear — that

$$(H_1 + H_2) \cdot M = H_1 \cdot M + H_2 \cdot M.$$

We used this when defining the stochastic integral for L^2-bounded things — we checked that on *elementary* processes it's linear, so it remains linear when you extend. And you can check that linearity still holds in this more general setting, again by some localization thing. (Or you can use the characterization; that's probably more direct.)

You also have the natural associativity property.

Proposition 14.23

If H and K are both progressive and locally bounded, then $H \cdot (K \cdot X) = (HK) \cdot X$.

(We use 'progressive and locally bounded' to make sure both sides are well-defined. If you want to be more general, you can say that if the left-hand side is well-defined then so is the right-hand side and vice versa, and they have to be equal.)

Again in SDE notation this is very natural — you can write $dY_t = H_t\,dX_t$ and $dZ_t = K_t\,dY_t$, and then this says $dZ_t = K_t H_t\,dX_t$.

The third property is that you have the same thing about putting stopping times in various places.

Proposition 14.24

For every stopping time T, you have

$$(H \cdot X)^T = (H\mathbf{1}_{[0,T]}) \cdot X = H \cdot (X^T).$$

Proposition 14.25

If X is a CLMG, or respectively a FV process, then the same is true for $H \cdot X$.

So if you integrate against a local martingale you get a local martingale; if you integrate against a FV process you get a FV process. That's another very nice thing about the stochastic integral, which we mentioned at the beginning — it respects martingale structure.

Proposition 14.26

If $H = \sum H_i \mathbf{1}_{(t_i, t_{i+1}]}$ where $H_i \in \mathcal{F}_{t_i}$, then $(H \cdot X)_t = \sum_{i=0}^{p-1} H_i(X_{t_{i+1}} - X_{t_i})$.

This is slightly more general than being an elementary process (elementary processes additionally require the H_i to be bounded), so there's something to check here.

So in summary, we've finished constructing stochastic integrals. In the remaining part of the course, we'll be exploring what you can do using stochastic integrals. Next week is Spring Break; when we return we'll talk about Itô's formula. Sadly Sky won't be here that day, but there will be a substitute.

§15 March 31, 2025

(Sky is not here, so Yuanzheng Wang is teaching.)

§15.1 Review

Let's first recall what we did before break.

Definition 15.1. A CSMG X_t is a sum $M_t + V_t$ where M_t is a CLMG and V_t is a FV process. We define $\langle X, X\rangle_t = \langle M, M\rangle_t$.

We have a discrete sum approximation of the QV — if we have a sequence of subdivisions with mesh size going to 0, then

$$\sum_{i=1}^{p_n} (X_{t_i} - X_{t_{i-1}})^2 \to \langle X, X\rangle_t$$

in probability. The week before break, we tried defining the stochastic integral with respect to a CSMG.

Definition 15.2. For a progressive and locally bounded process H, we define $\int_0^t H_s \, dX_s = \int_0^t H_s \, dM_s + \int_0^t H_s \, dV_s$.

(*Locally bounded* means that the local supremum of H is always fininte — i.e., $\sup_{0 \le s \le t} |H_s| < \infty$ for all t.)

If H_s is locally bounded, the second part is comparatively easier to define, and it turns out this is just a FV process. The week before break we tried defining $\int_0^t H_s \, dM_s$, which turns out to be a CLMG; thus this is a CSMG.

We took two steps to define this stochastic integral. The first step is we tried to define it for some 'nice' M and 'nice' H. And then we used a localization technique to extend it to general CLMGs and H.

In the first step, we considered the space $\mathbb{H}^2$ — the space of continuous martingales bounded in L^2. This space is a Hilbert space with inner product given by $\langle M, N\rangle_{\mathbb{H}^2} = \mathbb{E}[M_\infty N_\infty]$.

First, we considered any $M \in \mathbb{H}^2$. For such M, we considered the space $L^2(M)$, the space of all progressive H such that $\mathbb{E}[\int_0^\infty H_s^2\, d\langle M, M\rangle_s] < \infty$.

Then for such $M \in \mathbb{H}^2$ and such nice H, we defined $\int_0^t H_s\, dM_s$ by an Ito isometry mapping $H \mapsto \int_0^t H_s\, dM_s$. This is an L^2 isometry from $L^2(M)$ to $\mathbb{H}^2$. It turns out that in this case, the stochastic integral $\int_0^t H_s\, dM_s$ is a continuous martingale bounded in L^2, with QV given by $\int_0^t H_s^2 d\langle M, M\rangle_s$.

In the second step, we just used a localization technique. By introducing a bunch of stopping times, we were able to extend the definition of this integral $\int_0^t H_s\, dM_s$ to every CLMG M and every $H \in L^2_{\text{loc}}(M)$, where this is defined as the space of all progressive H such that for every $t > 0$, we have $\int_0^t H_s^2\, d\langle M, M\rangle_s < \infty$. We can see that if H is locally bounded, then this condition is always satisfied. So if we can define the integral for any such M and any such H, then we're able to define our desired stochstic integral.

So by combining these two steps we're able to define $\int H_s\, dM_s$, and therefore $\int H_s\, dX_s$.

This is what we've done before the spring break. Today our main goal is to introduce Ito's formula, which is kind of the most important topic in this stochastic calculus course.

§15.2 Some technical results

But before doing this, we need to introduce two technical results. The first is a stochastic version of the usual L^1 dominated convergence theorem. Here we're considering a general CSMG X, which has a canonical decomposition as $M + V$, and we fix t.

Proposition 15.3 (DCT for stochastic integrals)

Let $X = M + V$ be a CSMG, and let $t > 0$. Let $(H^n)_{n\geq 1}$ and H be locally bounded progressive processes, and let K be a nonnegative process. Assume that the following hold almost surely:

(i) $H_s^n \to H_s$ for every $s \in [0, t]$ (as $n \to \infty$).

(ii) We have $|H_s^n| \leq K_s$ for all n and all $s \in [0, t]$.

(iii) $\int_0^t K_s^2\, d\langle M, M\rangle_s < \infty$ and $\int_0^t K_s\, |dV_s| < \infty$.

Then $\int_0^t H_s^n\, dX_s \to \int_0^t H_s\, dX_s$ in probability.

We don't require K to be locally bounded; but since it's always nonnegative, the two integrals in (iii) are always defined.

To prove this, we want to use the definition of the stochastic integral. This is a sum of a CLMG part and a FV part. So to prove this convergence in probability, it suffices to prove convergence in probability for each part.

Proof. Let's first consider the FV part. By the usual L^1 dominated convergence theorem, we have that $\int H_s^n\, dV_s \to \int H_s\, dV_s$ almost surely. Here *a priori* we know the H^n, V, and H are always random; the way we're using DCT is that we *fix* an almost sure sample $\omega \in \Omega$ that satisfies (i)—(iii). Under this sample, H becomes a deterministic function $s \mapsto H_s^n(\omega)$, V becomes a deterministic process (which can be seen as a deterministic finite measure), and H also becomes a deterministic process; and we can use the usual L^1

dominated convergence theorem to say that

$$\int_0^t H_s^n(\omega)\, dV_s(\omega) \to \int_0^t H_s(\omega)\, dV_s(\omega).$$

(And this sample is almost sure, so we also get that this convergence is almost sure.)

Now it remains to show that the CLMG part also converges, i.e., that $\int_0^t H_s\, dM_s \to \int_0^t H_s\, dM_s s$ in probability. Recall that in the construction of this CLMG, we used two steps. In the first, we used an Ito isometry to define this integral for some nice M and H, and in the second step we used a localization technique. This is also the idea of hwo we are goign to prove the convergence in probability.

So we're going to first consider the case where M is a continuous martingale bounded in L^2, i.e., $M \in \mathbb{H}^2$. Assume that H^n, H, and K are all in $L^2(M)$. Then by Ito's isometry, we have

$$\mathbb{E}\left[\left(\int_0^t H_s^n\, dM_s - \int_0^t H_s\, dM_s\right)^2\right] = \mathbb{E}\left[\int_0^t (H_s^n - H_s)^2\, d\langle M, M\rangle_s\right].$$

The the right-hand side can again be interpreted as an L^1 integral, but here over two variables s (the time parameter and *omega* (the sample over probabilistic space). So here we have a double integral where we ifrst integrate against the QV, then take an expectaton nder the probability measure.

And under this L^1 integral, by (i) we know that $H_n - H \to 0$ almost surely for an alost sure pair of samples (s, ω). Then by (ii), we now that $(H_s^n - H_s)^2 \le (2Ks)$. And then because $K \in L^2$, we know that $H_s^n - H_s)^2$ is also integrable.

Then by the usual L^1 DCT, the RHS converges to 0.

This tells us that the left-hand side also converges to 0 in probablotu. SO our two stochastic integrals converge to each otheer in L^2, which implies convergence in probability.

Then to handle the general case, we will use thej localization technique. We'll define a sequence of stopping times as

$$T_m = \inf\{u \ge 0 \mid \int_0^u (1 + K_s^2)\, d\langle M, M\rangle_s \ge n\} \wedge t.$$

SO the first thing is exactly thee stopping time we encountered. And we fix t and we're only considering the stochastic integral up to time t.

Then under this stopping time, we know the stopped process M^{T_m} is in $\mathbb{H}^2$, and $H^n, H, K \in L^2(M^{T_m})$. Then for every m, we get that

$$\int_0^t H_s^n\, dM_s^{T_m} \to \int_0^t H_s\, dM_s^{T_m}$$

in probability.

Now because M is a CMLG (by assumption (3), which requires our last integral to be fininte)... we know that as $m \to \infty$, we should also have $\int_0^n (1+K_s^2)\, d\langle M, M\rangle_s \to \infty$, and therefore $T_m \to t$ (i.e., $\mathbb{P}[T_m = t] \to 1$).

Then the convergence in probability should follow from these two equations — namely, that

$$\mathbb{P}[]$$

(using $H \cdot M$ as an abbreviation of $\mathbb{P}[\left|(H^n \cdot M)_t - (H \cdot M)^t\right| \ge \varepsilon)$) — is at most

$$\mathbb{P}[T_m \ne t] + \mathbb{P}[\left|H^n \cdot M^{T_m}\right|_r - (H \cdot M^{T_m})_t].$$

And by what we talked aboiut before, for any fixed m the right-hand side converges to what we want. Then we can take $m \to \infty$; then the first part also goes to 0, and the convergence in probability should follow. □

There's one more technical result that we need to prove.

Proposition 15.4

Let X be a CSMG, and let H be an adapted process with continuous sample paths. Then for every $t > 0$ and every seequence $0 = t_0^n < \cdots < t_{p_n}^n = t$ of subdivisions of $[0, t]$ with mesh size tending to 0, we have

$$\sum_{i=0}^{p_n - 1} H_{t_i^n}(X_{t_{i+1}^n} - X_{t_i^n}) \to \int_0^t H_s \, dX_s$$

in probabiltiy as $n \to \infty$.

In short, thsi sauys that the stochastic integral with respect to a CSMG on the right can be approximated by some kind of a Riemann sum.

Remark 15.5. Mote that any continuous local martingale is locally bounded.

Note that when we take the Riemann sum, we're always taking the leftmost endpoint of each interval.

Proof. We're actually trying to use the stochastic version of the DCT here. So we'll define H^n as a sequence of step functions

$$H^n = H_0 \mathbf{1}_{\{0\}} + \sum_{i=0}^{p_n - 1} H_{t_i^n} \mathbf{1}_{(t_i^n, t_{i+1}^n)}.$$

So we can imagine graphing H and drawing t_1^n, t_2^n, and t_3^n; so we take the value H_0 from 0 to t_1^n, then value $H_{t_1^n}$ from T_1^n to T_2^n, and so on.

Now because H has continuous sample paths, we know that $H^n \to H$ pointwise by continuity. Also, because we're taking the *left* endpoints and H is adapted and continuous, we know every step function H^n is a progressive process.

Now to apply the DCT, we need to find a dominating function for all the step functions. In this case, we'll just take the dominating function

$$K_s = \sup_{0 \le r \le s} \sup |H_r| .$$

Clearly K dominates H^n; and because H has continuous sample paths, so does K, which means it is locally bounded. In particular K will satisfy the integrability condition (iii) from DCT.

So by the previous proposition, we can conclude that $\int_0^t H_s^n \, dX_s \to \int_0^t H_s \, dX_s$ in probability.

And then because of our definition of the step function H^n, the left-hand side is just the Riemann sum in our desired proposition,a nd the right-hand side is also the same. So the proof just follows. □

Remark 15.6. Here there's an important remark: It's crucial that we're approximating H by its values at the *left* endpoints. There are two reasons why this is important. First, suppose we took not the left endpoint, but e.g. the right endpoint t_{i+1}. Then the value of H^n at time t_i^n would depend on the value of H at time t_{i+1}^n, which might not be $\mathcal{F}_{t_i^n}$-measurable. And even if adaptedness was satisfied, taking a different point in the interval may give you a different value in the limiting Riemann sum. Actually for instance, if we took the midpoint (for each interval, we took $(t_i^n + t_{i+1}^n)/2$ here), this leads to a certain stochastic integral called the *Stratonovich integral*, which has different values than the Ito integral; we will see this later.

So it's important to take the left endpoint, or else the fluctuations of the CSMG might lead you to a different limit value.

§15.3 Ito's formula

Now we'll introduce the main topic of today's lecture, Ito's formula. In short, this formula shows that sufficiently regular functions of CSMGs are also CSMGs, and we can write out explicitly their final decompositions as a CLMG plus a FV process.

Theorem 15.7 (Ito's formula)

Let $X^1, \dots, X^p$ be CSMGs. Let $F : \mathbb{R}^p \to \mathbb{R}$ be $\mathcal{C}^2$. Then for every $t \geq 0$, we have

$$F(X_t^1, \dots, X_t^p) = F(X_0^1, \dots, X_0^p) + \sum_{i=1}^{p} \int_0^t \partial_i F(X_s^1, \dots, X_s^p)\, dX_s^i + \frac{1}{2} \sum_{i,j=1}^{p} \int_0^t \partial_{ij} F(X_s^1, \dots, X_s^p)\, d\langle X_i, X_j \rangle_s.$$

So we write $F(X_t^1, \dots, X_t^p)$ as the sum of three terms. The first is just a constant — the value of F at time 0. The second term is a sum of p stochastic integrals. In each, the integrand is a first-order derivatie of f (wh is icontinuous), and we're integrating against one CSMG. So each integral should again be a CSMG. And we can write out the explicit canoncail decomposition — it's $\int \partial_i F\, dM_s^i + \int \partial_i F\, dV_s$ (wheere the secodn part is FV and the third a CLMG). And the third is $1/2$ times a sum of p^2 terms. Each is a function of the X's integrated against the QV of X_i and X_j, which is an increasing process. So the third term is a FV process.

And in all, we can see that $F(x_1, \dots, x_p)$ is again CSJG. The continuous local part is given by the second term, and the FV part is given by both the second and third term.

In SDE notation, we can take a formal derivative and write

$$dF(X_t^1, \dots, X_t^p) = \sum_{i=1}^{p} \partial_i F(X_1, \dots, X_t)\, dX_s^i + \frac{1}{2} \sum_{i,j=1}^{p} \partial_{ij} F(X_1^s, \dots, X_p^s)\, d\langle X_i, X_j \rangle_s$$

(maybe all s's should be t's).

§15.4 Examples

Before going into the proof, let's first look at some examples.

Example 15.8

Take $p = 1$, and take $X_t^1 = B_t$ to be a standard 1-dimensional Brownian motion. Then Ito's formula tells us that

$$dF(B_t) = F'(B_t)\, dB_t + \frac{1}{2} F''(B_t)\, d\langle B, B \rangle_t.$$

Recall that $\langle B, B \rangle_t = t$, so we can replace this by

$$dF(B_t) = F'(B_t)\, dB_t + \frac{1}{2} F''(B_t)\, dt.$$

Example 15.9

In the higher-dimensional case, we can take $(X_t^1, \dots, X_t^p)$ as a standard p-dimensional BM $(B_t^1, \dots, B_t^p)$, which means that each X_t^i is a standard one-dimensional BM, and these p BMs are jointly independent. Then Ito's formula tells us that

$$dF(B_t^1, \dots, B_t^p) = \sum_{i=1}^{p} \partial_i F'(B_t^1, \dots, B_t^p)\, dB_t^i + \frac{1}{2} \sum_{i \neq j} \partial_{ij} F(B_1^p, \dots, B_t^p)\, d\langle B^i, B^j \rangle_t.$$

And we know that

$$\langle B_i, B_j \rangle = \begin{cases} t & i = j \\ 0 & i \neq j \end{cases}$$

(in the second case they're independent). So the left-hand side is just

$$\sum_{i=1}^{p} \partial_i F\, dB_t^i + \frac{1}{2} \sum_{i=1}^{p} \partial_{ii} F\, dt.$$

We xan write this in short as

$$\nabla F \cdot dB_t + \frac{1}{2} \Delta F\, dt$$

(where ∇F and dB_t are both p-dimensional, and we're taking a vector product; and ΔF is the Laplacian).

We can also read Ito's formula as a stochastic version of the chain rule. But it's different from the ordinary chain rule in calculus becaiuse of the extra second-ordre term.

In ordinary calculus, we know $(f(g(t)))' = f'(g(t))g'(t)\, dt$. In SDE notation, wee might write $d(f(g(t))) = f'(g(t))g'(t)\, dt$. And we can combine the last two terms together tow rite $d\, g$. This means if X_t is not a CSMG but a usual regular process, then we shoudl have $df(X_t) = f'(X_t)\, dX_t$ (this is the ordinary chain rule).

But now, when X has stronger fluctuations, we actually have

$$df(X_t) = f'(X_t)\, dX_t + \frac{1}{2} f''(X_t)\, d\langle X, X \rangle_t.$$

Recall that the QV can be approximated by a direct sum of squares — we could also write $d\langle X, X \rangle_t$ as $dX_t dX_t$.

What's the intuition behind why we have an extra second-order term here? Recall that for a CSMG, we have

$$\langle X, X \rangle_t = \langle M, M \rangle_t.$$

And we can take a formal derivative and write

$$d\langle X, X \rangle_t = d\langle M, M \rangle_t.$$

Then we know that because X is just a sum of a CLMG plus a FV process, the LHS is actually a sum of four terms — it's

$$d\langle M, M \rangle_t + d\langle M, V \rangle_t + d\langle V, M \rangle_t + d\langle V, V \rangle_t,$$

just using the linearity of QV to expand. And then because the QV can be approximated by a sum of squares, $d\langle M, M \rangle_t$ is somehow equal to

$$dM_t\, dM_t = \sum (M_{t_i} - M_{t_{i-1}})^2$$

(taking mesh size to 0).

Meanwhile, on the left we have four terms — $dM_t dM_t + dM_t dV_t + dV_t dM_t + dV_t dV_t$.

And in the lecture on CSMGs, Sky gave an intuitive explanation to the eequality here. We shoudl roughly think of $dM_t \sim (dt)^{1/2}$, and $dV_t \sim dt$. Then $dM_t dM_t$ on the right has order dt. And $dM_t dM_t$ has order dt, $dM_t dV_t$ and $dV_t dM_t$ are $(dt)^{3/2}$, and $dV_t dV_t$ $(dt)^2$. Then when summing over t, all lower-order terms (with power of dt bigger than 1) should vanish in the limit. So these three terms should vanish, because their powers are bigger than 1. So that's why when taking a limit, we only have the term $dM_t\, dM_t$ left on the left and right, and that's why we should have $\langle X, X\rangle_t = \langle M, M\rangle_t$.

And the same intuition actually applies here. We can use Taylor's formula to formally write

$$dF(X_t) = F'(X_t)\, dX_t + \frac{1}{2}F''(X_t)\, dX_t\, dX_t + \text{lower order terms}$$

(where the lower order terms are $o(dX_t\, dX_t)$). The first term $F'(X_t)\, dX_t$ is of order $(dt)^{1/2} + dt$. So this is not a lower-order term; when summing over t it shouldn't vanish, andi t gives us the first term.

The second term is of order $(dt)^{1/2\cdot 2} = dt$. So this also should not vanish, and it actually gives us the quadratic variation term in the formula.

And then all the lower order terms have a power of dt bigger than 1, so they all vanish when summing over t. So that's why Ito's term has a first-order and second-order term, but no others.

Meanwhile, in the usual chain rule, the first term is of order dt and should be kept, but the second term is already of order $(dt)^2$. That's why in the ordinary chain rule we only have one term on the right-hand side.

So that's kind of an intuitive explanation of Ito's formula. And now we're going to build a rigorous proof on this intuitive idea.

§15.5 Proof of Ito's formula

Let's first take the univariate case $p = 1$. Then to have this discrete sum story, we're going to take a sequence of subdivisions $\{t_i^n\}_{0\le i\le p_n}$ with mesh size going to 0. Here we'll also require the subdivisions are increasing (meaning any subdivision point remains a subdivision point as n gets larger).

Let's write

$$F(X_t) - F(X_0) = \sum_{i=0}^{p_n-1} F(X_{t_{i+1}}) - F(X_{t_i}).$$

And then we're going to use Taylor's expansion. This is kind of the $dF(X_t)$ term in the above intuition, and after Taylor expansion we should have that this is equal to

$$\sum_{i=0}^{p_n-1} F'(X_{t_i})(X_{t_{i+1}} - X_{t_i}) + \frac{1}{2}\sum_{i=0}^{p_n-1} F''(X_{t_i})(X_{t_{i+1}} - X_{t_i})^2 + \cdots,$$

wheere the rest are lower-order terms. To simplify the argument, we'll drop the lower-order terms. Instead of using the Peano form of Taylor expansion, we'll use the Lagrange form; so we instead write

$$\sum_{i=0}^{p_n-1} F'(X_{t_i^n})(X_{t_{i+1}^n} - X_{t_i}) + \frac{1}{2}\sum_{i=0}^{p_n-1} F''(\theta_i^n)(X_{t_{i+1}} - X_{t_i})^2,$$

where $\theta_i^n \in [X_{t_i^n}, X_{t_{i+1}^n}]$. Then we want to prove that the right-hand side here converges to the sum of our two terms as the mesh size goes to 0.

The first part is comparatively easier to handle — we can just apply the proposition above. This term is exactly a Riemann sum where we're taking the left endpoint, so this already converges to

$$\int_0^t F'(X_s)\,dX_s$$

as the mesh size goes to 0. So it only remains to show that the second term converges to the third term in Ito's formula.

So let's write the term there as a sum of two terms, as

$$\sum_{i=0}^{p_n-1} F''(\theta_i^n)(X_{t_{i+1}^n} - X_{t_i^n})^2 = \sum_{i=0}^{p_n-1} (F''(\theta_i^n) - F''(X_{t_i}^n))(X_{t_{i+1}^n} - X_{t_i^n})^2 + \sum_{i=0}^{p_n-1} F''(X_{t_i^n})(X_{t_{i+1}^n} - X_{t_i^n})^2.$$

The second term is again a Riemann sum taking left endpoints, and the first term is sort of an error term. So it suffices to show that the second term converges to the appropriate stochastic integral (as the mesh size goes to 0), and the first term converges to 0.

First, we know that the term $F''(\theta_i^n) - F''(X_{t_i^n})$ can be uniformly bounded — we have

$$\max_i \left|F''(\theta_i^n) - F''(X_{t_i^n})\right| \le V_{\Delta_n} := \sup_{s_1,s_2\in[0,t],|s_1-s_2|\le\Delta_n} \left|F''(s_1) - F''(s_2)\right|$$

(where Δ_n is the mesh size of the nth subdivision). And because of our assumption that $F \in C^2$, we know that F'' is uniformly continuous on the interval $[0,t]$. Then because our mesh size Δ_n is going to 0, we have $V_{\Delta_n} \to 0$ (almost surely).

So the first term here is bounded by

$$V_{\Delta_n} \cdot \sum_{i=0}^{p_n-1} (X_{t_{i+1}^n} - X_{t_i^n})^2.$$

And we know the sum converges to the QV of X at time t (in probability) while the first part converges to 0. So their product should converge to $0 \cdot \langle X, X\rangle_t = 0$.

So this part converges to 0 in probability. Now it just remains to show that the second term converges to

$$\sum_{i=0}^{p_n-1} F''(X_{t_i^n})(X_{t_{i+1}^n} - X_{t_i^n})^2 \to \int_0^t F''(X_t)\,d\langle X, X\rangle_t.$$

But recall that before, we learned a sequence of RVs X_n converge to X in probability if and only if for every subsequence X_{n_k}, there exists a further subsequence that converges almost surely. Here we're going to apply this principle, and we'll show that we have almost sure convergence along a suitable subsequence. This will suffice to give the convergence in probability.

Let's rewrite this sum as an integral — this is equal to

$$\int_0^t F''(X_s)\,d\mu_n(s),$$

where μ_n is a random measure given by a finite sum of Dirac measures — $\mu_n = \sum_{i=0}^{p_n-1}(X_{t_{i+1}^n} - X_{t_i^n})^2\,\delta_{\{t_i\}}$. (This is just a reformulation of the original discrete sum.)

We're also going to rewrite the right-hand side as

$$\int_0^t F''(X_s)\,d\mu(s),$$

where μ is the measure induced by the QV of X (which is itself an increasing continuous process).

Now what we want to show is that the expectation of F'' under μ_n converges, along some subsequence, almost surely, to $\mathbb{E}[F'']$ under the random measure μ. This should remind you of the notion of weak convergence — somehow we want to prove that the random measures μ_n converge to μ in the notion of weak convergence, almost surely along a subsequence. Then because $F \in C^2$, we'll have that the expectation converges correspondingly.

Let D be the union of all the subdivisions, i.e., $D = \bigcup_n \{t_0^n, \dots, t_{p_n}^n\}$. Because we're taking increasing subdivisions with mesh size tending to 0, we have that D is dense in $[0, t]$.

Now because we're taking *increasing* subdivisions, which means every subdivision point remains a subdivision point as n gets larger, this implies that for every $s \in D$, as long as n is sufficiently large, there exists an index $p_s^{(n)}$ such that $t^n_{p_s^{(n)}} = s$ (this means that in the nth subdivision, s is a subdivision point).

Then again using the fact that the QV can be approximated by a discrete sum, we have that

$$\sum_{i=0}^{p_s^{(n)}-1} (X_{t_{i+1}^n} - X_{t_i^n})^2 \to \langle X, X \rangle_s$$

in probability as $n \to \infty$.

Then by the definition of μ_n and μ, the LHS is the mass of $[0, s]$ under the measure μ_n (i..e, $\mu_n([0, s])$), and the RHS is exactly $\mu([0, s])$.

So combining these equations, we know $\mu_n([0, s]) \to \mu([0, s])$ in probability, for any $s \in D$. Because D is countable, we can use a diagonal argument to obtain a subsequence of μ_n — say μ_{n_k} — such that $\mu_{n_k}([0, s]) \to \mu([0, s])$ for every $s \in D$, almost surely. (This is kind of a combination of two steps. The first is because we have convergence in probability, we can extract a further subsequence which converges almost surely. Then for each s we have a subsequence which converges almost surely, and we can use a diagonal argument to extract a further subsequence where we have convergence almost surely for every $s \in D$.)

This is enough to deduce that $\mu_{n_k} \to \mu$, in the sense of weak convergence of finite measures (at least, on the interval $[0, t]$), almost surely. Here we're using a portmanteau theorem which gives you a bunch of equivalent definitions of weak convergence, and we're using the property that μ is almost surely absolutely continuous with respect to the Lebesgue measure; then because we have almost sure convergence on a dense set, we also have almost sure weak convergence.

And then on this almost sure event, we will have that $\mathbb{E}_{\mu_{n_k}}[F''(X_s)\mathbf{1}_{[0,t]}] \to \mathbb{E}_{\mu}[F"'(X_s)\mathbf{1}_{[0,t]}]$ (this is also the portmanteau theorem, which says that if we have weak convergence, then we have a convergence of expectation of any bounded continuous function, and vice versa).

Finally, by the definition of μ_n and μ, this expectation is just the integral

$$\int_0^t F''(X_s)\, d\mu_n(s)$$

from before, and the expectation with respect to μ is the integral with μ in place. So we've shown that these integrals, and therefore our discrete sums, converge almost surely along a subsequence.

So the univariate case of Ito's formula follows.

We will stop here, and next time Sky will continue to prove Ito's formula in the multivariate case.

§16 April 2, 2025

§16.1 Itô's formula

Today we'll start by finishing the proof of Ito's formula, whose statement we recall here.

Theorem 16.1 (Itô's formula)

Let $X^1, \dots, X^p$ be CSMGs. Let $F \in \mathcal{C}^2(\mathbb{R}^p, \mathbb{R})$. Then

$$dF(X_t^1, \dots, X_t^p) = \sum_i (\partial_i F)(X_t^1, \dots, X_t^p)\, dX_t^i + \frac{1}{2}\sum_{i,j} (\partial_{ij} F)(X_t^1, \dots, X_t^p)\, d\langle X^i, X^j\rangle_t.$$

It's basically the stochastic chain rule — as we tried to emphasize in previous classes, the usual chain rule in calculus doesn't hold when you have semimartingales. In particular, when you have the martingale part, which isn't C^1 but only $C^{1/2}$ (or just below 1/2-Hölder), you have to expand to second order in your chain rule. Usually this second term is called the *Ito correction*.

We should know this formula by heart. It's hopefully quite intuitive — where does it come from? You basically do a second-order Taylor expansion, and you basically use that in your Taylor expansion, whenever you see $dX_t^i\, dX_t^j$, this basically becomes a quadratic variation term $d\langle X_t^i, X_t^j\rangle$. (And the first-order term of the Taylor expansion gives the first thing — the usual term from the chain rule. But then because this thing is not necessarily 0 for martingales, you get the Ito correction term.)

Here we've written the formula in SDE notation; we're sometimes going to omit the summation over indices i and j, so we'll write this as

$$dF(X_t^1, \dots, X_t^p) = (\partial_i F)(X_t^1, \dots, X_t^p)\, dX_t^i + \frac{1}{2}(\partial_{ij} F)(X_t^1, \dots, X_t^p)\, d\langle X^i, X^j\rangle_t$$

(where the sums are implicit).

Last time we proved the univariate case, so now we'll finish with the multivariate case.

Proof. Take a sequence of subdivisions with mesh size tending to 0. Then as before, we want to Taylor-expand $F(X_{t_{i+1}^n}) - F(X_{t_i^n})$. If you do this to second-order (you always Taylor expand around the left endpoint), you get

$$(\partial_j F)(X_{t_i^n})(X_{t_{i+1}^n} - X_{t_i^n}) + \frac{1}{2}(\partial_{jk} F)(\theta_{ijk}^n)(X_{t_{i+1}^n}^j - X_{t_i^n}^j)(X_{t_{i+1}^n}^k - X_{t_i^n}^k),$$

where θ_{ijk}^n is on the line segment between $X_{t_{i+1}^n}$ and $X_{t_i^n}$ (but these points are close together because your mesh size is tending to 0, so you should basically think of it as the left endpoint $X_{t_i^n}$ — that's basically how we're going to use it).

As before, if you sum over i, you'll have that the first-order terms converge to the stochastic integral — we have

$$\sum_i (\partial_j F)(X_{t_i^n})(X_{t_{i+1}^n} - X_{t_i^n}) \to \int_0^t (\partial_j F)(X_t)\, dX_t^j$$

in probability as $n \to \infty$. And for the second-order terms, if you again sum over i, you'll have this term θ_{ijk}^n, and you want to approximate the thing by what you get if you replace θ with the left endpoint, so let's do that: we have

$$\sum_j \partial_{jk} F(\theta_{ijk}^n) - \partial_{jk} F(X_{t_i^n}) \cdot (X_{t_{i+1}^n}^j - X_{t_i^n}^j)(X_{t_{i+1}^n} - X_{t_i^n}^k) \to 0$$

as $n \to \infty$, by the same considerations as in the univariate case (because if we ignore the difference in $\partial_{jk} F$, the rest converges to something of order 1 — coming from the quadratic variation — and this difference converges to 0, because $\partial_{jk} F$ is continuous and θ_{ijk}^n gets arbirarily close to $X_{t_i^n}$). So to finish, all we need to show is that

$$\sum_i \partial_{jk} F(X_{t_i^n})(X_{\dots}^j - X_{\dots}^j)(X_{\dots}^k - X_{\dots}^k) \to \int_0^t \partial_{jk} F(X_t)\, d\langle X_t^j, X_t^k\rangle$$

in probability (at least, we'll show this along a subsequence). But basically you can just argue as before — you can find that there exists a subsequence, which we'll still denote as $\{(t_i^n)\}$, such that almost surely, for all j and k, if you define the measure

$$\mu_n^{jk} = \sum_{i=0}^{p_n-1} (X^j_{t^n_{i+1}} - X^j_{t^n_i})(X^k_{t^n_{i+1}} - X^k_{t^n_i})\delta_{\{t^n_i\}}$$

(basically, this is a discretized version of the quadratic variation), then as a measure, this converges to $d\langle X^j, X^k\rangle$ as $n \to \infty$. (The covariation is a FV process, so it always corresponds to some finite measure (at least on finite intervals); this convergence means you're converging to that measure in the sense of weak convergence.) And the sense of weak convergence is enough to imply the thing we want (since weak convergence by definition means that integrals of continuous functions converge.)

And why will this be true? (We will sometimes call the thing 'covariation,' but Le Gall calls it the 'bracket.')

We defined the bracket as a linear combination of QVs — specifically, you can use that

$$\langle X, Y\rangle = \frac{1}{2}(\langle X+Y, X+Y\rangle - \langle X, X\rangle - \langle Y, Y\rangle).$$

And if you look at any individual QV term, you can find a subsequence where we have this convergence (we did that in the univariate case). And there's only finitely many j and k, so there's only finitely many QV terms to consider. So you can find a subsequence which works for all of them, and along that subsequence you will have that this convergence holds.

And with that, that basically wraps up the proof of Itô's formula. □

Remark 16.2. This last part is maybe the most technical part, but modulo that it's very believable that this should be true — it's just second-order Taylor expansion, using the rules of stochastic rather than regular calculus.

§16.2 Some examples

Example 16.3

Suppose we take $F(x, y) = xy$. Then you'll have

$$d(X_tY_t) = X_t\,dY_t + Y_t\,dX_t + d\langle X, Y\rangle_t.$$

There's various ways you could see this. You can just think about applying Ito's formula — if you take the derivative in x you get the $Y_t\,dX_t$ term, and if you take the derivative in y you get the $X_t\,dY_t$ term, and when you take the derivative in both x and y you should be getting the $d\langle X, Y\rangle_t$ term. Specifically, we have $\partial_{xx}F = 0$, $\partial_{yy}F = 0$, and $\partial_{xy}F = \partial_{yx}F = 1$. Each one is 1, so there's two of them, and that cancels the $\frac{1}{2}$.

This is always kind of confusing. The way Sky would rather think about this type of calculation — and how you compute things in practice — is that you think about applying the rules of usual calculus, but now you allow yourself to take d twice. At least at a formal level, you can write

$$d(X_tY_t) = (dX_t)Y_t + X_t(dY_t) + dX_tdY_t.$$

(There's no terms with d^2X_t or d^2Y_t, where the d falls on the same variable twice, because these are 0.) And this last thing is going to give $d\langle X, Y\rangle_t$ (it's a different notation for precisely this).

The way you think about this calculation is that infinitesimally, this should mean that if I consider differences of the form

$$X_{t+\varepsilon}Y_{t+\varepsilon} - X_tY_t,$$

I should get something like

$$(X_{t+\varepsilon} - X_t)Y_{t+\varepsilon} + X_t(Y_{t+\varepsilon} - Y_t).$$

But this first thing is at $Y_{t+\varepsilon}$, while I wanted it to be at Y_t. So now I can write this as

$$(X_{t+\varepsilon} - X_t)Y_t + X_t(Y_{t+\varepsilon} - Y_t) + (X_{t+\varepsilon} - X_t)(Y_{t+\varepsilon} - Y_t).$$

The reason I want to do that is that the first two stochastic integrals are defined by taking the left endpoint of your integrand. So the first thing gives me the $Y\,dX$ term, the second the $X\,dY$ term, and the third gives the QV (because that's what the QV is on a discrete level).

As a special case of this example:

Example 16.4

If $X = Y$, we have $d(X_t^2) = 2X_t\,dX_t + d\langle X, X\rangle_t$.

As we mentioned before, we basically had to prove this when constructing the QV — so Ito's formula kind of recovers this, so this is kind of circuilar in some sense. But in some sense, what this says is that as soon as you prove Ito's formula for the square function, you get it in general — because all you needed was convergence for the stochastic integral and for the QV terms, and for the QV terms was by far the most technical part. And in the proof when we constructed the QV, we basically took this difference and showed that the martingale term converged to a martingale, which was basiclaly $2X_t\,dX_t$. When we constructed the QV, basically at a discrete level we identified a martingale, and we showed that as you take mesh size to 0 that martingale converges to this thing $2X_t\,dX_t$. Once you get that, you get that the discrete approximations to $d\langle X, X\rangle_t$ converge. So we said that the proof of QV would make more sense once you know Ito's formula, because it's kind of a special case of that.

More generally, if you have more products:

Example 16.5

We have $d(X^1 \cdots X^p) = \sum_{i=1}^p (\prod_{j\neq i} X_j)\,dX^i + \sum_{i<j} \prod_{k\neq i,j} X_k\,d\langle X^i, X^j\rangle$

Heuristically, when we take the d of the LHS, there's the terms where only one d appears, and it can appear on any of the X's. And then you have the terms where two d's appear. This is what you'd get if you think about putting the d on two individual terms. (The ordering is what removes the $\frac{1}{2}$ factor from Ito's formula.)

Example 16.6

If B_t is a d-dimensional Brownian motion and F is a function of d variaboes, then

$$dF(B_t) = \partial_j F(B_t)\,dB_t^j + \frac{1}{2}\Delta F(B_t)\,dt.$$

And usually when you have Brownian motion, you don't write this summation over indices; you just write this as

$$dF(B_t) = \nabla F(B_t)\cdot dB_t + \frac{1}{2}\Delta F(B_t)\,dt.$$

Example 16.7

If F is a function of $d+1$ variables and we compute $dF(t, B_t)$ (note that t is a CSMG — it's actually just a FV process), we get

$$dF(t, B_t) = \nabla F(B_t)\, dB_t + (\partial_t + \frac{1}{2}\Delta)F(t, B_t)\, dt..$$

Here in the first term the gradient is only over the last d variables (not t) — if you want to separate into a martingale part and a FV part, you want to group the dt part with the rest. You never see a derivative in t and in one of the spatial variables — because if you ever see something like that, then you're computing the QV $dt\, dX_t^i$, and this is 0 because the bracket of a FV process and a semimartingale is always 0. (So if it's the time variable, you only see one variable.)

What does this mean?

Fact 16.8 — If you have a *harmonic* function, with $\Delta F = 0$, then $F(B_t)$ is a local martingale.

And if F is C^2 with bounded first derivative, then it should actually be a martingale (not just a local martingale) — because the first term is bounded, so it's the stochastic integral of a bounded integrand, which should give you a martingale thing.

And now in the second example:

Fact 16.9 — If F solves the partial differential equation $(\partial_t + \frac{1}{2}\Delta)F = 0$, then $F(t, B_t)$ is a local martingale.

And if F is smooth and bounded, then it's actually a martingale.

You might recognize that this almost looks like the heat equation, but with a $+$ instead of $-$ — it's kind of a backwards heat equation. You can still relate solutions to the backwards heat equation to solutions to the heat equation, as follows.

If you have a solution to the heat equation $(\partial_t - \frac{1}{2}\Delta)u = 0$ where $u : [0, T] \times \mathbb{R}^d \to \mathbb{R}$, then if you define

$$F(t, x) = u(T - t, x)$$

and then you compute $\partial_t F + \frac{1}{2}\Delta F$, the time-derivative of F is the negative of that of u, so this is going to be $(-\partial_t u + \frac{1}{2}\Delta u)$ (where if the left was evaluated at (t, x), then this one is evaluated at $(T - t, x)$). So at least on finite intervals, you can get solutions to the backwards heat equation from ones to the forwards one.

And if you have the martingale part, that gives you a formula for F and therefore u — you're basically saying

$$\mathbb{E}F(T, B_T) = \mathbb{E}F(0, B_0).$$

If you let B start at some point x and we plug in our formula for F, the right-hand side is going to be $\mathbb{E}[u(0, B_T)]$. So we get the formula

$$u(T, x) = \int \rho_T(x - y)u(0, y),$$

where ρ_T is the density of $\mathcal{N}(0, T)$. This is a formula that basically relates the solution to the heat equation to the density of BM.

So assuming you actually have a martingale and not a local martingale (which shouldn't be too bad to justify if you assume $u \in C^\infty$ and all its derivatives are bounded), then you get a formula for the solution to the heat equation in terms of the original value. Usually with the heat equation you're solving some

evolution thing — you set some initial value at time 0, and want to understand the solution at later times. This is an explicit formula for that, where you're just convolving against Gaussian density.

This is one application of Ito's formula. (It's probably in Chapter 7 of Le Gall.)

If you think about doing something for the harmonic case $\Delta F = 0$, you can get similar formulas. It makes more sense to think about harmonic functions on some domain (e.g. a ball of radius 1) with some boundary conditions. And then you'll be able to find a formula for the value of the harmonic function inside the ball, related to its boundary values. (This is related to the mean equation that you can also prove by other means in a PDE class.)

§16.3 Stratonovich integral

Now that we've defined the Ito integral and proved Ito's formula, we'll talk about this other stochastic integral. It's maybe not mentioned in the textbook, but it's an important thing to know because sometimes it comes up.

First, Ito's formula is a stochastic chain rule. But you might ask a question:

> **Question 16.10.** Is there a stochastic integral which satisfies the *usual* chain rule?

So maybe you don't want to work with this Ito correction term; is there a way to define these integrals (differently from the Ito integral, of course) such that the actual chain rule holds? The answer is yes, nad it's called the Stratonovich integral. At some point (maybe next class), we'll give an example of why you would want the chain rule to hold.

How would you guess the definition of the Stratonovich integral — how would you guess a formula for the stochastic integral for which the chain rule holds?

Well, let's go back to why the usual chain rule doesn't hold — why you get the Ito correction term. At least in the univariate case, when you're proving Ito's formula, all you're doing is a second-order Taylor expansion

$$F(X_{t+\eta}) - F(X_t) = F'(X_t)(X_{t+\eta} - X_t) + \frac{1}{2}F''(X_t)(X_{t+\eta} - X_t)^2 + \text{lower-order terms}.$$

And the fact that the second thing might be order-1 in the limit gives you this first-order correction.

You see this because you tried expanding around the left endpoint. What if you tried expanding around the right endpoint — so we're expanding around $F(X_{t+h})$ instead? If you do that, you'll get

$$F(X_{t+h}) - F(X_t) = F'(X_{t+h})(X_{t+h} - X_t) - \frac{1}{2}F''(X_{t+h})(X_{t+h} - X_t)^2 + \text{lower-order terms}$$

(it's a $-$ because if you took $F(X_t) - F(X_{t+h})$ you'd get a $+$).

Now if you stare at these two identities, what you notice is that if you average the two identities, you get

$$F(X_{t+h})-F(X_t) = \frac{1}{2}(F'(X_t)+F'(X_{t+h}))(X_{t+h}-X_t)+\frac{1}{2}(F''(X_t)-F''(X_{t+h}))(X_{t+h}-X_t)^2+\text{lower-order terms}.$$

But now you notice that because you did this averaging, this difference $F''(X_t) - F''(X_{t+h})$ (assuming $F \in C^2$) — these two values are very close so this difference is going to be $o(1)$. Now if you repeated the proof of Ito's formula, more or less, you would actually get that this quadratic term disappears. Because before you had $F''(X_t)(X_{t+n} - X_t)^2$, which in the limit is $O(1)$. But now you have this difference of F''s, and that difference is $o(1)$, so the whole product is going to be $o(1)$.

So what that leads you to guess is that you want to define a stochastic integral such that

$$\int_0^t f(X_s) \circ dX_s = \lim_{n\to\infty} \frac{1}{2}(f(X_{t_i^n}) + f(X_{t_{i+1}^n}))(X_{t_{i+1}^n} - X_{t_i^n})$$

(where we use $\circ$ as new notation to indicate that this is different from Ito integration). Why? At a discrete level dF is whatever we had on the right-hand side. If we take limits, we expect the second term is disappearing, so we expect that

$$df(X_t) = f(X_t) \circ dX_t.$$

So at least in this specific for,ula where I'm integrating $f(X)$ against dX, I want the limit to be given by this discrete approximation of endpoints.

Why dos this limit exist? It's actually a very simple reason. What you just notice is that you can write the right-hand side as

$$\lim_{n\to\infty} \sum_i \cdots$$

(you have the average of two endpoints; you can write that as the sum of the left endpoint and the difference between the two endpoints). So you just use that

$$\frac{1}{2}(f'(X_{t_i^n}) + f'(X_{t_{i+1}^n})) = f'(X_{t_i^n}) + \frac{1}{2}(F'(X_{t_{i+1}^n}) - F'(X_{t_i}^n)).$$

Then you get that

$$\text{RHS} = \int_0^t F'(X_s)\, dX_s + \frac{1}{2}\langle F'(X)\, X\rangle_t\, dt.$$

(Suppose F is smooth, so there's no issues about regularity — then $F' \in C^2$, so Ito's formula tells you $\langle F(X), X\rangle$ is well-defined. And the difference term does converrge to this.)

So in summary, now we have a guess for what the Stratonovich integral should look like, at least for $f(X)$ integrated against X. More generally:

Definition 16.11 (Stratonovich integral)**.** If X and Y are CSMGs, the *Stratonovich integral* of X against (or with respect to) Y is defined as

$$\int_0^t X_s \circ dY_s = \int_0^t X_s\, dY_s + \frac{1}{2}\langle X, Y\rangle_t.$$

This is motivated by the fact that if $X = f(x)$ and F''.

In particular, because X and Y are both CSMGs, so is the Stratoniv j process as a process in p.

We've made this definition, but let's verify that the corresponding discrete approximation does converge. We saw the special case where you integrate$f'(x)$ against x, but now we'll see why oit converges in general.

Proposition 16.12

Let X and Y be continuous semimartingales, and take a sequence of subdivisions $\{t_i^n\}$ with mesh size tending to 0. Then

$$\sum_{i=1}^{p_n} \frac{X_{t_{i+1}} + X_{t_i}}{2} \cdot (Y_{t_i} - Y_{t_{i-1}}) \to \int_0^t X_s\, dY_s$$

(in probabiltty).

The proof is about the same — you just notice that

$$\frac{1}{2}(X_{t_{i-1}} + X_{t_i}).$$

This is what you shoudl remember about the Stratonovich integral — that the above is the discrete approximation that gives you the integral. So Ito is left endpoint, and Stratonovich is the average of endpoints. If you just remember this, then you will be able to recover the definition fo the Stratonovich integral, using the above algebraic identity.

Remark 16.13. In SDE notation,w eoften write $dZ = X \circ Y$ to denote that $Z_t = Z_0 + \int_0^t X_s \, dY_s$ (similarly to the Ito sense). This also means $dZ = X \, dY + \frac{1}{2} d\langle X, Y \rangle$.

One finaly thing we might write is that $\circ \, dZ$ basically means $\mathbf{1} \circ dZ$. ANd if you integrate 1 against intself, that's just the Ito integral $1 \, dZ = dZ$. Another way to say this is if you take your integrand to be 1, it does't mattter which endpoint you take.

The way you see this is that the bracket of 1 and anything else is just 0, so then you're just elft with the Iot erm.

§16.3.1 Chain rule

Those are some considerations on the Stratonovich integral; but now let's prove the chain rule. We guessed the definition fo the Stratonovich integral, but now starting from this integral we want to show that the chain rule is actually satisfied. So let's show that.

Proposition 16.14 (Chain rule)
Let $X \in \mathbb{R}^d$ be a CSMG, and let $f \in C^3(\mathbb{R}^d, \mathbb{R})$. Then

$$df(X) = \nabla f(X) \circ dX = \partial_j f(X) \circ dX^j.$$

Three derivatives might not be needed — kind of what you'd expect is all you should need for f is that the difference $F''(X_t) - F''(X_{t+h})$ should be lower-order, so you might just need F to be twice continuously differentiable. But the way we're going to prove this is by using Ito on F', and in Ito you assume your function is C^2, so we need to assume $F \in C^3$. But if you're more refined with the proof, you might be able to just prove this for C^2.

Proof. The first thing is you want to calculate df. Ito's formula tells ius that

$$df(X) = \partial_j f(X) \, dX^j + \frac{1}{2} \partial_{ij} f(X) \, d\langle X^i, X^j \rangle.$$

And somehow we have to say why this is goingto be equal to the thing we wrote. So now let's start from the other side of the identity we want to show — let's compute what

$$\partial_j f(X) \circ dX^j$$

is supposed to be. Well, by definition the Stratonovich integral is

$$\partial_j f(X) \circ dX^j = \partial_j f(X) \, dX^j + \frac{1}{2} \langle \partial_j f(X), dX^j \rangle.$$

How do we compute this bracket? For that, we need to be able to decompose $\partial_j f(X)$ into a martingale part and a FV part, and we only care about the martingale part (the FV part times dx will be 0). And that's exaclty what Ito's formula gives us (it gives a semimartingale and a FV part). So you just insert Ito's formula into this bracket — not this version of Ito's formula, but if we apply Ito's formula to $\partial_j f$ (that's the thing we want to decompose). So we apply Ito's formula to say

$$\partial_j f(X) = \partial_{jk} f(X) dX^k + \frac{1}{2} \partial_{jki} f(X) \, d\langle X^k, X^i \rangle$$

(we applied Ito's formula to $\partial_j f$ instead of f,which is why we have three derivatives). Now if I insert the RHS into the bracket $\langle \partial_j f(X), dX^j \rangle$, the second term is going to disappear (any FV thing times dX is 0). So the only thing I have to worry about when computing the bracket is the first term, which means

$$\langle \partial_j f(X), dX^j \rangle = \langle \partial_{jk} f(X)\, dX^k, dX^j \rangle = \partial_{jk} f(X)\, d\langle X^k, X^j \rangle$$

(where we pull things out using bilinearity). (This is where SDE notation really helps — imagine writing this out with integrals everywhere.)

And now this is precisely what we computed from Ito's formula for df (where k plays the role of i) — we wanted

$$\partial_{ij} f(X)\, d\langle X^i, X^j \rangle.$$

(Note that because you're C^2, the derivatives commute.) □

Remark 16.15. If you wanted to go into the details of the proof, you could probably improve the C^3 assumption to a C^2 assumption.

§16.4 Stochastic processes on manifolds

Why do you want to use Stratonovich? The general application is that in geometric settings, it's good to use it. What do we mean by geometric settings? If you want to define stochastic processes on manifolds, you want to use the Stratonovich integral because it satisfies the chain rule.

Example 16.16

Let's consider processes on the sphere — this is a very explicit case where your manifold is actually a subspace of Euclidean space.

So we let $d \geq 1$, and $\mathbb{S}^d = \{x \in \mathbb{R}^{d+1} \mid |x| = 1\} \subseteq \mathbb{R}^{d+1}$ (this is some sub-manifold of $\mathbb{R}^{d+1}$). Fo ra general $x \in \mathbb{R}^{d+1}$ which is nonzero, let's define the *projection operator* $P_x \in \mathrm{End}(\mathbb{R}^{d+1})$ (this means it's a linear operator on $\mathbb{R}^{d+1}$) as the projection onto the orthogonal complement $x^\perp$ (the linear subspace of $\mathbb{R}^{d+1}$ which is orthogonal to x). As an explicit formula, P_x is a linear operator, so it's a matrix. And the vector it sends a given y to is going to be

$$P_x y = y - \frac{x \cdot y}{|x|^2} x.$$

We'll say in a moment why we're introducing this thing, but first let's verify that it actually is a projection.

First of all, its image definitely lies in the orthogonal complement of x, because

$$x \cdot P_x y = x \cdot y - x \cdot \frac{x \cdot y}{|x|^2} x = x \cdot y - x \cdot y = 0.$$

And one can verify that $P_x^2 = P_x$ (this one we didn't prepare the calculations, so it's left to us to check). This is what it means to be an orthogonal projection (maybe you also check that it's symmetric, meaning $P_x^\intercal = P_x$).

So it's this projection map. Another way you can write this as, which is probably the better way, is that as a matrix,

$$P_x = \mathrm{id} - \frac{x x^\intercal}{|x|^2}$$

(because if you apply this matrix to y, you get $\mathrm{id} \cdot y$, and you also get $x x^\intercal y = x(x \cdot y)$). If you write it this way, it's easier to verify that it's symmetric; and why $P_x^2 = P_x$ should just be some calculation where you expand out the square and get P_x back.

What's this good for? If you think of $\mathbb{S}^d$ as a submanifold of $\mathbb{R}^{d+1}$, and say you take a point x in $\mathbb{S}^d$, then P_x is precisely the projection onto the tangent space of your manifold. So for all $x \in \mathbb{S}^d$ (all x which are unit vectors), $P_x$mis a projection from $\mathbb{R}^{d+1}$ to the tangent space. Very concretely, the tangent space is the set of vectors which are orthogonal to x, so it precisely is the orthogonal complement of x. So in this explicit example, you have this projection to the tangent space.

And using this projection, you can construct curves which lie in your manifold. How do I construct curves which lie in $\mathbb{S}^d$? Let $y(t)$ be a smooth curve in $\mathbb{R}^{d+1}$, and maybe assume $y'(t)$ is never 0; so we have some very nice smooth curve. Then we can define $x(t)$, which we claim is going to always lie in $\mathbb{S}^d$, as follows. First, we take $x(0) = x_0 \in \mathbb{S}^d$ to be some fixed point in the unit sphere. (This is usual calculus so far; we're not talking about stochastic calculus yet.) Then you define this curve as an ODE, where

$$\dot{x}(t) = P_{x(t)}\dot{y}(t).$$

So I start with a curve in $\mathbb{R}^{d+1}$, and all I care about are its increments; and I want to use the icnrements of my y curve to drive the x curve. If you have the y curve, it's going to have some increment that's not orthogonal to x. But if you want your x to always lie in the sphere, you have to project this increment onto the tangent space of x.

Claim 16.17 — Assuming x exists, we have $x(t) \in \mathbb{S}^d$ for all t.

(This is an equation, so you have to say why there exist solutions.) Why? What it means to lie in the unit sphere is that $|x|$ is always 1. So we want to compute the evolution of the squared norm and show that

$$\frac{d}{dt}|x(t)|^2 = 0$$

(because then I always stay norm 1). But what is this time derivative? Again x is smooth, so I can write it as

$$\frac{d}{dt}(x^j(t)x^j(t)) = 2x^j(t) \cdot \dot{x}^j(t) = 2x(t) \cdot \dot{x}(t).$$

And the claim is that this is supposed to be 0. And I'm specifying the time derivative $\dot{x}(t)$ — this is

$$2x(t) \cdot P_{x(t)}\dot{y}(t)$$

(the inner product of x and something that has been projected to the orthogonal complement of x). And by definition this is 0 (since the second thing lies in the orthogonal complement of x).

This is just a reminder of why you want the chain rule to hold in usual calculus — if you want to define a curve in an embedded manifold in Euclidean space, you have to specify its increments. Its increments are supposed to lie in the tangent space; and when you have an embedded manifold, you just want to project onto that tangent. And what the chain rule tells you is that when you project a general increment to your tangent space, you're going to stay in the manifold.

Remark 16.18. This generalizes to more general settings (where you have general Riemannian manifolds).

Now let's talk about stochastic processes on this manifold. Say B is a Brownian motion in $\mathbb{R}^{d+1}$. Consider a CSMG X satisfying the SDE $X_0 \in \mathbb{S}^d$ and $dX_t = P_{X_t} \circ dB_t$. What does this even mean? This means X is supposed to solve the integral equation

$$X_t = X_0 + \int_0 6tP_{X_s} \circ dB_s.$$

Note that P_{X_s} is a matrix-valued process; so what does it mean to have your integrand be a matrix? But you can kind of think of dB as a vector; and if we think of a matrix times a vector, this thing should itself

be a vector. So it should have $d+1$ compoments, and I define the ith component (which should be a real number) as

$$(P_{X_s})_{ij} \circ dB_s^j.$$

Because P is a matrix-valued process, any matrix entry is going to be a real-valued process. So this is a CSMG if X is; so we're back in the scalar world. That's how you would interpret a stochastic integral where you have a matrix-valued integrand against a vector-valued integrator — you just go back to the scalar case. This is supposed to be the analog of the ODE $\dot{x}(t) = P_{x(t)}\dot{y}(t)$, where $\dot{y}$ is replaced by dB.

You do have to worry about what happens if X reaches 0 (because if it does, then P_X is ill-defined); but for the purposes of this example, let's assume it never does. So we'll assume that a process X satisfying this integral equation exists, and that $X_t \neq 0$ for all t. (In practice you can actually prove this, but we won't talk about that at least now.) The point we want to make is that assuming this process exists, it actually has to lie in the sphere.

Claim 16.19 — We have $X_t \in \mathbb{S}^d$ for all t (almost surely).

Proof. As before, we want to evaluate the evolution $d(|X_t|^2)$; we want to show this evolution is 0, so that $|X_t|$ is constant (and by definition it started at 1).

Because you specified this SDE as a Stratonovich integral, you have the chain rule; so you have

$$d(|X_t|^2) = 2X_t^i \circ dX_{t,i} = 2X_t^i \circ (P_{X_t} \circ dB_t)_i.$$

Now let's be very expllicit and write this differential as a sum of scalar things, as

$$2X_t^i \circ ((P_{X_t})_{ij} \circ dB_t^j).$$

Here's one thing that we didn't really say yet: The Stratonovich integral has the same associativity property as the Ito integral, so I can write this as

$$2(X_t^i (P_{X_t})_{ij}) \circ dB_t^j.$$

And the point now is that the $X_t^i (P_{X_t})_{ij}$ term is always 0. Why? This is the jth entry of $X^\intercal P_X$ (this is a row vector). But forgetting about the jth entry, remember that P_X is the porjection matrix, so $X^\intercal P_X = X^\intercal - X^\intercal = 0$ (by how we defined P_X).

So this calculation shows you that $d\,|X_t|^2 = 0$, and therefore $|X_t| = 1$ for all t. □

This is an example of why you want the Stratonovich integral. If you think about it, we never actually used the fact that B is a BM; this all works if B is just a general CSMG. If you do take B to be a BM, then X is going to be a BM on the unit sphere. Because in general, how do you construct BM on a manifold? Kind of what you want to say is, what is BM really? It's just that if you're at a point x, you want to take a random direction in your tangent space to move, and this random direction is going to be dictated by some sort of Euclidean BM. You can extend this to get constructions of BM on a general Riemannian manifold if you want.

§17 April 7, 2025

Today we'll begin the new notes on applications of Itô's formula. That'll be 2–3 lectures. We'll see a couple of applications. In the textbook, he talks about exponential local martingales (which we'll talk about today), then martingale representations. We'll skip over the fact that every CLMG is a reparametrization of BM,

but if you're interested you can read about it. We'll talk about BG (which is some inequality) and then Girsanov's theorem (which will take a little longer).

The reparametrization thing is the natural statement that you know BM has a certain QV (it's just t, as a function of t), if you're given a general CLMG and reparametrize time so that at time t the QV is t (in general it's some increasing function, but you can reparametrize to make it t), then that's a BM. We'll skip that, but the way you prove it is using Levy's characterization, which we will prove today.

§17.1 Exponential local martingales

Suppose we're in classical calculus, and you want to solve the ODE

$$\dot{z} = z(t)\dot{x}(t)$$

(assume x is smooth or something). The natural way to solve an ODE of this form is to bring $z(t)$ to the left to get

$$\frac{\dot{z}(t)}{z(t)} = \dot{x}(t).$$

And the left-hand side is the derivative of $\log z(t)$, so you get

$$\frac{d}{dt}\log z(t) = \dot{x}(t).$$

Then you can integrate and get

$$z(t) = z(0)e^{x(t)}.$$

And indeed, if you differentiate this you get that it satisfies the ODE — e^x is its own derivative, so when you differentiate it in t you get precise $e^x \cdot \dot{x}$.

Now suppose we want to solve a SDE like

$$dZ_t = Z_t\, dM_t$$

where M is a CLMG (think BM). You may be tempted to just write

$$Z_t = Z_0 e^{M_t},$$

because that works in usual calculus. But the whole point we've been emphasizing over and over is that this is not true — the usual rules of calculus do not apply in the stochastic world. What we talked about last time at the end with the Stratonovich integral — if you actually wanted to find a solution to $dZ_t = Z_t \circ dM_t$ (using the Stratonovich integral instead of the Ito one), then $Z_t = Z_0 e^{M_t}$ would actually solve the Stratonovich SDE, because Stratonovich integrals do satisfy the chain rule. But often you do want to work with Ito because you get martingales.

Another reason there's no way e^{M_t} will be a solution to the Ito SDE is because the Ito SDE is going to give you a martingale (you just have a CLMG part), and there's no way e^{M_t} is a martingale (its expectation would have to be constant, but e.g. e^{B_t} depends on t).

How do you actually find a solution to a SDE of this form? We maybe don't have a great systematic way, but you kind of want to play around with Ito's formula. In particular, if you apply Ito's formula to e^{M_t}, it doesn't satisfy the Ito SDE. But if you play around with it enough, you'll actually see that

$$Z_t = e^{M_t - \frac{1}{2}\langle M,M\rangle_t}$$

is the solution. At least this is kind of consistent with BM (and maybe this is a way to guess the form of the solution). We talked about the exponential martingale of BM, where you had to put $\frac{1}{2}t$ here; and $t = \langle B, B\rangle_t$. So if you think of M as BM at some random time, you might guess the solution to the SDE is of this form.

This motivates the following definition.

Definition 17.1 (Exponential local MG)**.** Let M be a CLMG. For $\lambda \in \mathbb{C}$, we define the *exponential local martingale* of M, denoted by $\mathcal{E}(\lambda M)$, as the process defined by

$$\mathcal{E}(\lambda M)_t = \exp\left(\lambda M_t - \frac{\lambda^2}{t}\langle M, M\rangle_t\right).$$

We allow complex parameters, so in general this is a complex-valued process. When we talk about complex-valued processes, we still have a notion of a martingale:

Definition 17.2. We say a $\mathbb{C}$-valued process (X_t) is a (local) martingale if both $(\operatorname{Re} X_t)$ and $(\operatorname{Im} X_t)$ are (local) martingales.

(This is basically the definition you would guess.)

It's very important that you just talk about martingales; if you talk about submartingales, you can't come up with as canonical a definition.

In particular, $\mathcal{E}(\lambda M)$ is going to be a local martingale — we're going to show that it satisfies the SDE above. Ito's formula gives you the canonical decomposiiton of a semimartingale; and that equation tells you there's no FV part, only a local martingale part, so it's actually going to be a local martingale.

Remark 17.3. Something we're skipping over: we'll show this process satisfies the SDE $dZ_t = Z_t\, dM_t$. But is it the unique solution? In regular calculus, under reasonable assumptions on x, our z was actually the unique solution (you can go from the ODE to the solution, and this is the only one). But why would that be true in this case? That we'll come to probably later, at the end of the course, if we have time. But for now, it suffices to define this process and verify it's a local martingale, and we'll just use that; but actually it will be the unique solution to that SDE.

Proposition 17.4

The process $\mathcal{E}(\lambda M)$ is a CLMG, and moreover,

$$d\mathcal{E}(\lambda M)_t = \lambda \mathcal{E}(\lambda M)_t\, dM_t.$$

(You have a λ because it's λM.)

Proof. The proof is you just apply Ito's formula. You can kind of cook up a function of two variables — the spatial variable and time variable — and think about applying Ito's formula. But the way you do this in practie kind of directly is that we can think about this as a product of two processes $e^{\lambda M_t}$ and $e^{-\frac{\lambda^2}{2}\langle M,M\rangle_t}$. And basically you can put the d on either of them, so we get

$$d(e^{\lambda M_t}\cdot e^{-\frac{\lambda^2}{2}\langle M,M\rangle_t}) = d(e^{\lambda M_t})e^{-\frac{\lambda}{2}\langle M,M\rangle_t} + e^{\lambda M_t}d(e^{-\lambda\langle M,M\rangle_t}) + d(e^{\lambda M_t})d(e^{-\frac{\lambda^2}{2}\langle M,M\rangle_t}).$$

And the last term is just 0, because $\langle M, M\rangle_t$ is a finite variation process.

For the first term, we can again apply Ito's formula to get

$$\left(\lambda e^{\lambda M_t}\, dM_t + \frac{\lambda^2}{2}e^{\lambda M_t}\, d\langle M, M\rangle_t\right) e^{-\frac{\lambda^2}{2}\langle M,M\rangle_t}.$$

And for the second term, this thing is actually finite variation, so the usual rules of calculus apply, and we get

$$-\lambda e^{\lambda M_t}e^{-\lambda^2\langle M,M\rangle_t}\, d\langle M, M\rangle_t.$$

And now you see the cancellation — that's precisely what the correction term of $-\frac{\lambda^2}{2}\langle M, M\rangle_t$ is giving you (you have to add in this term in order to cancel out the second term that arises when you apply Ito's formula to the first term).

So this is equal to

$$\lambda e^{\lambda M_t} e^{-\frac{\lambda^2}{2}\langle M,M\rangle_t}\, dM_t.$$

And this is precisely $\lambda \mathcal{E}(\lambda M)\, dM_t$. □

It's always kind of slow if you think about what the function is supposed to be of my two variables to apply Ito's formula; it's faster to just compute like this. So if you'll be working with Ito's formula in the future you want to be able to compute quickly. Computations are kind of part of math — if you really understand something deeply, you should be able to do the computations quickly, instead of having to look up Ito's formula all the time. But that just comes with applying Ito's formula a bunch of times and getting used to working with it.

Example 17.5

In the case of Brownian motion, we've seen the exponential martingale before — we have

$$\mathcal{E}(\lambda B_t) = \exp\left(\lambda B_t - \frac{\lambda^2}{2} t\right).$$

This is actually a martingale (since this thing satisfies very nice tail bounds, e.g., it has finite expectation).

§17.2 Levy's characterization of Brownian motion

This is the first big application of Ito's formula.

Proposition 17.6 (Lévy's characterization of BM)

Let $X = (X^1, \dots, X^d)$ be an adapted d-dimensional process with continuous sample paths. Then the following are equivalent:

(i) X is a d-dimensional $(\mathcal{F}_t)$-Brownian motion.

(ii) X is a CLMG with $\langle X^i, X^j\rangle = \delta^{ij} t$ (where $\delta^{ij} = \mathbf{1}[i = j]$).

In particular, in the 1-dimensional case, a CLMG is an $(\mathcal{F}_t)$-Brownian motion if and only if $\langle M, M\rangle_t = t$ for all $t \geq 0$, or equivalently, if $M_t^2 - t$ is a CLMG.

As usual, we're imagining we've fixed a filtration at the very beginning (and 'adapted' is with respect to that fixed filtration). A d-dimensional BM means that these processes X^i are mutually independent, adapted to $\mathcal{F}_t$, have independent increments, and have the same finite-dimensional distributions as BM.

For (ii), X is vector-valued, but a vector-valued local martingale just means each of its coordinates is a local martingale; so extending the notion of local martingales to a vector space is immediate. (Again, it wouldn't be as canonical for submartingales.)

(The last part is because one characterization of the QV is the unique process you have to subtract from M_t^2 to get a local martingale.)

The fact that (i) implies (ii) is basically immediate — the X^is are independent, so you can kind of just do this calculation and get the QV. The main thing is really why (ii) implies (i). (And the last statement is a

direct consequence of the more general one, because in the 1-dimensional case there's just one component, and this is the characterization of the QV.) So for the proof, we'll mainly just focus on why (ii) implies (i).

Proof. We'll show that (ii) implies (i). Fix a vector $\xi \in \mathbb{R}^d$. Then — this is kind of related to how you prove the (multivariate) central limit theorem. You have a vector $X = (X^1, \dots, X^d)$, and if it's multivariate normal, that means if you dot-product this with any d-dimensional vector, you get a multivariate Gaussian. So here we're going to again dot-product with fixed vectors ξ; we write

$$(\xi \cdot X)_t = \xi_i X_t^i$$

(there's an implicit summation over the index i, so this is a dot product). This is a CLMG (since the space of CLMGs is a vector space), and you can compute that its QV is

$$\langle (\xi \cdot X), (\xi \cdot X) \rangle_t = \xi_i \xi_j \langle X^i, X^j \rangle_t$$

(you can pull out the ξ's using bilinearity). And using (ii), this becomes

$$\xi_i \xi_j \delta^{ij} t.$$

Here you're summing over all indices, but δ^{ij} essentially enforces a restriction that $i = j$, so this is actually just going to be $|\xi|^2 t$ (where $|\xi|$ is the norm of ξ).

Now we use the proposition from before (on exponential martingales). Take $\lambda = i$ (this means the square root of -1). Then by that proposition, the process

$$\exp\left(i(\xi \cdot X)_t + \frac{1}{2} |\xi|^2 t \right)$$

is a CLMG. (We're taking $\lambda = i$, so with the $-\frac{\lambda^2}{2}$ we get $+\frac{1}{2}$.)

And actually, this is going to be a martingale — the first term is bounded by 1 (it's just a complex phase). And the second term is bounded on any $[0, t_0]$, and you can apply the fact that if a local martingale is bounded, it's an actual martingale (at least on any finite interval, and then you can take an arbitrary interval). So again we're just saying if you have a CLMG which is also bounded, that's actually a martingale — we proved this a while ago. And this thing actually is bounded, at least on a finite interval.

Now we get an actual martingale; what does that mean? If we look at the martingale property, it tells us that

$$\mathbb{E}\left[\exp\left(i(\xi \cdot X)_t + \frac{1}{2} |\xi|^2 t \right) \mid \mathcal{F}_s \right] = \exp\left(i(\xi \cdot X)_s + \frac{1}{2} |\xi|^2 s \right)$$

(the conditional expectation at time t given $\mathcal{F}_s$ is the thing at time s). And if we rearrange this, we actually get a Gaussian characteristic function — moving things around (the second term in the conditional expectation is constant, so you can move it over), you get

$$\mathbb{E}\left[\exp(i\xi \cdot (X_t - X_s)) \mid \mathcal{F}_s \right] = \exp\left(-\frac{1}{2} |\xi|^2 (t - s) \right).$$

In particular, if you take expectations, you immediately get that the increment $X_t - X_s$ is a d-dimensional Gaussian distributed as

$$X_t - X_s \sim \mathcal{N}_d(0, (t - s) \cdot \mathrm{id}).$$

This is already halfway there (to showing that X is going to be a BM); at the very least, the increments have the correct distributions.

But now you have to show the increments are independent of each other — once you know these two things, you'll have that all finite-dimensional distributions of X are the same as BM, so X will be a BM.

To get this independence, we have to use the aboev identity

$$\mathbb{E}\left[\exp(i\xi \cdot (X_t - X_s)) \mid \mathcal{F}_s\right] = \exp\left(-\frac{1}{2}|\xi|^2 (t-s)\right)$$

before taking expectations. Right now, our goal is to show that $X_t - X_s$ is independent of $\mathcal{F}_s$. This is almost immediate from the above statement depending on how much measure theory you're willing to assume — this statement essentially tells me the conditional law of X_t given $\mathcal{F}_s$ is this thing, which doesn't depend on $\mathcal{F}_s$. And that should give me that the law of X_t doesn't depend on $\mathcal{F}_s$, which should give the independence. In papers you could probably just say this, but we'll make it more formal.

Fix some $A \in \mathcal{F}_s$. Then we'll multiply both sides by $\mathbf{1}_A$ and take expectations; so

$$\mathbb{E}\left[\mathbf{1}_A \exp(i\xi \cdot (X_t - X_s))\right] = \mathbb{P}[A] \exp\left(-\frac{|\xi|^2}{2}(t-s)\right)$$

(the fact that $A \in \mathcal{F}_s$ means we can take A on the inside when taking expectations; and the right-hand side is just constant, and $\mathbb{E}[c\mathbf{1}_A] = \mathbb{P}[A] \cdot c$).

Now, at least if you assume $\mathbb{P}[A] > 0$, then you can define a 'conditioned probability measure'

$$\mathbb{P}_A[B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[A]}$$

(this is basically the conditional probability of B given that A occurs). Basically what the above identity is telling you is — you could have moved $\mathbb{P}[A]$ to the left-hand side, so what this tells you is that under $\mathbb{P}_A$, the law of $X_t - X_s$ is still the same Gaussian $\mathcal{N}_d(0, (t-s)\,\mathrm{id})$ (because if you move $\mathbb{P}[A]$ to the left, you get the characteristic function of $X_t - X_s$, but with the probability measure of $\mathbb{P}_A$; and we're saying the characteristic function is $\exp(\cdots)$, which is what comes from this Gaussian distribution).

And that tells us that if we now integrate any bounded measurable function f under this new probability measure, we get

$$\frac{\mathbb{E}[\mathbf{1}_A f(X_t - X_s)]}{\mathbb{P}[A]},$$

and using the fact that we have the law of $X_t - X_s$, this is just

$$\mathbb{E}[f(X_t - X_s)]$$

(here this is again because we showed that under our original probability measure, $X_t - X_s$ is just distriuted like a Gaussian).

Now if you move $\mathbb{P}[A]$ to the right-hand side, you basically get the desired independence. Basically we've shown that for any $A \in \mathcal{F}_s$ and any bounded measurable function f, we have

$$\mathbb{E}[\mathbf{1}_A f(X_t - X_s)] = \mathbb{P}[A]\mathbb{E}[f(X_t - X_s)].$$

And once you arrive here, this basically is just independence. So this implies $X_t - X_s$ is independent of $\mathcal{F}_s$ (in particular, if you take f to be an indicator function, this is basically the definition of independence).

Then if you iterate this, you obtain that any sequence of increments X_{t_1}, $X_{t_2} - X_{t_1}$, $X_{t_3} - X_{t_2}$, ... are mutually independent, and that basically finishes the proof. □

Remark 17.7. Sky mentioned a while ago (when we first introduced local martingales) the fact that you should think of the QV as telling you how much time you ran the BM for. That's basically restating this reparametrization of BM result that we're skipping. But it also basically comes out of this calculation here. If you actually knew the QV of your local martingale at time t was (deterministically) t, this argument tells you that it's a Gaussian with variance t.

§17.3 BDG inequalities

Now we'll talk about another application of Ito's formula. Before we get to that, let's recall that if you have $Z \sim \mathcal{N}(0,t)$, its $2p$th moments — there will be an explicit formula

$$\mathbb{E}[Z^{2p}] = (2p-1)(2p-3)\cdots 1 \cdot t^p$$

(and odd moments are just 0). And actually, if you track the asymptotics carefully, this should be

$$\mathbb{E}[Z^{2p}] \sim p^p \mathbb{E}[Z^2]^p.$$

It turns out that you have this type of asymptotics, in that higher moments of the Gaussian are basically controlled by the variance, times some powers of p.

This actually implies some very strong concentration bounds. In general, if a random variable X satisfes this type of moment bound

$$\mathbb{E}[X^{2p}]^{1/2p} \lesssim p^{1/2}\mathbb{E}[X^2]^{1/2}$$

(this is what happens if we take $1/2p$th powers of both sides; and $\lesssim$ denotes 'up to some constant independent of p'), this is a very strong statement on the tails — it actually implies that X is *sub-Gaussian* (we're not going to prove this here), which in particular means the tails of X decay like a Gaussian distribution — specifically,

$$\mathbb{P}[|X| > x] \lesssim \exp(-cx^2),$$

where c depends on the implicit constant in the earlier $\lesssim$. So this is another characterization of sub-Gaussians — that the L^{2p}-norm is bounded by the L^2-norm.

More generally, if you take the exponent in $p^{1/2}$ to not be $1/2$, but something larger, then you'll get 'stretched exponential tails.' If

$$\|X\|_{L^{2p}(\Omega)} \lesssim p^{2k} \|X\|_{L^2(\Omega)}$$

(the left-hand side is another way of writing $\mathbb{E}[X^{2p}]^{1/2p}$), then X hsas *stretched-exponential tails*. In other words, its tail behavior $\mathbb{P}[|X| > x]$ (think of x as being very large) should decay like $\exp(-cx^\alpha)$, where α will be related to k (if $k = 1/2$ then $\alpha = 2$; if $k > 1/2$ then $\alpha < 2$, because then this is a weaker inequality, which should mean you get less tail decay).

This type of concentration is very important in various situations — basically anything that satisfies this tpe of exponential tail, you can treat as $O(1)$ (it's never going to be too big) — as soon as x is big, this thing is going to dominate.

We're thinking of local martingales as some time-change of BM (though we're not going to prove this). And BM is distributed like a standard Gaussian — so BM does satisfy this type of bound

$$\|B\|_{L^{2p}(\Omega)} \lesssim p^{1/2} \|B_t\|_{L^2(\Omega)}.$$

And if BM satisfies this bound and local martingales are reparametrizations of BM, you might think that LMs should also satisfy bounds like this. That's what the following inequalities are.

Proposition 17.8 (Burkholder–Davis–Grundy inequalities (BDG))

Let M be a CLMG with $M_0 = 0$. Then for all $p \geq 1$, there is some constant C_p such that for every stopping time T, we have

$$\mathbb{E}\left[\sup_{0\leq t\leq T} |M_t|^{2p}\right] \leq C_p \mathbb{E}\left[\langle M, M\rangle_T^p\right].$$

We didn't track C_p carefuly — it should grow with p — but you should be able to.

Here the powers have to work out — we have $2p$ powers of M on the left-hand side, and on the right-hand side $\langle M, M\rangle$ is already quadratic. In particular, this equality is invariant up to scaling.

Le Gall proves a more general statememt, but for instance here we always have $2p \geq 2$; he also proves this statement between 1 and 2 and for odd powers. But in practice, you don't rea.ly care about when the power os between 1 and 2 — you realy just care about the asymptotic behavior. But maybe there are situations where the expectation is only finite if $p \in (1, 2)$, and then maybe you care; but for that you can just read Le Gall.

This type of proof is very reminiscent to some type of energy inequality if you ever took PDEs before. Basically what it's going to boil down to is if you have some inequality like

$$x^{2p} \leq x^{2(p-1)} y$$

(where these are just real numbers, and $x, y > 0$), this is obviously going to imply $x^2 \leq y$.

The philosophy is on the LHS you have $2p$ powers. You really worry if X is super large (if $X \leq 1$, you can just bound this by 1). So you're worried about the case X is very large. The philosophy is on the LHS you have $2p$ powers of something that may be large. You try to bound it by something that has less powers, and that lets you absorb stuff.

So how do we do that in this scenario? First there's a bunch of simplifications you can make. We can assume $T = t_0$ is deterministic (in general, you can apply the inequality to $M_{t_0}^T$ and then send $t_0 \to \infty$). So basically, if you had this inequality where T was always just some fixed time t_0, then you could apply this inequality to the stopped martingale and send $t_0 \to \infty$. (The limits are fine — these are all monotone under increasing T, so you can always apply the MCT.)

And also, by localizing M, we can assume that M is bounded. It's a continuous local martingale, so you can always localize by stopping at the first time it goes above n (that's a sequence of stopping times we've sxeen before). If you have this inequality for bounded martingales and want to extend to CLMGs, you just replace M by a stopped version and send $n \to \infty$. The thing is again going to be monotone in n.

So all this is to say that you can assume M i bounded and the stopping tiem is deterministic.

> **Remark 17.9.** This is analogous to how when you prove things in analysis, you prove them first for smooth functions and then argue by density — so you can take as many derivatives as you want. Here we do this so that you can take as many moments as you want.
>
> Note that in general, the right-hand side might be ∞ (these are all positive quantities so their expectations are still well-defined).

And now we can apply Ito's formula — you can think of Ito as giving you the evolution of the process. So if we forget about the sup, we want to understand how the $2p$th moment of M evolves in time, because that's what we need to understand. This is where it's simpler to take an even power — the function $x \mapsto x^{2p}$ is smooth, but if you had an odd power (e.g., $x \mapsto |x|$), it's not smooth at the origin. You can still apply Ito's but you have to be a bit more careful.

But at least for this function, it's just a smooth function, so there's no problem in applying Ito's, and you'll get

$$d(M_t^{2p}) = 2pM_t^{2p-1}\, dM_t + \frac{1}{2} 2p(2p-1) M_t^{2(p-1)}\, d\langle M, M\rangle_t$$

(the martingale term plus the Ito correction term). This is one illustration of why it's so nice to have martingales — why you want to actually decompose into martingales and FV terms.

The fact that M is bounded means the first term will actually be a martingale, not just a local martingale. This means if you take expectations of both sides (first you integrate both sides, since this thing is really

shorthand notation for an integral equation), the first thing is a martingale starting at 0, so its expectation is 0 and we don't have to worry about it. So the only thing we care about is the expectation of the second term, and we can write

$$\mathbb{E}[M_t^{2p}] = p(2p-1)\mathbb{E}\left[\int_0^t M_s^{2(p-1)}\, d\langle M, M\rangle_s\right].$$

Remark 17.10. Often the way you want to understand the evolution of some expectation is you apply Ito's formula, and then you can ignore the martingale part (assuming it's a martingale) and just try to understand the evolution of the FV part.

Now that you've arrived here, one thing you can do is bound the integrand by the sup over the interval; if you do that, you get

$$\mathbb{E}[M_t^{2p}] \le p(2p-1)\mathbb{E}\left[\sup_{0\le s\le t} M_s^{2(p-1)}\langle M, M\rangle_t\right]$$

(we're writing t instead of t_0) — you just bound the integrand by the max it could be, and now when we integrate we just get the QV term.

But now you look at this and it seems kind of bad — if I want to chain things like $x^{2p} \le x^{2(p-1)y}$, you want the two things to look the same. But the RHS has a sup and the LHS doesn't, so this looks kind of hopeless.

But then you remember you have the L^p maximal inequality, so the expectation of the $2p$th power of the endpoint does control the expectation of the sup. You really have to use that in a crucial way, or else this is hopeless.

So by the L^{2p} maximal inequality, we get that there is some constant C_p' such that

$$\mathbb{E}\left[\sup_{0\le s\le t} M_s^{2p}\right] \le C_p'\mathbb{E}[M_t^{2p}].$$

And now you plug in the inequality we had, so we get

$$\mathbb{E}\left[\sup_{0\le s\le t} M_s^{2p}\right] \le C_p'\mathbb{E}[M_t^{2p}] \le p(2p-1)\mathbb{E}\left[\sup_{0\le s\le t} M_s^{2(p-1)}\langle M, M\rangle_t\right].$$

And now this is more hopeful. The philosophy on how you would come up with this argument is first you want to compute the evolution of the $2p$th power, and Ito's formula gives you that. But if you want to come back to this kind of $x^{2p} \le x^{2(p-1)}y$ type inequality, you need to control the RHS by the LHS, and that's where you use the L^{2p} maximal inequality.

We're not done yet because we have an expectation of a product, and we have to separate it. But for that, we can use Hölder:

Fact 17.11 (Hölder) — We have $\mathbb{E}[XY] \le \mathbb{E}[X^q]^{1/q}\mathbb{E}[X^r]^{1/r}$ where $1/q + 1/r = 1$.

In the notes, we finish slightly differently; the way we're presenting it here is maybe slightly more direct. Basically, the thing we actually control on th eleft-hand side is just $\mathbb{E}[\sup]$, but here we have an expectation of a product, so we want to separate these. Here's where you use Hölder, thinking of the sup as X and the $\langle M, M\rangle_t$ term as Y. We want to choose these parameters such that $2(p-1)q = 2p$ — because if I choose q this way, then the first factor here will be precisely something that's on the left-hand side. And that means I have to solve for r now — this is going to tell me that $q = \frac{p}{p-1}$, so $r = p$. Then you apply Hölder with this choice (which is informed by that you want to get whatever's on the LHS into the RHS). Then by Hölder you get

$$\mathbb{E}\left[\sup_{0\le s\le t} M_s^{2p}\right] \le C_p' p(2p-1)\mathbb{E}\left[\sup_{0\le s\le t} M_s^{2p}\right]^{(p-1)/p} \cdot \mathbb{E}[\langle M, M\rangle_t^p]^{1/p}.$$

And now we're basically done. Actually the right inequality to use (with x and y) would have been $x \leq x^{1-1/p}y$, and that would give $x^{1/p} \leq y$.

And what's crucial is that we have a strictly less than 1 power of the LHS on the RHS — we have $\frac{p-1}{p} < 1$ (if it were 1, this would be hopeless). And this gives me an estimate on the sup I'm trying to estimate — we get

$$\mathbb{E}\left[\sup_{0\leq s\leq t} M_s^{2p}\right]^{1/p} \leq C_p' p(2p-1)\mathbb{E}\left[\langle M, M\rangle_t^p\right]^{1/p}.$$

If we raise both sides to the pth power, we get what we want.

Remark 17.12. We probably have

$$C_p' = (\frac{2p}{2p-1})^{2p}$$

or something like this — whatever comes from the L^{2p} maximal inequality.

If this is the right constant, the thing we were saying before is that if the thing grows by a constant p^k where k is independent of p, then you get subexponential tails. This thing actually might work out — this C_p' should be bounded by a constant independent of p, which means you get p^2 dependence, which gives you subexponential tails. (But there could be a mistake here.)

Remark 17.13. If you've taken a PDE course, this type of energy estimate may be familiar — you try to estimate some quantity by a strictly less than 1st power of that same quantity, times some other quantity (and then you get a bound on that quantity). In this case, the way you do that is by computing the evolution, and trying to understand the RHS. The good thing about the RHS is it involves strictly less powers of the original quantity (only $2(p-1)$ instead of $2p$). So you can hope to estimate the thing by a strictly-less-than-1 power of the thing. But in order to get that, you have to use the L^p maximal inequality and Hölder. And the choice of the parameters is always involved because you want to get the thing on the LHS onto the RHS, so you just have to make this $2(p-1)$ into a $2p$ and then you're done.

So that's the BDG inequality. In applications, it is quite useful — not that we're really going to give any applications of it, but at least with these subexponential tails, any time you want a concentration inequality for a local martingale (or a martingale) you would try to apply BDG to bound its moments, because an estimate on its moments is going to lead to some type of concentration.

Other martingale concentration inequalities you see in discrete-time are like Azuma–Hoeffding or the bounded differences inequality; BDG is kind of a *continuous*-time martingale concentration inequality.

§17.4 The martingale representation theorem

This is another separate application of Ito's formula.

This martingale representation theorem basically gives you a way to represent random variables as terminal values of a martingale. It's in the special setting where you're working with a BM and its completed filtration.

Theorem 17.14 (Martingale representation theorem)

Suppose $(\mathcal{F}_t)$ is the completed canonical filtration of a Brownian motion B (started from 0). Then for all random variables $Z \in L^2(\Omega, \mathcal{F}_\infty, \mathbb{P})$, there exists a unique progressive process $h \in L^2(B)$ such that

$$Z = \mathbb{E}[Z] + \int_0^\infty h_s\, dB_s.$$

(If we don't say otherwise, B starts from 0 and is $\mathbb{R}$-valued. This same theorem holds for vector-valued BMs, but we'll just state it for real-valued ones, at least for now.)

Your random variables Z are in principle functions of your *entire* Brownian path — you can observe the entire path and make some function out of that (and we're assuming it's in L^2).

Recall that for a progressive process to be in $L^2(B)$, in the special case of Brownian motion, this means $\mathbb{E}[\int_0^\infty h_s^2\,ds] < \infty$ (more generally, instead of ds you'd take the QV of your martingale, but for BM that's just s).

So in other words, for any Z which is a function of your Brownian path (in L^2), you can write it as a stochastic integral along the path — you can basically recover the function by following this martingale to ∞. That's the first part of the theorem. Here's the second part.

Corollary 17.15

Consequently, for every martingale M which is bounded in L^2 (respectively, for every CLMG M), there exists a unique process $h \in L^2(B)$ (respectively, $h \in L^2_{\text{loc}}(B)$) and a constant C such that

$$M_t = C + \int_0^t h_s\,dB_s.$$

Recall that L^2_{loc} means that if you integrate on a finite interval, the thing is finite.

Remark 17.16. One remark to emphasize here is that in the L^2-bounded case, M is not assumed to be continuous. In other words, what the second part of this theorem is saying is that any L^2-bounded martingale with respect to this completed filtration of BM actually has a continuous modification. As usual, this is up to almost sure equality; and the continuous modification is the RHS. (Recall that when $h \in L^2(B)$, we get this continuous L^2-bounded martingale $\int_0^t h_s\,dB_s$.) So actually every martingale with respect to this filtration has a continuous modification. This is quite strong — why should that be true?

So that's the statement of the theorem. We won't prove it today, but Sky will start by stating this technical lemma that we need.

Lemma 17.17

Under the assumptions of the theorem, the vector space generated by random variables of the form

$$\exp\left(i\sum_{j=1}^n \lambda_j(B_{t_j} - B_{t_{j-1}})\right)$$

is dense in $L^2(\Omega, \mathcal{F}_\infty, \mathbb{P})$ (where $\lambda_j \in \mathbb{R}$, and $t_0 < t_1 < \cdots$).

(Here we're using $\mathbb{C}$-valued L^2 — these are in general complex-valued functions.)

The way the proof of the martingale representation theorem will go is that we'll verify explicitly that the statement holds for such random variables (think of this random variable as a possible Z — it's bounded so it's definitely in L^2, and it's definitely in $\mathcal{F}_\infty$ because it's a function of your Brownian path up to t_n). So you ca write down explicitly some h such that when you integrate against BM you get this process. (This almost looks like the exponential martingale of Brownian motion, up to some scalar factor; and you allow scalar factors, because it's a generated vector space.)

And then if you just had that these RVs are dense, and also that this statement is 'closed' (in that you

can take limits and it still holds), then you'll be able to prove the theorem. So that's what we're trending towards, and we'll finish that next time.

§18 April 9, 2025

§18.1 The martingale representation theorem

We'll start today by proving the martingale representation theorem. The setting is you take the completed canonical filtration of a BM; and it turns out you can represent *any* random variable which is a function of your entire Brownian path as some sort of integral along the path.

Theorem 18.1 (Martingale representation theorem)

Suppose that $(\mathcal{F}_t)$ is the completed canonical filtration of a Brownian motion B. Then for all $Z \in L^2(\Omega, \mathcal{F}_\infty, \mathbb{P})$, there exists a unique progressive process $h \in L^2(B)$ such that

$$Z = \mathbb{E}[Z] + \int_0^\infty h \, dB.$$

In the proof of the theorem, we'll actually show that this will imply that for any martingale with respect to this filtration, you can write it as some stochastic integral.

Corollary 18.2

For every martingale M bounded in L^2 (respectively, every CLMG M), there exists a unique progressive process $h \in L^2(B)$ (respectively, $h \in L^2_{\text{loc}}(B)$) and a constant $c \in \mathbb{R}$ such that

$$M_t = c + \int_0^t h_s \, dB_s \quad \text{for all } t.$$

This is quite a strong statement — there's no reason why it should be true. But it basically says that in the case of the completed canonical filtration of a BM, these stochastic integrals are the only L^2-bounded martingales we have.

As mentioned last time, the way we'll prove this is by some density argument, which is why we have the following technical lemma.

Lemma 18.3

Under the assumptions of Theorem 18.1, the vector space generated by

$$\exp\left(i \sum_{j=1}^n \lambda_j (B_{t_j} - B_{t_{j-1}})\right)$$

is dense in $L^2_{\mathbb{C}}(\Omega, \mathcal{F}_\infty, \mathbb{P})$.

Here the λ_j's are real, and the t_j's are some finite sequence of times. So we consider any random variable of this form; and if you take the vector space generated by such things, you get a dense subspace of this space $L^2_{\mathbb{C}}(\Omega, \mathcal{F}_\infty, \mathbb{P})$. Here it's natural that the filtration has to be $\mathcal{F}_\infty$ and not $\mathcal{F}$ (the abstract filtration associated with your probability space), because these random variables only look at your Brownian path. So $\mathcal{F}_\infty$ is the best we can hope for, but that's all you need.

Proof. The main step of the proof we'll actually kind of be assuming (you can Google the needed prerequisites if you want).

It suffices to show that if $Z \in L^2_{\mathbb{C}}(\Omega, \mathbb{P}, \mathcal{F}_\infty)$ is orthogonal to every random variable of this form — meaning that

$$\mathbb{E}\left[Z \cdot \exp\left(i \sum_{j=1}^{n} \lambda_j (B_{t_j} - B_{t_{j-1}})\right)\right] = 0$$

for all λ_j's and t_j's — then $Z = 0$ almost surely. This is just some abstract Hilbert space theory; the point is that if this is true, you can't have a kind of picture where this vector space spanned by all these random variables is some lower-dimensional subspace and you could take something in the orthogonal complement. This picture can't occur if you could prove a property like this — then you could take Z in the orthogonal complement. That's the picture in finite dimensions, at least. We used something like this previously (and probably drew a similar picture).

So let's prove this. Why is this going to be true? Let's fix a set of times $\{t_j\}_{j=1}^n$, and then consider the complex measure (what is a complex measure? it's a measure with a real part and imaginary part which are real-valued measures) μ on $\mathbb{R}^n$ (where n is the number of time points we have), defined by

$$\mu(F) = \mathbb{E}\left[Z\mathbf{1}(B_{t_1}, B_{t_2} - B_{t_1}, \ldots, B_{t_n} - B_{t_{n-1}}) \in F\right].$$

So given a Borel set $F \subseteq \mathbb{R}^n$, we're taking the increments of our BM and seeing whether they're in F.

Then if you integrate functions against this measure, you'll get

$$\int \mu(dx) f(x) = \mathbb{E}\left[Z f(B_{t_1}, \ldots, B_{t_n} - B_{t_{n-1}})\right]$$

(you can verify by standard arguments that it's true on indicator functions and you can extend to bounded measurable functions). In particular, μ is a finite measure — if you take $f = 1$ you're computing the total mass, which is going to be $\mathbb{E}[Z]$, and your assumption is that $Z \in L^2$, so its expectation is finite. So there's no infinities to worry about.

As a consequence of this, you can compute the Fourier transform of μ. If we have $\xi \in \mathbb{R}^n$, by definition the Fourier transform of a measure is defined by

$$\widehat{\mu}(\xi) = \int \mu(dx) e^{ix \cdot \xi}.$$

And now if you use this integration formula, our f is of a very particular form, so we get

$$\widehat{\mu}(\xi) = \mathbb{E}\left[Z \cdot \exp\left(i \sum \xi_j (B_{t_j} - B_{t_{j-1}})\right)\right]$$

(we just used this integration formula for our particular choice of f). But anything like this is just 0 by our assumption. So we actually have $\widehat{\mu} = 0$.

And here's where you can Google the result — this is going to have to imply that the measure itself is 0, i.e., $\mu = 0$. Basically, the Fourier transform has this property — you can define it on the space of finite complex measures, and if the Fourier transform of a measure is 0, so is the actual measure. (You probably saw this in 18.675 for probability measures — we saw characteristic functions, which are the Fourier transform of a probability measure, and you can do Fourier inversion to recover the measure from the Fourier transform. Something similar works here.)

And this implies $\mathbb{E}[Z\mathbf{1}_A] = 0$ for all $A \in \sigma(B_{t_1}, \ldots, B_{t_n})$ (because for any event of this form, you can write it as the indicator that these increments of B are in some Borel set).

But the set of times was arbitrary; so by more measure theory stuff, this implies that $\mathbb{E}[Z\mathbf{1}_A] = 0$ for all $A \in \sigma(B_t \mid t \geq 0)$ (if it holds for any finite collection of times, then it holds for the generated σ-algebra on all times; this is again just some abstract measure theory).

And by assumption $\mathcal{F}_\infty$ is the completion of this σ-algebra $\sigma(B_t \mid t \geq 0)$. So any event in $\mathcal{F}_\infty$ can be obtained from an event in this σ-algebra by adding or intersecting with measure-0 or 1 events. Anything about expectations is unchanged if you modify A on a measure-0 event, so it's going to be also true for all $A \in \mathcal{F}_\infty$. (There was some exercise earlier in the semester giving an explicit formula for what the completion of a σ-algebra looks like, and one can verify that if $\mathbb{E}[Z\mathbf{1}_A] = 0$ is true for all events in the original, it's also true for all in the completion.)

But since Z is $\mathcal{F}_\infty$-measurable, this implies $Z = 0$ almost surely — for example, you could take $A = \{Z > 0\}$; this is in $\mathcal{F}_\infty$ by assumption, and that tells you $Z^+ = 0$ (and similarly $Z^- = 0$). □

Student Question. *Why is the Fourier transform* 0*?*

Answer. The Fourier transform is a function $\mathbb{R}^n \to \mathbb{C}$, and the way we defined our measure is such that we have this integration formula for $\int \mu(dx) f(x)$. And by definition the Fourier transform is the integral of $e^{ix\cdot\xi}$; and using this integral formula, we get this $\mathbb{E}[Z\exp(\cdots)]$ formula.

And we assumed that Z was orthogonal to every random variable of the form $\exp(\cdots)$, which means this expectation is 0.

Now let's prove the martingale representation theorem.

Proof of Theorem 18.1. As usual, uniqueness is the easier part: if you had two processes h and $\widetilde{h}$, that would tell you

$$\int_0^\infty h\, dB = \int_0^\infty \widetilde{h}\, dB,$$

which by linearity would tell you

$$\int_0^\infty (h - \widetilde{h})\, dB = 0.$$

We want to conclude $h - \widetilde{h} = 0$. But by the Ito isometry, if we take the second moment of this, we get

$$\mathbb{E}\left[\left(\int_0^\infty (h - \widetilde{h})^2\, ds\right)^2\right] = 0.$$

(In general, you're integrating against the QV of your BM, but here that's just ds.)

This tells us that $h = \widetilde{h}$ in $L^2(B)$ (since we're just saying that the norm of their difference is 0).

So that's uniqueness. Now let's talk about existence. For this, let $\mathcal{H}$ be the vector space of all random varaibles Z for which the theorem statement is true. (You first define $\mathcal{H}$ as the *set* of random variables for which it's true, and you have to verify it's a vector space. But that's just by the linearity of the stochastic integral — if you have Z_1 and Z_2, then they have associated h_1 and h_2, and the h that'll work for $Z_1 + Z_2$ is just $h_1 + h_2$. So it's actually a vector space.) We're going to show that:

(1) $\mathcal{H}$ is a closed subspace.

(2) $\mathcal{H}$ contains a dense subspace, which will be precisely the one that the lemma is about.

So this is the roadmap. Once you show (1) and (2), then we're done — that'll imply $\mathcal{H}$ contains all of L^2.

First, how do you show $\mathcal{H}$ is a closed subspace? Well, what does that mean? It means if I have a sequence $\{Z_n\} \subseteq \mathcal{H}$ which converges in L^2 to some random variable Z, then we want to show that actually $Z \in \mathcal{H}$. (That's what it means to be closed.)

Well, by assumption — by the definition of $\mathcal{H}$, you have

$$Z_n = \mathbb{E}[Z_n] + \int_0^\infty h^n\, dB$$

(that's how we defined $\mathcal{H}$), where h^n is some progressive process. And you also have that $\mathbb{E}[Z_n] \to \mathbb{E}[Z]$ (from the L^2 convergence). Eventually we want to write Z as $\mathbb{E}[Z]$ plus a stochastic integral, so at least the $\mathbb{E}[Z_n]$'s converge to the right thing. Now we want h whose stochastic integral is the limit of those of the h^ns; how do we do that?

First, using the L^2 convergence of the Z_n's and the convergence of the expectations, you get that

$$\int_0^\infty h^n \, dB$$

converges in $L^2(\Omega, \mathcal{F}_\infty, \mathbb{P})$ as a sequence of random variables in n. Here's again where you use the Ito isometry — you can compute

$$\mathbb{E}\left[\left(\int_0^\infty (h^n - h^m)\, dB\right)^2\right] = \mathbb{E}\left[\int_0^\infty (h^n - h^m)^2 \, ds\right]$$

by the Ito isometry. Why am I taking these differences? I want to produce a process h from this statement that these stochastic integrals converge in $L^2(\Omega, \mathcal{F}_\infty, \mathbb{P})$. So how am I going to produce this process h? Well, by this identity and the L^2 convergence, we get that the sequence (h^n) is Cauchy in $L^2(B)$ — because the norm in $L^2(B)$ of the difference between any two h^ns is precisely the right-hand side, and I know the left-hand side goes to 0 if you take n and m large enough (by the L^2 convergence of these random variables).

And by completeness in L^2, this means there exists a limit $h \in L^2(B)$ such that $h^n \to h$. And then now you use this Ito isometry again, and you get that actually $\int_0^\infty h^n \, dB \to \int_0^\infty h \, dB$ (again, this is in $L^2(\Omega, \mathcal{F}_\infty, \mathbb{P})$).

And now we're basically done. Our starting assumption was that $Z_n \to Z$ in L^2. The terms $\mathbb{E}[Z_n] \to \mathbb{E}[Z]$ as constants, and we just concluded that the $\int_0^\infty h^n \, dB$ terms converge to $\int_0^\infty h \, dB$. So by the uniqueness of limits, we get

$$Z = \mathbb{E}[Z] + \int_0^\infty h \, dB.$$

This finishes showing (1), that $\mathcal{H}$ is a closed subspace (because this means $Z \in \mathcal{H}$).

Now we'll do (2) — why does $\mathcal{H}$ contain this subspace? This is a bit imprecise, because that subspace is a space of complex random variables. What we mean by (2) is really that $\mathcal{H}$ contains the real and imaginary parts of all of these.

So basically, we need to cook up some integrand h which gives me precisely this random variable

$$\exp\left(i \sum_j \lambda_j (B_{t_j} - B_{t_{j-1}})\right),$$

at least up to constants. And basically, you kind of see this is similar to the exponential local martingales from last time.

So we fix some λ_j's and t_j's, and then we want to look at this random variable and show it's in $\mathcal{H}$. So we first define an integrand

$$g = \sum_{j=1}^n \mathbf{1}_{(t_{j-1}, t_j]} \lambda_j.$$

This integrand g is actually deterministic, so it's definitely a progressive process. Then we can define a martingale

$$M_t = \int_0^t g_s \, dB_s.$$

Based on this explicit form of the integrand, you can compute this exactly, and it's going to be

$$M_t = \sum_{j=1}^n \lambda_j (B_{t_j \wedge t} - B_{t_{j-1} \wedge t})$$

(you have to be slightly careful and take a min with your time t). Already this looks similar to the thing in your exponential — if you take $t = \infty$, it is precisely the thing in the exponential (up to the factor of i).

Now let's look at the exponential martingale

$$\mathcal{E}(iM)_t = \exp\left(iM_t + \frac{1}{2}\langle M, M\rangle_t\right).$$

And it's actually going to be a martingale because of this i, which means it's basically bounded (the first part is bounded by 1; the second part isn't bounded, but we can still control it). Substituting our formula for M, we get that this is

$$\exp\left(i\sum_{j=1}^{n}\lambda_j(B_{t_j\wedge t} - B_{t_{j-1}\wedge t}) + \frac{1}{2}\sum_{j=1}^{n}\lambda_j^2(t_j\wedge t - t_{j-1}\wedge t)\right)$$

(the second part is the QV, which you can explicitly compute).

And what is nice about the exponential martingale? Well, where is the integrand going to come from? We're trying to show this thing is a stochastic integral. So far, we just succeeded in writing it as an exponential local martingale, so where does the integral come from?

That's just the fact from last time that you have a SDE for the exponential local martingale, that

$$d\mathcal{E}(iM)_t = i\mathcal{E}(iM)_t\, dM_t.$$

And SDEs are interpreted in integral form, so what this SDE really means is that

$$\mathcal{E}(iM)_\infty = \mathcal{E}(iM)_0 + \int_0^\infty i\mathcal{E}(iM)_t\, dM_t.$$

One final step is I wanted an integral with respect to B, not M. But the way we defined M as a stochastic integral, basically $dM = g\, dB$. So I can substitute $g\, dB$ into here. The initial value is just going to be 1 (it's an exponential, and M starts at 0), so we get

$$\mathcal{E}(iM)_\infty = 1 + \int_0^\infty \mathcal{E}(iM)_t g_t\, dB_t.$$

And so now we're basically done. Why? We've shown that if you plug in this formula with time ∞, I get that

$$\exp\left(i\sum_{j=1}^{n}\lambda_j(B_{t_j} - B_{t_{j-1}})\right) = \exp\left(-\sum_{j=1}^{n}\frac{1}{2}\lambda_j^2(t_j - t_{j-1})\right)\cdot\left(1 + \int_0^\infty i\mathcal{E}(iM)_t\, g_t\, dB_t\right)$$

(I used this identity but moved the second part in $\mathcal{E}(iM)_t$ to the RHS).

So I've succeeded in writing the LHS as a constant times a stochastic integral (or 1 plus a stochastic integral). This stochastic integral is definitely in my vector space (you can check that $\mathcal{E}(iM)_t\, g_t \in L^2(B)$, but that's true). You also have to check that 1 is in the vector space, but that's trivially true (we can just take $h = 0$ and write $1 = \mathbb{E}[1] + 0$). I add these two and that's still in my vector space, and then I multiply by some constant and it's still in my vector space.

Technically I arrive at this identity and then I have to take real and imaginary parts (because the vector space is real-valued). But the real and imaginary parts of this thing lie in my vector space, and that's enough for density (since if we have a dense subspace of $L^2_{\mathbb{C}}$, then its real and imaginary parts form a dense subspace of $L^2_{\mathbb{R}}$). □

Corollary 18.2 is a direct consequence of Theorem 18.1; so let's see that.

Proof. First let's see this for L^2-bounded martingales — now let M be an L^2-bounded martingale. Because it's L^2-bounded, it's uniformly integrable; so there exists $M_\infty \in L^2(\Omega, \mathcal{F}_\infty, \mathbb{P})$ such that

$$M_t = \mathbb{E}[M_\infty \mid \mathcal{F}_t].$$

(The fact that it's actually in L^2 comes from the fact that you're L_2-bounded.)

Now it's kind of natural what you want to do — by the first part of the theorem, there exists $h \in L^2(B)$ such that you can write M_∞ as a stochastic integral

$$M_\infty = \mathbb{E}[M_\infty] + \int_0^\infty h_s \, dB_s.$$

Now you just take conditional expectations of both sides, and you get

$$\mathbb{E}[M_\infty \mid \mathcal{F}_t] = \mathbb{E}[M_t] + \mathbb{E}\left[\int_0^\infty h_s \, dB_s \mid \mathcal{F}_t\right].$$

And the conditional expectation of this integral is just going to be $\int_0^t h_s \, dB_s$. Why? Heuristically this is your integral up to time t; and the integral after that only looks at increments of BM after $\mathcal{F}_t$, which are independent of $\mathcal{F}_t$. Alternatively, this is just by definition — the integral up to ∞ is just $(h \cdot B)_\infty$, and the process $(h \cdot B)_t$ is a martingale. If you write out the martingale property, that just tells you this identity.

And this is exactly what the claim was, at least for L^2-bounded martingales.

Remark 18.4. Note that in particular, you're not assuming M is continuous. So you've actually shown that any L^2-bounded martingale (in the completed canonical filtration of BM) has a continuous modification — technically this is always up to almost sure equalities, and the continuous modification is given by the RHS (since by construction $\int_0^t h_s \, dB_s$ is continuous in t).

For contrast, for local martingales we're actually assuming it's a continuous local martingale. One reason is because we only defined continuous local martingales.

Now let's talk about local martingales — let M be a CLMG. As usual, we want to localize — so for instance, you can take

$$T_n = \inf\{t \geq 0 \mid |M_t| \geq n\}.$$

If you stop your local martingale at time T_n, it's bounded, so it's also L^2-bounded; thus applying this L^2-bounded result to the stopped martingale M^{T_n}, we obtain a progressive process $h^n \in L^2(B)$ such that

$$M_t^{T_n} = c_n + \int_0^t h^n \, dB_s$$

for all $t \geq 0$. In principle, this constant c_n might also depend on n, but we're going to show that it doesn't. Why? You just apply the $t = 0$ version of this identity and get $M_0^{T_n} = c_n$; but this is always just M_0. So c_n is just going to be M_0. (It's kind of implicit here that M_0 has to be almost surely constant — M is adapted to the completed canonical filtration of BM, where $\mathcal{F}_0$ is trivial — you're assuming the BM starts at 0, so $\mathcal{F}_0$ is just literally $\{\Omega, \emptyset\}$; and then when you complete it you add all your null sets, but it still contains only probability-0 or 1 events, so any random variable measurable with respect to it is almost surely constant.)

Now you want to stitch these processes together — we have an h^n for every n, but we want to show they agree with each other on common intervals. So why is that going to be true? Let $m \geq n$. You start with the identity

$$(M^{T_m})_t^{T_n} = M_t^{T_n}.$$

(This is because $T_m \geq T_n$.) That has to imply an identity for these integrals — because plugging in our formulas for these things, we get

$$(h^m \cdot B)_t^{T_n} = (h^n \cdot B)_t^{T_n}$$

(we're going back to the notation in Le Gall where we write stochastic integrals as $h \cdot B$; we may as well put in an additional stopping by T^n on the right, since the process is constant after T^n anyways).

And we proved that you can put this stopping time onto the integrator, so this implies

$$(h^m \cdot B^{T_n})_t = (h^n \cdot B^{T_n})_t$$

(in principle you can put the stopping time anywhere you want; we're putting it on the integrator, and the reason we want to do that is now we have the same integrators in the two places). So this actually implies

$$(h^m - h^n) \cdot B^{T_n} = 0.$$

This is true for all times t. And we basically already saw this earlier: You want to go from this statement to a statement that h^m and h^n are equal (at least on some interval). And for that, you can basically just use the Ito isometry — this implies that

$$\mathbb{E}\left[\int_0^{T_n} (h_s^n - h_m^s)^2 \, ds\right] = 0.$$

But basically, I'm claiming by Ito isometry, when you take the second moment of this thing, it's going to be equal to this (the fact you only integrate from 0 to T_n is reflected by the fact that your integrator becomes constant after T_n). But then this implies that almost surely $h^n = h^m$ (at least, almost everywhere) on $[0, T_n]$. (The almost sure comes from the fact that the expectation is 0, and 'almost everywhere' from the fact that the integral of the difference squared is 0, which only really lets us conclude almost everywhere equivalence.)

But this holds for all $m \geq n$; and we can swap the 'almost sure' and the 'for all.'

And this is basically enough to stitch — up to null modifications, you just define

$$h = \lim_{n \to \infty} h^n.$$

The point is that if you look at any time t, eventually the right-hand side becomes constant — eventually they just all agree with each other.

So we've found a candidate integrand h. And with this candidate integrand, we want to show why the stated result is true. And we want to verify it's actually in $L^2_{\text{loc}}(B)$.

By construction, this stitched process h satisfies

$$\int_0^{T_n} h_s^2 \, ds < \infty$$

for all n. Why? Well, on this time interval, h is equal to h^n. And you recall that h^n is in $L^2_{\text{loc}}(B^{T_n})$. And then if you unwind what it means to be in that, that's precisely that the expectation of this thing is finite, which means this thing is finite almost surely.

This is true for all n, and you have $T_n \uparrow \infty$. Combining these two things, you get that indeed $h \in L^2_{\text{loc}}(B)$.

> **Remark 18.5.** How would you come up with it? You essentially want to find a way to stitch things together, and after playing around with things enough, you get this.

So we've verified $h \in L^2_{\text{loc}}(B)$; now why does it give us this representation of our local martingale? Let's start with the thing that we have by construction — that

$$M_t^{T_n} = M_0 + \int_0^t h_s^n \, dB_s.$$

We can write this as

$$M_t^{T_n} = M_0 + \int_0^{T_n \wedge t} h_s^n \, dB_s$$

(that's kind of playing around with the fact that I can put my stopping time into the evaluation time). And the point of all this is that for times before my stopping time, by definition this is going to be the process h itself, so I get

$$M_t^{T_n} = M_0 + \int_0^{T_n \wedge t} h_s \, dB_s$$

(because as long as $s \le T_n$, all the h's become the same, which means $h_s = h_s^n$).

This is true pointwise for all t; then if you take $T_n \ge t$, this implies

$$M_t = M_0 + \int_0^t h_s \, dB_s$$

(because now this $T_n \wedge t$ becomes t). □

Remark 18.6. We didn't say much about uniqueness for Corollary 18.2, but that should be similar to uniqueness in Theorem 18.1.

§18.2 Some consequences

Let's now talk about some consequences of this martingale representation theorem. One consequence is the following.

Proposition 18.7

Let $(\mathcal{F}_t)$ be the completed canonical filtration of BM. Then $(\mathcal{F}_t)$ is both left and right continuous.

Recall this basically means that $\mathcal{F}_t = \bigcup_{s<t} \mathcal{F}_s$ and $\mathcal{F}_t = \bigcap_{s>t} \mathcal{F}_s$. So the information that comes just before time t is equal to the information all the way up to time t; and the information infinitesimally up to time t is the same as up to time t. This comes from the completion process somehow — the canonical filtration of BM itself does not satisfy this, but once you complete it to add in all these null events, it basically does.

Proof. We'll just talk about right-continuity (left-continuity is similar). Suppose we have some $Z \in \mathcal{F}_t^+$ with $Z \in L^2$ (recall that $\mathcal{F}_t^+ = \bigcap_{s>t} \mathcal{F}_s$). Now we want to say why Z is measurable with respect to $\mathcal{F}_t$. (Once you show that, you show these two σ-algebras are equal — of course $\mathcal{F}_t \subseteq \mathcal{F}_t^+$, so you just have to show the reverse direction. And if we can show this statement, then we can take Z to be indicator functions of events, and that'll imply the containment.)

We have that Z is in $\mathcal{F}_t^+ \subseteq \mathcal{F}_\infty$, so by the martingale representation theorem, there exists some integrand $h \in L^2(B)$ such that

$$Z = \mathbb{E}[Z] + \int_0^\infty h_s \, dB_s.$$

The point now is that Z is measurable with respect to this much smaller σ-algebra, so that should tell you that you don't actually have to integrate all the way up to ∞ — since $Z \in \mathcal{F}_{t+\varepsilon}$ for any $\varepsilon > 0$, if you fix any ε then you get

$$Z = \mathbb{E}[Z \mid \mathcal{F}_{t+\varepsilon}],$$

and then if you use this representation for Z and take conditional expectations, now you're only integrating up to time $t+\varepsilon$ — so we get

$$Z = \mathbb{E}[Z] + \int_0^{t+\varepsilon} h_s \, dB_s.$$

(All of this is almost sure equality.) Now taking $\varepsilon \to 0$, our first construction of stochastic integrals gives you a continuous process (a continuous L^2-bounded martingale). So this integral is continuous in your time endpoint, and taking $\varepsilon \to 0$, you get

$$Z = \mathbb{E}[Z] + \int_0^{t} h_s \, dB_s.$$

But this right-hand side is measurable with respect to $\mathcal{F}_t$ (since by construction the stochastic integral is an adapted process). Thus Z is almost surely equal to a random variable in $\mathcal{F}_t$. But because $(\mathcal{F}_t)$ is complete, this implies Z itself is in $\mathcal{F}_t$. (We're saying if I modify Z up to some null event then I get something in $\mathcal{F}_t$; and that event is itself contained in $\mathcal{F}_t$, because it's complete.) □

That's the first consequence of the martingale representation theorem. Now let's talk about the second one. This second consequence is simply saying that any martingale with respect to this filtration has a continuous modification. In the martingale representation theorem we proved this for L^2-bounded martingales, but now we're saying it'll be true for any martingale.

Proposition 18.8

Let $(\mathcal{F}_t)$ be the completed canonical filtration of BM. Then all martingales with respect to $(\mathcal{F}_t)$ have a continuous modification.

Proof. Somehow we just want to reduce to the L^2-bounded case. So how do we do that? First, instead of L^2-bounded, let's suppose that M is uniformly integrable. If you can prove it for UI martingales, then you'll also prove it for general martingales — because in general you can replace M by a stopped version of M at a deterministic time t_0, i.e., M^{t_0}. And we had this result from when we talked about martingales that if you have a martingale and stop it at a deterministic fixed time, then it's always going to be UI — because M_{t_0} 'closes' your martingale. And then you can take $t_0 \to \infty$ and get a continuous modification for all times.

So it suffices to just prove this for uniformly integrable martingales. And now we have

$$M_t = \mathbb{E}[M_\infty \mid \mathcal{F}_t]$$

for some random variable M_∞ which is $\mathcal{F}_\infty$-measurable.

Now let's take a sequence of bounded random variables (M_∞^n) (which you think of as approximating the endpoint M_∞) such that $M_\infty^n \to M_\infty$ in L^1. (This is the best kind of convergence you could hope for, because we only assumed $M_\infty \in L^1$. But the good thing is you can always produce such a sequence of bounded things.)

Then each M_∞^n is going to give you a L^2-bounded martingale. One thing to note is first of all that

$$M_t^n = \mathbb{E}[M_\infty^n \mid \mathcal{F}_t] \to M_t$$

in L^1. Now by the L^2 result, we can write this endpoint M_∞^n as a stochastic integral (because it's bounded, so it's definitely in L^2), giving

$$M_\infty^n = \mathbb{E}[M_\infty^n] + \int_0^\infty h^n \, dB.$$

And now, by the L^2 result, the martingale (M_t^n) has a continuous modification for all n (this is how we're going to apply the L^2 result). So let's replace M^n by its continuous modification.

Now we kind of want to say why — M^n is a sequence of continuous functions, and we want to show that as $h \to \infty$, this converges to a continuous function. As soon as you can show that (at least, on a subsequence), the limiting thing has to be a continuous modification of your original martingale (because you've shown this converges pointwise in L^1 to M_t, and if you can show it converges also pointwise to a continuous function, then that limit has to be almost surely equal to M_t).

We'll do this by Borel–Cantelli-type arguments. To show (M_t^n) converges uniformly in the space of continuous functions, we want to find a Cauchy sequence (or rather, a Cauchy subsequence). And the way you find that is by proving about

$$\mathbb{P}\left[\sup_{t\geq 0} |M_t^n - M_t^n| \geq \lambda\right].$$

First, by Doob's maximal inequality you can bound this by

$$\frac{3}{\lambda}\mathbb{E}\left[|M_\infty^n - M_\infty^m|\right]$$

(if you have a UI martingale, the tails of the sup are bounded by the L^1 norm of the endpoint, with a Markov-type bound; and the difference of two martingales is a martingale).

The thing that eventually lets you reduce to a convergent subsequence is you know the RHS is Cauchy, because by construction $M_\infty^n \to M_\infty$ in L^1. So it's Cauchy, and then you can construct a very good subsequence. Basically, since $M_\infty^n \to M_\infty$ in L^1, there's going to exist a subsequence $\{n_k\}$ such that

$$\mathbb{E}[\left|M_\infty^{n_k} - M_\infty^{n_{k+1}}\right|] \leq 2^{-2k}.$$

For this subsequence, applying Doob's maximal inequality gives you that

$$\mathbb{P}\left[\sup_{t\geq 0}\left|M_t^{n_{k+1}} - M_t^{n_k}\right| \geq 2^{-k}\right] \leq 32^k 2^{-2k} = 3\cdot 2^{-k}$$

(the exact constants don't matter; I just want both the difference and the probability to be exponentially small). And now if I sum over k, these probabilities are summable, so by Borel–Cantelli I I get that with probability 0 these events happen infinitely often — in other words, eventually almost surely the sup of the difference of successive martingales is less than 2^{-k}, i.e.,

$$\sup_t \left|M_t^{n_{k+1}} - M_t^{n_k}\right| < 2^{-k}$$

for all sufficiently large k. This implies the subsequence converges — for instance, now if you sum all these successive differences, you get

$$\sum_k \sup_t \left|M_{n_{k+1}} - M_t^{n_k}\right| < \infty$$

(there's finitely many terms you have no control over, but eventually you get 2^{-k} decay, so it's summable). And this implies (M^{n_k}) is Cauchy in the spce of continuous functions $\mathcal{C}([0,\infty))$ (with the sup norm) — it's not even just uniform convergence on compact sets, it's actually uniform convergence on the entire infinite real line, which is quite strong.

And this implies there exists $\widetilde{M}$ such that $M^{n_k} \to \widetilde{M}$ in $\mathcal{C}([0,\infty))$. So it converges uniformly; in particular, that also means it converges pointwise, i.e., almost surely $M_t^{n_k} \to \widetilde{M}_t$.

And now you recall that $M_t^{n_k} \to M_t$ in L^1. And by uniqueness of these limits, that implies $M_t = \widetilde{M}_t$ almost surely. In other words, $\widetilde{M}$ is a modification of M, and it's continuous by construction. So you have a continuous modification. □

Remark 18.9. We've seen this type of argument several times now — it's this technical thing where you want to extract a subsequence that converges uniformly to get a continuous modification.

Next time we'll start with Girsanov's theorem, which is the second half of applications of Ito's formula (it's quite separate from this stuff); it's about what happens to martingales when you change your underlying probability measure.

Remark 18.10. One thing to emphasize about the martingale representation theorem: I know there exists h, but is there actually a formula for it? It turns out that there is, and it should be given by something like

$$h = \mathbb{E}[\mathcal{D}_t Z \mid \mathcal{F}_t],$$

where $\mathcal{D}_t$ denotes the 'Malliavin derivative of Z.' We probably won't have time to discuss this in the course, but building on stochastic calculus you can build a notion of differentiation for random variables in L^2. Essentially, we'll have $Z = f(B)$ — it's a function of your Brownian motion (some complicated function defined via stochastic integration). So you can think about differentiating this with respect to BM — if you vary the BM in some direction, how does the function change? This is all taking place in infinite dimensions — B itself lies in a space of functions, so you have to talk about directional derivatives in some infinite space. But you can make sense of it and you get the Malliavin derivative. A more refined version of the martingale representation theorem is saying that for this integrand h, you have to take a derivative of Z. And maybe you also have to condition on $\mathcal{F}_t$ because after all h had better be adapted — in principle $\mathcal{D}_t Z$ might not be $\mathcal{F}_t$-measurable, which is why you need this.

If you're interested in learning more, this is one thing that occurs in stochastic analysis — differentiation on infinite spaces with respect to BM.

This is also related to analysis and mathematical physics and quantum field theory; Malliavin calculus appears there. It's also related to geometry.

§19 April 14, 2025 — Girsanov's theorem

Today we'll start on Girsanov's theorem, which is the last main application of Ito's formula that we'll talk about. This might take 1.5 classes; we probably won't finish everything today.

Girsanov's theorem has to do with change of measure. To start, we won't go in the order of the notes; we'll first discuss an analogy with Gaussian change of measure to set the stage.

§19.1 Gaussian change of measure

What is Gaussian change of measure? Suppose you have $X \sim \mathcal{N}_d(0, \Sigma)$ (this means X is a Gaussian random variable in d dimensions with mean 0 and covariance matrix Σ). Then if you fix $\mu \in \mathbb{R}^d$, you can define a new probability measure $\mathbb{Q}$ on our underlying space $(\Omega, \mathcal{F})$ (where we originally had a probability measure $\mathbb{P}$). We'll define it by its density with respect to $\mathbb{P}$ — we set

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = \exp\left(\mu \cdot X - \frac{1}{2}\mu^\intercal \Sigma \mu\right).$$

This is linear in X; the second term $\mu^\intercal \Sigma \mu$ is just there to make the total mass 1 (since the total mass of $\mathbb{Q}$ is going to be the expectation of this random variable; then you recall the formula for the exponential generating function and see that $\mathbb{E}[\exp(\mu \cdot X)] = \exp(\frac{1}{2}\mu^\intercal \Sigma \mu)$, so you have to divide by this).

Claim 19.1 — Under $\mathbb{Q}$, we have $X \sim \mathcal{N}_d(\Sigma\mu, \Sigma)$.

So X is still normal; its mean has shifted, but its covariance matrix is the same.

What does this mean? By definition, a random variable is a measurable function $X : (\Omega, \mathcal{F}) \to \mathbb{R}$. Whenever you write $X \sim \mathcal{N}_d(0, \Sigma)$, you're saying that when you have probability measure $\mathbb{P}$, its law is distributed like this. Here we've changed the underlying probability measure, so the law may have changed. And Gaussian change of variables says that when you reweight by this exponential-linear thing, you just get a mean shift.

Proof. This is just a computation — you can compute $\mathbb{E}_{\mathbb{Q}}[f(X)]$ (where f is some bounded measurable function). If the claim is true, then in the end, we want this to be equal to

$$\int_{\mathbb{R}^d} f(x) p_{\Sigma\mu,\sigma}(x)\, dx,$$

where $p_{\Sigma\mu,\Sigma}$ is the density of $\mathcal{N}_d(\Sigma\mu, \Sigma)$. What does it mean for X to have a certain law? That means if I compute expectations of any function f, then I should get the integral of f against the density.

And because we defined $\mathbb{Q}$ by its density, you can compute the expectation over $\mathbb{Q}$ as an expectation over $\mathbb{P}$, but adding in this density term — so

$$\mathbb{E}_{\mathbb{Q}}[f(X)] = \mathbb{E}\left[f(X) \exp\left(\mu \cdot X - \frac{1}{2}\mu^{\intercal}\Sigma\mu\right)\right].$$

(Whenever we write $\mathbb{E}_{\mathbb{Q}}$, we're taking an expectation with respect to the new probability measure; when we write $\mathbb{E}$, we're using the original underlying one.)

But the assumption is that under $\mathbb{P}$, X has the explicit law $\mathcal{N}_d(0, \Sigma)$. That means we can write this expectation as an explicit d-dimensional integral

$$\int_{\mathbb{R}^d} f(x) p_{0,\Sigma}(x) \exp\left(\mu \cdot x - \frac{1}{2}\mu^{\intercal}\Sigma\mu\right)\, dx.$$

But now I want to actually write out the explicit formula for this density, and this becomes

$$\frac{1}{(2\pi)^{d/2}\det(\Sigma)^{1/2}} \int_{\mathbb{R}^d} f(x) \exp\left(-\frac{1}{2}x^{\intercal}\Sigma^{-1}x\right) \exp\left(\mu \cdot x - \frac{1}{2}\mu^{\intercal}\Sigma\mu\right)\, dx$$

(the prefactor and the first term form the density). Now you want to complete the square — you want to interpret this product of exponentials as itself a Gaussian density, and the reason you hope to do so is that this thing is still quadratic in x. There's now a quadratic term and a linear term, and you can absorb the linear term by shifting x. If you do the algebra correctly, this becomes

$$\frac{1}{(2\pi)^{d/2}\det(\Sigma)^{1/2}} \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} f(x) \exp\left(-\frac{1}{2}(x - \Sigma\mu)^{\intercal}\Sigma^{-1}(x - \Sigma\mu)\right).$$

This extra term precisely allows you to complete the square — the quadratic terms are the same as before, the linear term in x is $\frac{1}{2} \cdot 2(\Sigma\mu)^{\intercal}\Sigma^{-1}x = \mu \cdot x$, and the last term expands out to $\frac{1}{2}\mu^{\intercal}\Sigma\mu$.

But now you're done — this prefactor times this exponential is precisely the density $p_{\Sigma\mu,\Sigma}$. □

So this is a Gaussian calculation; the point is that under this change of measure, the mean shifts by $\Sigma\mu$ (so you have to look at the covariance and the μ that you're reweighting by), but then the covariance stays the same. So that's the key thing to take out of this example.

§19.2 Analogy to Girsanov's theorem

And basically what Girsanov's theorem is — you can think of it conceptually as an 'infinitesimal version' or analog of Gaussian change of measure. In our world of stochastic calculus, you have stochastic processes naturally indexed by time. Here you had d-dimensional Euclidean space as your index set, where there isn't really a natural time-ordering. But with martingales you do have this natural notion of time; so infinitesimal refers to infinitesimally in time.

It helps in all of this to think of a mean zero Gaussian as being analogous to being a martingale. Here we started with a mean 0 Gaussian; in Girsanov's theorem, we're going to start with a martingale (or local martingale).

And you can kind of think of the covariance as a quadratic variation — that's kind of the analogy.

Basically what Girsanov tells you is that if you have a martingale and use that to reweight your probability measure, then any martingale under the new probability measure is going to pick up an extra mean term — so the mean is going to be shifted (it'll no longer be a martingale). And the shift is some analog of $\Sigma\mu$, which has to do with the QV.

§19.3 Preliminaries

With that, let's now begin with some preliminary stuff. We can't just directly state Girsanov's theorem; we need a couple of preliminary results.

Throughout, we'll assume that $(\mathcal{F}_t)$ is complete and right-continuous. (Right-continuity — probably Le Gall assumes it for technical reasons, and we're skipping over some of the technical stuff, so we might not see why it's needed.)

Proposition 19.2 (Likelihood ratio process)

Let $\mathbb{Q}$ be a probability measure on $(\Omega, \mathcal{F})$ which is absolutely continuous (with respect to $\mathbb{P}$) on $\mathcal{F}_\infty$.

For every $t \in [0, \infty]$ (allowing ∞), let

$$D_t = \left.\frac{d\mathbb{Q}}{d\mathbb{P}}\right|_{\mathcal{F}_t}.$$

Then (D_t) is a uniformly integrable martingale.

Furthermore, suppose that (D_t) is continuous. Then for all stopping times T, you can write

$$D_T = \left.\frac{d\mathbb{Q}}{d\mathbb{P}}\right|_{\mathcal{F}_T}.$$

Finally, if $\mathbb{P}$ and $\mathbb{Q}$ are mutually absolutely continuous with respect to $\mathcal{F}_\infty$, then

$$\inf_{t \geq 0} D_t > 0 \quad \mathbb{P}\text{-almost surely.}$$

The way we write 'absolutely continuous' is as $\mathbb{Q} \ll \mathbb{P}$. What it means to be absolutely continuous on $\mathcal{F}_\infty$ is for any $A \in \mathcal{F}_\infty$, if $\mathbb{P}[A] = 0$, then $\mathbb{Q}[A] = 0$.

What does D_t mean? If $\mathbb{Q}$ is absolutely continuous with respect to $\mathbb{P}$ on $\mathcal{F}_\infty$, this is also true on every $\mathcal{F}_t$. And whenever you have this absolute continuity, you can write its Radon–Nikodym derivative — this will be some $\mathcal{F}_t$-measurable function with the property that if you integrate the Radon–Nikodym derivative (with respect to $\mathbb{P}$) against any event in your σ-algebra, you just get the mass of it under $\mathbb{Q}$ — i.e., for every $A \in \mathcal{F}_t$, we have

$$\mathbb{E}[D_t \mathbf{1}_A] = \mathbb{Q}[A].$$

The way we're going to show that (D_t) is a UI martingale is by showing it's the expectation with respect to $\mathcal{F}_t$ of D_0.

A priori you don't know (D_t) is continuous. This is perhaps where right-continuity of the filtration helps you (then (D_t) should have a continuous modification). But we're skipping over that technical stuff, so we'll just assume that (D_t) is continuous.

Mutually absolutely continuous means that $\mathbb{P}$ is also absolutely continuous with respect to $\mathbb{Q}$.

Why is this called 'likelihood ratio'? You can think of D as a density — the density of $\mathbb{Q}$ with respect to $\mathbb{P}$, which is another name of the likelihood ratio. So you're saying this is one way to get martingales — you have an absolutely continuous measure $\mathbb{Q}$ with respect to $\mathbb{P}$, and condition on the sub-σ-algebras.

Proof. First, we claim that

$$D_t = \mathbb{E}[D_\infty \mid \mathcal{F}_t].$$

Note that D_∞ is nonnegative (by the definition of a Radon–Nikodym derivative) and integrates to 1; so as long as you show this, you get a UI martingale.

To see this, first note that $D_t \in \mathcal{F}_t$ (by the definition of a Radon–Nikodym derivative). That's the first property of a conditional expectation. Moreover, for $A \in \mathcal{F}_t$, you get that

$$\mathbb{E}[D_t \mathbf{1}_A] = \mathbb{Q}[A].$$

But $\mathcal{F}_t \subseteq \mathcal{F}_\infty$, so we could also write $\mathbb{Q}[A]$ as

$$\mathbb{Q}[A] = \mathbb{E}[D_\infty \mathbf{1}_A].$$

But this is the defining property of conditional expectations, showing that this is true.

Now we'll move on to the second part, about stopping times. The reason you need to assume you have continuous (or right-continuous) sample paths is because you want to apply the optional stopping theorem (which has some assumption of continuity). Then you can say that

$$D_T = \mathbb{E}[D_\infty \mid \mathcal{F}_T]$$

(because this is a UI martingale). But you can also show that

$$\mathbb{E}[D_\infty \mid \mathcal{F}_T] = \left.\frac{d\mathbb{Q}}{d\mathbb{P}}\right|_{\mathcal{F}_T}.$$

This is basically the same argument we just did — if you have some $A \in \mathcal{F}_T$, then we have that

$$\mathbb{E}[D_T \mathbf{1}_A] = \mathbb{E}[\mathbb{E}[D_\infty \mid \mathcal{F}_T]\mathbf{1}_A] = \mathbb{E}[D_\infty \mathbf{1}_A] = \mathbb{Q}[A].$$

And because $A \in \mathcal{F}_T$ by assumption, I can also write this using the RN derivative as

$$\mathbb{E}\left[\left.\frac{d\mathbb{Q}}{d\mathbb{P}}\right|_{\mathcal{F}_T} \mathbf{1}_A\right].$$

This is true for every A, and it's enough to pove what we want (since conditional expectations are unique).

For the third part, the argument in the notes is slightly different from the book, so you can also read that if you want. The one in the notes assumes something about RN derivatives — namely that since $\mathbb{Q} \ll \mathbb{P}$ and $\mathbb{P} \ll \mathbb{Q}$ with respect to $\mathcal{F}_\infty$, what you really should have then is that

$$D_t^{-1} = \left.\frac{d\mathbb{P}}{d\mathbb{Q}}\right|_{\mathcal{F}_t}.$$

In the notes we didn't write any justification. But the reason it should be intuitive is that the defining property of D_t is that

$$\mathbb{E}\left[D_t Z\right] = \mathbb{E}_{\mathbb{Q}}[Z]$$

for any nonnegative random variable Z (previously we wrote indicator functions, but we can instead replace the indicator with Z). But now you can just say that if I'm allowed to have any nonnegative random variable, I can plug in $Z = Y D_t^{-1}$. That should give me the identity

$$\mathbb{E}[Y] = \mathbb{E}_{\mathbb{Q}}[Y D_t^{-1}]$$

(at least, formally). And if you have this identity, that's precisely the statement

$$D_t^{-1} = \left.\frac{d\mathbb{P}}{d\mathbb{Q}}\right|_{\mathcal{F}_t},$$

since now you're reducing derivatives with respect to $\mathbb{P}$ into ones with respect to $\mathbb{Q}$.

Of course, there's something you have to make precise here – you have to use mutual absolute continuity somewhere, and we didn't really explain where. But we'll just assume this is true.

Then under $\mathbb{Q}$, we have that (D_t^{-1}) is a UI martingale — we're basically applying the first result, but with the roles of $\mathbb{P}$ and $\mathbb{Q}$ reversed. And we're still assuming D_t is continuous, so D_t^{-1} is also continuous. Applying the proof of the first result, you get the martingale structure

$$\mathbb{E}_{\mathbb{Q}}[D_\infty^{-1} \mid \mathcal{F}_t] = D_t^{-1}.$$

Then we have

$$D_t^{-1} \to D_\infty^{-1} \quad \mathbb{Q}\text{-almost surely}$$

(when you have a UI martingale, you converge almost surely to the limit (and also in L^1)). But we also have

$$D_\infty^{-1} < \infty \quad \mathbb{Q}\text{-almost surely.}$$

(Why is that true? If I compute $\mathbb{E}_{\mathbb{Q}}[D_\infty^{-1}]$, this should be the same as computing $\mathbb{E}_{\mathbb{P}}[1]$ — this is again one of the consequences of the Radon–Nikodym derivative. But this is just 1. So $\mathbb{E}_{\mathbb{Q}}[D_\infty^{-1}]$ is finite, which means $D_\infty^{-1} < \infty$ almost surely.)

If you combine these two things (this is again why we need to assume your sample paths are continuous), we get that

$$\sup_{t \geq 0} D_t^{-1} < \infty \quad \mathbb{Q}\text{-almost surely}$$

(since you have a continuous function in time that converges at ∞ to something finite, so the entire path has to be bounded).

Then you use mutual absolute continuity to go from $\mathbb{Q}$-almost surely to $\mathbb{P}$-almost surely (probability 1 events are the same under both measures).

But now if you take inverses, you get that

$$\inf D_t^{-1} > 0 \quad \mathbb{P}\text{-almost surely.}$$ □

Somehow, another kind of intuitive thing is that this uniform integrability condition sort of tells you you're not losing any probabilitiy mass as you go to ∞, because $D_t \to D_\infty$ in L^1 as well as almost surely. This is saying it's somehow intimately related to absolute continuity in some way.

The main thing about this proposition is that if you have absolutely continuous probability measures, that gives you a natural martingale.

Now let's talk about one more preliminary result before we get to Girsanov.

Proposition 19.3

Let D be a CLMG taking strictly positive values. Then there exists a unique CLMG L such that

$$D_t = \exp\left(L_t - \frac{1}{2}\langle L, L\rangle_t\right).$$

Moreover, we have an explicit formula for L — we have

$$L_t = \log D_0 + \int_0^t D_s^{-1}\, dD_s.$$

Here D is a CLMG; because it takes strictly positive values, D_s^{-1} is finite (on any finite interval, D^{-1} is going to be bounded); that's why this stochastic integral will be well-defined. So that's the statement of the theorem.

Proof. As usual, uniqueness kind of just follows — if you had L_t and $\widetilde{L}_t$ which worked, that implies that both become D once you take this exponential thing, but that implies

$$L_t - \frac{1}{2}\langle L, L\rangle_t = \widetilde{L}_t - \frac{1}{2}\langle \widetilde{L}, \widetilde{L}\rangle_t.$$

But that implies $L - \widetilde{L}$ is both a CLMG (because each is a CLMG) and a FV process (because you can move the CLMGs to the LHS and the QVs to the RHS, so you get that this difference is also a difference of QVs), which means $L - \widetilde{L} = 0$.

For existence, you can recall that this form looks very familiar — it's actually just $\mathcal{E}(L)_t$. So you want to find L whose exponential local martingale gives you D.

But the exponnetial local martingale satisfies $d\mathcal{E}(L)_t = \mathcal{E}(L)_t\, dL_t$. And you want $\mathcal{E}(L)_t$ to be D. So you want to find L such that

$$dD_t = D_t\, dL_t.$$

And then you make the guess that you should have $dL_t = D_t^{-1}\, dD_t$. That's where your guess comes from.

Now if you define dL_t in this way, you just have to verify why this identity is true. And why is it going to be true? The way Sky did it in the notes is that you can compute

$$d(\log D_t)$$

(the evolution of $\log D$ — here D takes strictly positive values, so going back to the Remark put on the homework, technically log is not C^2 on all of $\mathbb{R}$, but if you assume your process takes strictly positive values, you can do some localization argument to still apply Itô's formula to $\log D$), the derivative of $\log x$ is $\frac{1}{x}$ and the second is $-\frac{1}{x^2}$, so you get

$$d(\log D_t) = \frac{dD_t}{D_t} - \frac{1}{2}\frac{1}{D_t^2} d\langle D, D\rangle_t.$$

In the notes, this was maybe written a bit confusingly; but you just want to verify that the right-hand side itself you can write as $dL_t - \frac{1}{2}d\langle L, L\rangle_t$. Because if I can now say that the RHS is given by this evolution, then I just integrate both sides, and then I get that

$$\log D_t - \log D_0 = L_t - L_0 - \frac{1}{2}\langle L, L\rangle_t.$$

And then you see I just need to set $L_0 = \log D_0$ (which is completely consistent with our identity) — because the QV at time 0 starts at 0, and you want $D_0 = e^{L_0}$, which is why you want this.

So we just need to verify that

$$\frac{dD_t}{D_t} - \frac{1}{2}\frac{1}{D_t^2}d\langle D, D\rangle_t = dL_t - \frac{1}{2}d\langle L, L\rangle_t.$$

For this, you see that dL is already the $\frac{dD_t}{D_t}$ term by definition, so we just want to say why the second terms are equal — so I just want to compute

$$\frac{1}{2}d\langle L, L\rangle_t.$$

And going back to properties of the QV, the way you internally think about it is that $d\langle L, L\rangle_t = dL_t\, dL_t$. And if you think about it this way, you get

$$\frac{1}{2}D_t^{-2}\, dD_t\, dD_t = \frac{1}{2}D_t^{-2}d\langle D, D\rangle_t.$$

The more formal way to say this is that

$$\langle (H \cdot K), M\rangle = H \cdot \langle K, M\rangle,$$

and similarly if both your things are stochastic integrals, applying this thing twice gives you

$$\langle H \cdot K, H \cdot K\rangle = H^2 \cdot \langle K, K\rangle.$$

All we're doing is writing this in differential notation, which is the same as this integral identity. (This second way is how Le Gall writes it, but the first way is how at least Sky does calculations.) □

Basically, the outcome of this is that the first proposition gives us a strictly positive martingale (it's not just a local martingale, but an actual martingale; and it's strictly positive because we're going to assume mutual absolute continuity). That means we can now write that UI martingale as the exponential martingale of some L. That's kind of how you combine these two propositions. Now with that, we can talk about Girsanov's theorem.

§19.4 Girsanov's theorem

Theorem 19.4 (Girsanov)

Let $\mathbb{P}$ and $\mathbb{Q}$ be mutually absolutely continuous with respect to $\mathcal{F}_\infty$. Let

$$D_t = \left.\frac{d\mathbb{Q}}{d\mathbb{P}}\right|_{\mathcal{F}_t}.$$

Assume that (D_t) has continuous sample paths, and let (L_t) be the CLMG such that $D = \mathcal{E}(L)$.

Then if M is a CLMG under $\mathbb{P}$, then $N - \langle M, L\rangle$ is a CLMG under $\mathbb{Q}$.

In all our applications, the fact that (D_t) has continuous sample paths is going to be guaranteed. The second proposition guarantees us that we can go from D to L (using the explicit formula stated).

We've erased the analogy with Gaussian change of measure; but you think of $\mathbb{Q}$ as a reweighting of $\mathbb{P}$ by this exponential local martingale of L. So L before was $\mu \cdot X$ in the Gaussian change of measure analogy. And now you're saying if I start with a martingale under $\mathbb{P}$ (which you think of as something with mean 0), when I change the measure, it no longer has mean 0 — M itself is no longer a martingale. But you subtract out by the mean shift; and if you subtract out the mean shift $\langle M, L\rangle$, it really is a martingale. You should think of $\langle M, L\rangle$ as analogous to $\Sigma\mu$ in the Gaussian change of measure. So this is saying under a change of measure, martingales have a mean-shift, and if you subtract that out you get back a martingale.

So that's the statement; now let's prove it. It's really an exercise in Ito's formula.

Remark 19.5. Before the proof, we'll remark that the notion of a local martingale depends on your underlying probability measure (even for the notion of a martingale, you're always taking expectations with respect to your probability measure). On the other hand, the notion of a FV process does *not* depend on the underlying measure. So basically the QV (or really, the bracket), which by construction is a FV process, is FV under both $\mathbb{P}$ and $\mathbb{Q}$. Because it's a pathwise statement — that every sample path is a FV function — so there's nothing about expectations in that definition.

As a consequence, M is still going to be a semimartingale under $\mathbb{Q}$ — $M - \langle M, L\rangle$ is going to be a CLMG, and its FV part is going to be $\langle M, L\rangle$.

Proof. This is an exercise in Ito's formula; let's see why. The claim is that $M - \langle M, L\rangle$ is going to be a martingale under $\mathbb{Q}$; this means I have to think about taking expectations like

$$\mathbb{E}_{\mathbb{Q}}[(M_t - \langle M, L\rangle_t)].$$

At least informally, I'd expect this to be 0. But $\mathbb{Q}$ is defined by a Radon–Nikodym derivative, so you can transform expectations over $\mathbb{Q}$ to ones over $\mathbb{P}$ — basically, you want

$$\mathbb{E}[D_t(M_t - \langle M, L\rangle_t)] = 0.$$

This is just informal. What you'd actually expect is that when you compute the Ito differential under $\mathbb{P}$, you should have

$$d(D_t(M_t - \langle M, L\rangle)_t) = d\text{CLMG}.$$

The point is when you compute Ito differentials, you have martingale parts and finite variation terms; and the claim is that all the FV terms cancel out, so you get a CLMG. We'll see that this will imply when you back to $\mathbb{Q}$, you'll have just the first part being a CLMG.

So in summary:

Claim 19.6 — Under $\mathbb{P}$, we have that $D_t(M_t - \langle M, L\rangle_t)$ is a CLMG.

Proof. Like we said, you want to apply Ito's formula to a product of two semimartingales; as usual, you get

$$d(D_t(M_t - \langle M, L\rangle_t)) = (dD_t)(M_t - \langle M, L\rangle_t) + D_t(dM_t - d\langle M, L\rangle_t) + dD_t(dM_t - d\langle M, L\rangle_t).$$

Now we have to find some sort of cancellation using the definition of D.

Any martingale term we're okay with — so for the first term, $dD_t \cdot (\cdots)$ is a martingale term, because D is a CLMG. So this term we're okay with — it's integration against a CLMG, which gives you a CLMG. So we can ignore this whole first term.

For the second term, $D_t\, dM_t$ is also fine (because M is by assumption a CLMG). But we do have to care about the $D_t\, d\langle M, L\rangle_t$ term.

For the last term, $dD_t\, d\langle M, L\rangle_t$ is 0 because $\langle M, L\rangle$ is FV — you can only have nonzero brackets when things have martingale parts.

So you're going to get

$$-D_t\, d\langle M, L\rangle_t + dD_t\, dM_t + \text{CLMG terms}.$$

And we have to say why this first thing is 0. But for this, recall that $D = \mathcal{E}(L)$, which means $dD_t = D_t\, dL_t$ (since the exponential local martingale satisfies this SDE). Now you plug this in and get that the first line is

$$-D_t\, d\langle M, L\rangle_t + D_t\, dL_t\, dM_t.$$

And now you see you can combine $dL_t\, dM_t$ into $d\langle M, L\rangle_t$, and this is exactly the cancellation you want.

So all you're left with in this computation is local martingale terms, which you're okay with. □

So we've proved Claim 1, that this is a local martingale. Now how do we finish? Assuming no issues with integrability — basically, just assume everything is bounded, or at least has all moments, so there are no issues with integrability (so we're assuming e.g. that it's not just a CLMG, but actually a martingale) — then the martingale identity says that

$$\mathbb{E}[D_t(M_t - \langle M, L\rangle_t) \mid \mathcal{F}_s] = D_s(M_s - \langle M, L\rangle_s).$$

Now you want to use this to conclude why $M - \langle M, L\rangle$ is a martingale under $\mathbb{Q}$. But if we have $A \in \mathcal{F}_s$, then

$$\mathbb{E}_{\mathbb{Q}}[(M_t - \langle M, L\rangle_t)\mathbf{1}_A],$$

by the definition of conditional expectation I need to say why I can replace the t with the s here (that'll give me the martingale identity). For this, we convert $\mathbb{Q}$-expectations into $\mathbb{P}$-expectations by introducing D — we get that

$$\mathbb{E}_{\mathbb{Q}}[(M_t - \langle M, L\rangle_t)\mathbf{1}_A] = \mathbb{E}[D_t(M_t - \langle M, L\rangle_t)\mathbf{1}_A].$$

And now using the martingale property for this, we can replace t with s to say this is

$$\mathbb{E}[D_s(M_s - \langle M, L\rangle_s)\mathbf{1}_A].$$

And now I can go back, and this is also

$$\mathbb{E}_{\mathbb{Q}}[(M_s - \langle M, L\rangle_s)\mathbf{1}_A].$$

So that gives the martingale property, in the ideal scenario where this thing is a martingale.

Now we have to handle the general case, where this thing is just a CLMG. That's just the usual localization arguments, though it's not super immediate.

So to localize, we let

$$T_n = \inf\{t \geq 0 \mid \langle L, L\rangle \geq n \text{ or } |L_t| \geq n \text{ or } |M_t| \geq n \text{ or } \langle M, L\rangle_t \geq n\}$$

(we basically want everything in the expectations to be bounded; the first two things ensure that D is bounded, and the last two things ensure that $M_t - \langle M, L\rangle_t$ is bounded). Then for each n, the stopped process

$$D_t^{T_n}(M_t^{T_n} - \langle M, L\rangle_t^{T_n})$$

is a bounded CLMG, and thus in fact a bounded martingale. This combines several results — one that if you start with a CLMG and then add in a stopping time, you still have a CLMG. And by definition of this stopping time, it's going to be bounded; and we have a result saying any bounded CLMG is actually a martingale.

So that's the first step of the localization. Now all you want to do is now that you have a finite n, you get some martingale identity like the above thing; and then you just send $n \to \infty$.

But you have to be kind of careful, for reasons we'll explain. In the end, you want to verify that $M - \langle M, L\rangle$ is a local martingale; now that you've localized and gotten a bounded martingale, you want to show that under $\mathbb{Q}$, this stopped process $M^{T_n} - \langle M, L\rangle^{T_n}$ is a martingale. (Once we show this, we're done — the definition of CLMG just means you can find a sequence of stopping times $T_n \uparrow \infty$ such that for every n you have a UI martingale).

Your first try would just be to mimic what we did before — you want to fix $s \leq t$, take an event $A \in \mathcal{F}_s$, and then you want to compute

$$\mathbb{E}_{\mathbb{Q}}[(M_t^{T_n} - \langle M, L\rangle_t^{T_n})\mathbf{1}_A].$$

And then you'd do the usual thing where you convert this into an expectation with respect to $\mathbb{P}$, as

$$\mathbb{E}_{\mathbb{Q}}[D_\bullet \cdot (M_t^{T_n} - \langle M, L\rangle_t^{T_n})\mathbf{1}_A].$$

But you have to be careful what you put in $D_\bullet$ — you can only really put in D_t. What you really want to put in is $D_t^{T_n}$, since you know that the stopped process itself is a $\mathbb{P}$-martingale. But you don't actually have this identity *a priori* where you can put in $D_t^{T_n}$. Why? In general, you worry about if $T_n \le t$. What we want to say is that the time you put in here has to be larger than the times in the rest of the thing — $\mathbf{1}_A$ is going to be $\mathcal{F}_s$-measurable, and $(M_t^{T_n} - \langle M, L\rangle_t^{T_n})$ is going to be $\mathcal{F}_{T_n \wedge t}$-measurable. So we need to make sure our time in D is larger than both of these. So the problem is if for some reason you have $T_n \le s$ — in that case you cannot put $D_{T_n \wedge t}$ here, because you need the time to be greater than both the times $T_n \wedge t$ and s (kind of by definition of what the density is).

So at best, what you can hope for is probably D_t (since definitely $t \ge s$ and $t \ge T_n \wedge t$). But then you have a problem — because this is not the thing you know is a martingale; you only know it's a martingale with $D_t^{T_n}$.

So that's why we said you have to be careful. But you can get around this. How? You first just take your event $A \in \mathcal{F}_{T_n \wedge s}$. For such an event, you *can* actually write

$$\mathbb{E}_{\mathbb{Q}}[(M_t^{T_n} - \langle M, L\rangle_t^{T_n})\mathbf{1}_A] = \mathbb{E}[D_t^{T_n}(M_t^{T_n} - \langle M, L\rangle_t^{T_n})\mathbf{1}_A],$$

because now you're not worried about if $T_n \le s$ (then $\mathbf{1}_A$ is going to be $\mathcal{F}_{T_n}$-measurable, not just $\mathcal{F}_s$-measurable). And now that you've got this thing, which is a bounded martingale, you can apply the martingale identity to replace t with s; we definitely have $\mathcal{F}_{T_n \wedge s} \subseteq \mathcal{F}_s$, so we can turn this into

$$\mathbb{E}[D_s^{T_n}(M_s^{T_n} - \langle M, L\rangle_s^{T_n})\mathbf{1}_A].$$

And then we can write this as

$$\mathbb{E}_{\mathbb{Q}}[(M_s^{T_n} - \langle M, L\rangle_s^{T_n})\mathbf{1}_A].$$

So the message is that the same computation as before works, but only if you restrict $A \in \mathcal{F}_{T_n \wedge s}$.

What does this give you? It gives you some identity of conditional expectations, but with respect to the σ-algebra $\mathcal{F}_{T_n \wedge s}$ instead — the outcome is that

$$\mathbb{E}_{\mathbb{Q}}[(M_t^{T_n} - \langle M, L\rangle_t^{T_n}) \mid \mathcal{F}_{T_n \wedge s}] = M_s^{T_n} - \langle M, L\rangle_s^{T_n}.$$

But basically this allows you to conclude. Why? Now if I take a conditional expectation

$$\mathbb{E}_{\mathbb{Q}}[M_t^{T_n} - \langle M, L\rangle_t^{T_n} \mid \mathcal{F}_s],$$

what I really want to show is that this thing on the inside is the martingale, so I want to show this thing is equal to the stopped process at s. But what you do now is just split based on whether the stopping time is above or below s — for simplicity, we'll write $M_t^{T_n} - \langle M, L\rangle_t^{T_n} = X_t^{T_n}$. Then we have

$$\mathbb{E}_{\mathbb{Q}}[X_t^{T_n} \mid \mathcal{F}_s] = \mathbb{E}_{\mathbb{Q}}[\mathbf{1}(T_n \ge s)X_t^{T_n} \mid \mathcal{F}_s] + \mathbb{E}_{\mathbb{Q}}[\mathbf{1}(T_n < s)X_t^{T_n} \mid \mathcal{F}_s].$$

Then you notice that the second term is kind of a trivial conditional expectation — if $T_n < s$, then $X_t^{T_n}$ is just equal to $X_s^{T_n}$. And then the conditional expectation doesn't do anything (because the thing is already $\mathcal{F}_s$-measurable); so we can write

$$\mathbb{E}_{\mathbb{Q}}[\mathbf{1}(T_n < s)X_t^{T_n} \mid \mathcal{F}_s] = X_s^{T_n}\mathbf{1}(T_n < s).$$

For the first term, we want to use the identity we just arrived at; so we want to say that it's equal to

$$\mathbb{E}_{\mathbb{Q}}[\mathbf{1}(T_n > s)X_t^{T_n} \mid \mathcal{F}_{T_n \wedge s}].$$

At least intuitively, why is this believable? On the event $T_n > s$, the $T_n \wedge s$ becomes s, which is what you're conditioning on originally. So that's the heuristic, but you actually have to verify why it's true. We skipped

it in the notes, but you can try verifying it yourself; it's a good exercise (you just have to play with the definition of conditional expectation).

But now that you can actually have this, we can pull out $\mathbf{1}[T_n > s]$ (since it's measurable with respect to the σ-algebra we're conditioning on) to get

$$\mathbf{1}(T_n > s)\mathbb{E}_{\mathbb{Q}}[X_t^{T_n} \mid \mathcal{F}_{T_n \wedge s}] + X_s^{T_n}\mathbf{1}[T_n < s].$$

And now we can use the identity we just proved to get that this is

$$\mathbf{1}(T_n \geq s)X_s^{T_n} + X_s^{T_n}\mathbf{1}(T_n < s) = X_s^{T_n}.$$

And now if we plug in what X was, we see that $(X_t^{T_n}) = (M_t^{T_n} - \langle M, L\rangle_t^{T_n})$ is a $\mathbb{Q}$-martingale.

And once we know the stopped process is a martingale, the unstopped process $M - \langle M, L\rangle$ is a CLMG. □

That's it for Girsanov. The main takeaway is if you ignore this technical stuff about integrability, it's just Ito's formula and realizing the two FV terms cancel each other — to show something's a martingale, you compute its evolution using Ito and show the FV terms cancel. That makes you believe things should work, but in the general case you then need this localization argument, which you have to think through.

Remark 19.7. One comment is that it might have been better to organize the proof this way, to make things more clear: What we just showed is that if you have some process X times this likelihood ratio martingale D, then if the stopped process $(XD)^T$ is a $\mathbb{P}$-martingale, then the stopped process X^T by itself is a $\mathbb{Q}$-martingale. That's basically what we showed, for this specific case of $X = M - \langle M, L\rangle$. It's somewhat believable because D is what lets you convert expectations iwth respect to $\mathbb{Q}$ into ones with respect to $\mathbb{P}$.

Next time we'll continue talking about Girsanov's theorem.

Remark 19.8. One thing Sky wanted to mention at the beginning of class, a remark on the martingale representation theorem — last time, there was a question about how this theorem doesn't contradict the kind of process where you have a martingale that starts at 0, is constant until time 1, and then you flip a coin and if it's heads it becomes 1 from then on forever, and if tails it becomes -1? This is a martingale; let's call it M_t. And it's kind of clear this cannot have a continuous modification — there's no way to resolve the jump here.

But we said that every martingale has a continuous modification...

The point is there's no way you can write this as $M_t = \mathbb{E}[f(B) \mid \mathcal{F}_t]$. Any martingale that's bounded and adapted to your Brownian filtration, you can write as a function of your entire Brownian path B, conditioned on $\mathcal{F}_t$. And the point is there's no way to find a function f such that you get a process that looks like this. So the point is that it's impossible to realize such a martingale as being adapted to the filtration of your BM.

§20 April 16, 2025

§20.1 Some consequences of Girsanov

Let's quickly discuss some immediate consequences of Girsanov's theorem.

Recall that in the setting of Girsanov, you define a measure $\mathbb{Q}$ with respect to $\mathbb{P}$, and it's basically coming out of some sort of exponential martingale of a martingale L, i.e.,

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = \mathcal{E}(L)_\infty$$

(we assume $\mathbb{Q}$ and $\mathbb{P}$ are mutually absolutely continuous, so there are some restrictions on L).

Under this change of measure, Girsanov's theorem basically tells you how martingales change. It tells you the drift a martingale picks up when you change the measure, because it tells you what you have to subtract to get a martingale again. Specifically, a CLMG under $\mathbb{P}$ is a CSMG under $\mathbb{Q}$ — the reason you know that is Girsanov gives you teh explicit decomposition

$$M = (M - \langle M, L\rangle) + \langle M, L\rangle.$$

We mentioned last time that the concept of a FV process doesn't depend on the underlying probability measure, so $\langle M, L\rangle$ is still FV; and Girsanov says that $M - \langle M, L\rangle$ is a CLMG under $\mathbb{Q}$.

It also follows that CSMGs under $\mathbb{P}$ are also CSMGs under $\mathbb{Q}$ (again because the concept of FV processes doesn't depend on the probability measure), and vice versa. So the notion of CSMG doesn't depend on your probability measure.

That's one comment. The second is:

Fact 20.1 — Let X and Y be CSMGs under $\mathbb{P}$ (and therefore also $\mathbb{Q}$). Then actually, $\langle X, Y\rangle$ is the same under $\mathbb{P}$ and $\mathbb{Q}$.

The reason for this is that you could have computed the probability as a limit in probability of discrete approximations, as

$$\langle X, Y\rangle_t = \lim \sum_{i=1}^{p_n} (X_{t_i^n} - X_{t_{i-1}^n})(Y_{t_i^n} - Y_{t_{i-1}^n})$$

(where the limit is in probability). And since $\mathbb{P}$ and $\mathbb{Q}$ are mutually absolutely continuous, convergence in probability under $\mathbb{P}$ is equivalent to convergence in probability under $\mathbb{Q}$. It's not immediately obvious why this is true, because absolute continuity is a statement about measure-0 sets, whereas convergence in probability is a statement about probabilities *tending* to 0. But one way you could see the equivalence is to use that $X_n \xrightarrow{\text{prob}} X$ is equivalent to the statement that for all subsequences X_{n_k}, there exists a further subsequence $X_{n_{k_j}}$ such that $X_{n_{k_j}} \xrightarrow{\text{a.s.}} X$. (This is a fact we won't prove.) So the point is that convergence in probability is equivalent to this statement about almost sure convergence; and *this* is about probability 0 or 1 events, so it's going to be the same under $\mathbb{P}$ and $\mathbb{Q}$.

The final comment is that if B is an $(\mathcal{F}_t)$-Brownian motion under $\mathbb{P}$, then $\widetilde{B} = B - \langle B, L\rangle$ is a CLMG under $\mathbb{Q}$, and moreover by the above fact (and the fact that the bracket doesn't care about the FV term $\langle B, L\rangle$), we have

$$\langle \widetilde{B}, \widetilde{B}\rangle_t = \langle B, B\rangle_t = t.$$

Then Lévy's characterization implies that $\widetilde{B}$ is an $(\mathcal{F}_t)$-Brownian motion under $\mathbb{Q}$.

§20.2 Conditions for uniform integrability of $\mathcal{E}(L)$

Now we'll talk about something a bit technical. We want to talk about conditions which allow you to apply Girsanov's theorem. In Girsanov's theorem, the assumption is that $\mathbb{Q}$ and $\mathbb{P}$ are mutually absolutely continuous. What you really need is that $\mathcal{E}(L)$ is uniformly integrable. So you want to understand conditions on L for which $\mathcal{E}(L)$ is actually a uniformly integrable martingale — if you know $\mathcal{E}(L)$ is a UI martingale, that'll imply $\mathbb{Q}$ and $\mathbb{P}$ are mutually absolutely continuous (one needs to prove this, but we'll skip that).

Let's first suppose that L is a CLMG with $L_0 = 0$ (at least, almost surely), and $\langle L, L\rangle_\infty < \infty$ almost surely. This latter condition implies that L_∞ exists almost surely. This was probably on the homework one or two weeks ago. In a couple of sentences, why would this be true? First, if $\langle L, L\rangle_\infty$ is integrable, then you have a L^2-bounded martingale; then it's UI integrable, and you can apply the convergence theorem. But if this is finite almost surely, you can do some truncation argument to basically pretend it's bounded — you take

a stopping time which is the first time this goes above some level. Then you're bounded. And because it's finite almost surely, the limit of these stopping times goes to ∞.

The point of all this is that under these conditions, $\mathcal{E}(L)_\infty$ is well-defined (since it's defined in terms of L and its QV; and if both of these things exist at time ∞, then you can talk about $\mathcal{E}(L)_\infty$).

So in other words, then $\mathcal{E}(L)$ is a nonnegative CLMG which converges almost surely to the corresponding thing

$$\mathcal{E}(L)_\infty = \exp\left(L_\infty - \frac{1}{2}\langle L, L\rangle_\infty\right).$$

And in particular, by Fatou, this implies that

$$\mathbb{E}[\mathcal{E}(L)_\infty] \leq 1.$$

We might also want to assume here that $\mathcal{E}(L)$ is a true martingale. At least if you assume that, then this is true (by Fatou's). In principle, you want $\mathbb{E}[\mathcal{E}(L)_t] \leq 1$ for all t to conclude by Fatou that inequality. And it will be exactly equal to 1 if this thing is a martingale, because $L_0 = 0$.

And now the claim is that actually if this is an equality, then this martingale is a UI martingale.

Claim 20.2 — If $\mathbb{E}[\mathcal{E}(L)_\infty] = 1$, then $\mathcal{E}(L)$ is a UI martingale.

Proof. Intuitively, why that would be true is the fact it might be strictly less than 1 is kind of saying you're losing mass as you go off to ∞. The classic example of something that's L^1-bounded but not UI is where you take a sequence of thinner and thinner functions (rectangular boxes) where the width is $\frac{1}{n}$ and the height is n. Then the almost sure limit is going to be 0, but for every finite n it has L^1-norm 1. So you're losing mass to ∞. The point of this condition that it's exactly equal to 1 is more or less ruling out this scenario.

To see why this implies uniform integrability, you can first apply conditional Fatou to say that

$$\mathbb{E}[\mathcal{E}(L)_\infty \mid \mathcal{F}_t] \leq \mathcal{E}(L)_t.$$

But if you now combine this with $\mathbb{E}[\mathcal{E}(L)_\infty] = 1$, you obtain that this is actually an equality almost surely — i.e.,

$$\mathbb{E}[\mathcal{E}(L)_\infty \mid \mathcal{F}_t] = \mathcal{E}(L)_t.$$

The point is that if the expectation of the LHS is 1, well, we also know $\mathbb{E}[\mathcal{E}(L)_t] = 1$ (since we're assuming this thing is a martingale, so the expectation at every finite time is 1). But if you have this inequality, then there's no way for it to be strict on some positive-probability event (then there's no way their expectations would be equal).

So combining this with the assumption, you get that actually $\mathbb{E}[\mathcal{E}(L)_\infty \mid \mathcal{F}_t] = \mathcal{E}(L)_t$; and that gives you a UI martingale. □

We want to give conditions under which $\mathbb{E}[\mathcal{E}(L)_\infty] = 1$, and that's going to imply we get a UI martingale. Now if we define this new measure $\mathbb{Q}$ in that way, we get a situation where we can apply Girsanov's theorem.

Theorem 20.3

Let L be a CLMG such that $L_0 = 0$. Consider the following three properties:

(i) (Novikov's criterion) $\mathbb{E}[\exp(\frac{1}{2}\langle L, L\rangle_\infty)] < \infty$.

(ii) (Kazamaki's criterion) L is a UI martingale and $\mathbb{E}[\exp(\frac{1}{2}L_\infty)] < \infty$.

(iii) $\mathcal{E}(L)$ is a UI martingale.

Then $(i) \implies (ii) \implies (iii)$.

So this gives you criteria under which you get that $\mathcal{E}(L)$ is a UI martingale. In practice, (i) is the one we're going to see most often, because $\langle L, L\rangle_\infty$ is probably actually going to be bounded for us, so (i) will be trivial to verify.

The proof is kind of playing around with technical things.

Proof of $(i) \implies (ii)$. The first thing to notice is that (i) implies

$$\mathbb{E}[\langle L, L\rangle_\infty] < \infty$$

(the assumption on the exponential is much stronger). Once you have this, you know at least that L is a UI martingale (actually it's L^2-bounded, which is even stronger). And then what you now say is, we're trying to bound $\mathbb{E}[\exp(\frac{1}{2}L_\infty)]$. So you take this and kind of want to introduce the thing in (i), which you have control of; you can write

$$\mathbb{E}\left[\exp\left(\frac{1}{2}L_\infty\right)\right] = \mathbb{E}\left[\mathcal{E}(L)_\infty^{1/2} \cdot \exp\left(\frac{1}{4}\langle L, L\rangle_\infty\right)\right].$$

And now you can use Cauchy–Schwarz to bound this by

$$\mathbb{E}[\mathcal{E}(L)_\infty]^{1/2}\mathbb{E}\left[\exp\left(\frac{1}{2}\langle L, L\rangle_\infty\right)\right]^{1/2}.$$

But we just showed the first thing is at most 1, and the second part is finite by assumption. □

Remark 20.4. To justify $\mathbb{E}[\mathcal{E}(L)_\infty] \le 1$, we might need to justify why this condition implies $\mathcal{E}(L)$ is a true martingale.

But actually we can get away without assuming $\mathcal{E}(L)$ is a true martingale — there's a result that a nonnegative CLMG is actually a supermartingale (basically the proof is actually Fatou's), so you will have $\mathbb{E}[\mathcal{E}(L)_t] \le 1$ because it's a supermartingale.

So this part wasn't too bad; really the work is in showing that (ii) implies (iii). This is where you have to be quite clever.

Proof of $(ii) \implies (iii)$. First, since L is a UI martingale, by the optional stopping theorem you have that for all stopping times T, we have

$$L_T = \mathbb{E}[L_\infty \mid \mathcal{F}_T].$$

And then you use conditional Jensen's to bound

$$\exp\left(\frac{1}{2}L_T\right) \le \mathbb{E}\left[\exp\left(\frac{1}{2}L_\infty\right) \mid \mathcal{F}_T\right]$$

(L_T is itself a conditional expectation; Jensen's inequality lets you put a convex function inside the expectation, and conditional Jensen's lets you put it inside the conditional expectation).

Now, your assumption is that $\mathbb{E}[\exp(\frac{1}{2}L_\infty)] < \infty$. Something that was mentioned at some point when we talked about UI martingales is that if you have an integrable function, then the collection of all random variables formed by taking conditional expectations with all sub-σ-algebras is a UI collection — so the collection $(\mathbb{E}[\exp(\frac{1}{2}L_\infty) \mid \mathcal{F}_T])$ is UI (where this collection is indexed by stopping times T).

Let's also quickly write the definition of uniform integrability for a collection, since we're going to use it later:

Definition 20.5. For an arbitrary collection of random variables $(X_\lambda)_{\lambda\in\Lambda}$, we say it's uniformly integrable if for all $\varepsilon > 0$, there exists $\delta > 0$ such that for any event $\mathbb{P}[A] < \delta$, we have

$$\sup_{\lambda\in\Lambda} \mathbb{E}[|X_\lambda| \mathbf{1}_A] \le \varepsilon.$$

This is some sort of uniform control over the tails of all your random variables — the key point is you can find some condition so that as long as you satisfy $\mathbb{P}[A] < \delta$, it works for *all* random variables in your collection. This is quite a stringent condition.

But we have that this collection is UI, and that $\exp(\frac{1}{2}L_T)$ is bounded by a random variable in this collection, and it's also nonnegative. So if you look at the definition of uniform integrability, you also get that the collection of random variables $(\exp(\frac{1}{2}L_T))$ (again indexed by stopping times T) is UI.

That's the first thing to note that we'll use.

And now, you want to introduce an auxiliary parameter $0 < a < 1$. You should think of it as close to 1 — we'll eventually take $a \to 1$ — but the purpose of this is to give you some room in applying Hölder, as we will see.

So fixing this parameter a, we define

$$Z_t^{(a)} = \exp\left(\frac{a}{1+a} \cdot L_t\right).$$

As a comment, since $a < 1$, this $\frac{a}{1+a}$ is always strictly less than $\frac{1}{2}$, which means we'll be able to bound the expectation of this by the expectation of $\exp(\frac{1}{2}L_t)$, by Hölder. In particular, these random variables are integrable; and they're also in L^p for some $p > 1$ which depends on a.

One thing you realize is that you can write

$$\mathcal{E}(aL)_t = \exp\left(aL_t - \frac{a^2}{2}\langle L, L\rangle_t\right)$$

(so instead of $\mathcal{E}(L)$, we have $\mathcal{E}(aL)$), by definition. And it turns out that you can write this as

$$\mathcal{E}(aL)_t = \mathcal{E}(L)_t^{a^2} \cdot (Z_t^{(a)})^{1-a^2}.$$

This is some algebra — the point is that the a^2 in $\mathcal{E}(L)_t$ takes care of the QV term, so you just have to say why when you combine the appropriate power of $Z_t^{(a)}$ plus $a^2 L$, you're going to get aL. The identity you use is basically

$$a^2 + (1-a^2) \cdot \frac{a}{1+a} = a^2 + (1-a)a = a.$$

So that's one thing to note. Next, what we're going to show is that, actually this is going to define for us a UI collection of random variables:

Claim 20.6 — The collection of random variables $(\mathcal{E}(aL)_T)$, indexed by stopping times T, is uniformly integrable.

Proof. Well, we have this identity for $\mathcal{E}(aL)_t$, so somehow we want to bound this in terms of our two terms. And because $\frac{a}{1+a} < \frac{1}{2}$, we'll be able to use control of $\exp(\frac{1}{2}L)$ to control Z; and the $\mathcal{E}(L)_t^{a^2}$ term will be fine, because its expectation is going to be at most 1 for similar reasons to before.

We're going to try to directly verify the condition of UI. So let $\Gamma \in \mathcal{F}$ (this plays the role of the event A from the event). Our goal is to estimate

$$\mathbb{E}\left[\mathbf{1}_\Gamma \mathcal{E}(aL)_T\right].$$

Now you use the identity for $\mathcal{E}(aL)_t$ combined with Hölder to estimate this by

$$\mathbb{E}\left[\mathcal{E}(L)_T\right]^{a^2}\mathbb{E}\left[\mathbf{1}_\Gamma Z_T^{(a)}\right]^{1-a^2}.$$

Basically, what you did here was you used the fact that $\mathbb{E}[XY] \le \mathbb{E}[X^p]^{1/p}\mathbb{E}[Y^q]^{1/q}$ where $1/p + 1/q = 1$. Then you just need to choose the right p and q. You want to get exactly one power of $\mathcal{E}(L)_T$ in the expectation, while in the bound you have a^2. So you want to choose p such that $pa^2 = 1$, i.e., $p = 1/a^2$. That imposes $1/q = 1 - a^2$, so in other words $q = 1/(1-a^2)$. Then you see that when you raise $(Z_t^{(a)})^{1-a^2}$ to the qth power, this also becomes 1.

And $\mathcal{E}(L)$ is a nonnegative supermartingale, so $\mathbb{E}[\mathcal{E}(L)_T] \le 1$; and you're just left with the second factor, which is at most

$$\mathbb{E}[\mathbf{1}_\Gamma Z_T^{(a)}]^{1-a^2}.$$

Now we're going to use Hölder again to bound this by

$$\mathbb{E}\left[\mathbf{1}_\Gamma \exp\left(\frac{1}{2}L_T\right)^{(1-a^2)\cdot\frac{2a}{1+a}}\right].$$

Where does this come from? Here we used the fact that $\mathbb{E}[X] \le (\mathbb{E}X^p)^{1/p}$, and you just want to again choose the right p. And since $a < 1$, we have $\frac{a}{1+a} < \frac{1}{2}$. So you want $p \cdot \frac{a}{1+a} = \frac{1}{2}$ (since the thing with $\frac{1}{2}$ is what you control — we said the thing with $\frac{1}{2}$ is UI). And we have this equation for p; that means we should have $p = \frac{1+a}{2a}$. And thus when we take Z^p, we're goin gto get this $\frac{1}{2}$ by construction, and when we take $1/p$ on the outside, we get $\frac{2a}{1+a}$.

But now we use the fact that the collection $\exp(\frac{1}{2}L_T)$ is UI; that's going to imply our claim, because you just look at the definition of uniform integrability. We can bound $\mathbb{E}[\mathbf{1}_\Gamma\mathcal{E}(aL)_T]$ by this, and if this is UI then we're done. □

Student Question. *Where did we use $a < 1$?*

Answer. The first step seems fishy if $a = 1$ — specifically, where we bound

$$\mathbb{E}[\mathbf{1}_\Gamma\mathcal{E}(aL)_T] \le \mathbb{E}[\mathcal{E}(L)_T]^{a^2}\mathbb{E}[\mathbf{1}_\Gamma Z_T^{(a)}]^{1-a^2}.$$

The problem is here you need $p = 1$ and $q = \infty$; then you get the L^∞ norm of Z, and we don't control that. We control the expectation of Z to some power, but we don't know that Z is actually bounded.

Now we're almost done, but there's still some work left to do. Now we know this collection $(\mathcal{E}(aL)_T)_{T \text{ a stopping time}}$ is UI. Finally, since $\mathcal{E}(aL)$ is a CLMG — basically, we want to show that $\mathcal{E}(aL)$ is UI (this is what we're working towards). You can then believe that if we know it's UI for all $a < 1$, by some limiting procedure as $a \to 1$, you would get the uniform integrability of the thing we actually want, i.e., $\mathcal{E}(L)$.

Why is this going to be a UI martingale? Well, we just showed it's UI, so we just need to show it's a martingale. But a CLMG which is UI is not necessarily a martingale, so we have to be careful.

At least, since it's a CLMG, there is a sequence of stopping times (T_n) such that the localized versions

$$(\mathcal{E}(aL)_{t\wedge T_n})_t$$

are UI martingales. And then for $s \le t$, I want to verify the martingale identity. So if I take the conditional expectaiton at time t given $\mathcal{F}_s$, i.e.,

$$\mathbb{E}[\mathcal{E}(aL)_t \mid \mathcal{F}_s],$$

the value at time t is the pointwise limit of these stopped processes as $n \to \infty$, so this is

$$\mathbb{E}\left[\lim_{n\to\infty} \mathcal{E}(aL)_{t\wedge T_n} \mid \mathcal{F}_s\right].$$

And now you use the fact that if you fix t and consider this as a collection of random variables in n, it's UI (because of the above claim). And because it's UI, you can commute the expectation and limit (this is some sort of conditional version of a limit theorem), so this is

$$\lim_{n\to\infty} \mathbb{E}[\mathcal{E}(aL)_{t\wedge T_n} \mid \mathcal{F}_s].$$

This is very important — it's the step that doesn't hold in general for a UI CLMG. The fact you have a CLMG tells you this $(\mathcal{E}(aL)_{t\wedge T_n})_t$ is UI in t for a fixed n; but even if you assume the original process is UI, you don't know that if you fix t, the thing is UI in n — that's not true in general, but for us it is true, because we were able to show it.

And because $\mathcal{E}(aL)$ is a UI martingale, this is equal to

$$\lim_{n\to\infty} \mathcal{E}(aL)_{s\wedge T_n} = \mathcal{E}(aL)_s.$$

So we've shown that $\mathcal{E}(aL)$ is a martingale.

Now that you have this martingale identity, you might try to conclude by taking limits in a. That's not what we do (you probably run into some issue). Instead, you recall that it suffices just to lower-bound the expectation of the thing at time ∞ by 1 — to show that $\mathcal{E}(L)$ is a UI martingale, it suffices to show that $\mathbb{E}[\mathcal{E}(L)_\infty] \geq 1$, because we basically have the argument we just erased.

So we're going to try to lower-bound this expectation by 1, and here is where you'll be able to take limits in a. We have a UI martingale for $a < 1$, which means

$$\mathbb{E}[\mathcal{E}(aL)_\infty] = \mathcal{E}(aL)_0 = 1.$$

And then you again use this identity

$$\mathcal{E}(aL)_t = \mathcal{E}(L)_t^{a^2} (Z_t^{(a)})^{1-a^2},$$

now at time ∞, to write this as a product

$$1 = \mathbb{E}[\mathcal{E}(aL)_\infty] = \mathbb{E}\left[\mathcal{E}(L)_\infty^{a^2} (Z_\infty^{(a)})^{1-a^2}\right].$$

(We're just inserting our previous identity.) And then you just do the same thing where you apply Hölder in exactly the same way, which should give that this is at most

$$\mathbb{E}[\mathcal{E}(L)_\infty]^{1/a^2} \mathbb{E}[Z_\infty^{(a)}]^{1-a^2}.$$

This is true for all $0 < a < 1$. And now you take the limit. You want to claim that $\mathbb{E}[Z_\infty^{(a)}]^{1-a^2} \to 1$ as $a \to 1$. Why is that going to be true? Well, we need one more step of Hölder here to replace this expectation by $\mathbb{E}[\exp(\frac{1}{2}L)]$, precisely as we did before, because that's the thing we control; so applying Hölder again, we get that this is at most

$$\mathbb{E}[\mathcal{E}(L)_\infty]^{1/a^2} \mathbb{E}\left[\exp\left(\frac{1}{2}L_\infty\right)\right]^{(1-a^2)\cdot\frac{2a}{1+a}}.$$

Now the good thing is that $\mathbb{E}[\exp(\frac{1}{2}L_\infty)]$ doesn't depend on a; the only dependence on a is in the exponent. And it's finite by assumption. So now we can send $a \uparrow 1$, and we get $\mathbb{E}[\mathcal{E}(L)_\infty] \geq 1$, which is precisely what we wanted. □

Student Question. *Where does the first equality*

$$\mathbb{E}[\mathcal{E}(aL)_t \mid \mathcal{F}_s] = \mathbb{E}\left[\lim_{n\to\infty} \mathcal{E}(aL)_{t\wedge T_n} \mid \mathcal{F}_s\right]$$

come from?

Answer. It's just true pointwise, even before you take the conditional expectations, since $T_n \uparrow \infty$. The line *after* that uses the uniform integrability of the collection $(\mathcal{E}(aL)_{t\wedge T_n})_n$, which is the thing that you had to prove.

§20.3 Applications of Girsanov

Now we can talk about more applications of Girsanov.

Suppose b is a bounded measurable function of time and space — i.e., on $\mathbb{R}_+ \times \mathbb{R}$ (where $\mathbb{R}_+$ is your time variable, and $\mathbb{R}$ your space variable). Assume $g \in L^2(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+), dt)$ is L^2 with respect to the usual Lebesgue measure, such that you have a pointwise bound

$$|b(t,x)| \le g(t).$$

Example 20.7

As a model scenario which satisfies this assumption, for instance if b is bounded on some finite intervla of time $[0,A] \times \mathbb{R}$ and 0 outside this interval — so for times before A it's bounded on this space, and after that it's 0. Then you can make g also be 0 after time A, and it's also going to be bounded; so you get a L^2 function that way. So that's one scenario where you'll have this assumption be satisfied.

Now let B be an $(\mathcal{F}_t)$-Brownian motion. We'll define a process

$$L_t = \int_0^t h(s, B_s)\, dB_s.$$

In principle, one has to check that $b(s, B_s) \in L^2(B)$. The fact that it's going to be in $L^2(B)$ should just come from the assumptions on b and g — we have to bound something like

$$\mathbb{E}\left[\int_0^\infty |b(s,B_s)|^2\, ds\right].$$

But you have this pointwise bound by g, which is actually a (deterministic) L^2-function; so this is bounded by

$$\int_0^\infty g(s)^2\, ds < \infty.$$

So the norm bound is fine. But you also have to check that it's progressive, because $L^2(B)$ is the set of progressive processes satisfying this bound. This we're just going to skip (usually with these measurability things, it should be true — at least if b is continuous you'd definitely believe it, because any continuous process which is adapted is also progressive, so at least if the thing is continuous it should be fine; and if b is just a general measurable function, maybe you just approximate it by continuous functions or something). So we're not going to check this.

Then we can look at the associated exponential martingale

$$\mathcal{E}(L)_t = \exp\left(\int_0^t b(s,B_s)\, dB_s - \frac{1}{2}\int_0^t b(s,B_s)^2\, ds\right)$$

(the first term is L, and the second is $\langle L, L\rangle$).

We already estimated $\int b(s, B_s)^2$ by $\int g(s)^2$, and the estimate is true even before you take expectations (because it comes from the pointwise bound). So the first condition in Theorem 20.3 is true — we have

$$\langle L, L\rangle_\infty = \int_0^\infty b(s, B_s)^2\, ds \le \int_0^\infty g(s)^2\, ds < \infty.$$

So this is actually a bounded random variable, which means (i) is satisfied.

So Theorem 20.3 says that $\mathcal{E}(L)$ is a UI martingale (which we'll call D). Now we can apply Girsanov's theorem — let a new measure $\mathbb{Q}$ be defined by

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = D_\infty = \mathcal{E}(L)_\infty.$$

By Girsanov — we started with a Brownian motion B, so if we apply Girsanov's theorem with the local martingale M being B, we fet that

$$B - \langle B, L\rangle$$

is an $(\mathcal{F}_t)$-Brownian motion under $\mathbb{Q}$. This is combining Girsanov with the third consequence mentioned in class today — this is a CLMG and its QV is the same as that of BM, so it's actually a BM by Levy.

Moreover you can compute this bracket, because L is actually just a stochastic integral against B — so we have

$$B - \langle B, L\rangle = B - \int_0^t b(s, B_s)\, ds.$$

So let's define this process

$$\beta_t = B_t - \int_0^t b(s, B_s)\, ds.$$

Then under $\mathbb{Q}$, (β_t) is a Brownian motion.

You can put this another way — under $\mathbb{Q}$, there exists an $(\mathcal{F}_t)$-Brownian motion (β_t) such that the process $X = B$ (this is more for notational reasons) satisfies the SDE

$$dX_t = b(t, X_t)\, dt + d\beta_t.$$

Why does X satisfy this SDE? Well, recall that SDEs are just integral equations; written in integral form, we're supposed to have

$$X_t = \int_0^t b(s, X_s)\, ds + \beta_t.$$

But then you just look at the definition of β — β is your BM minus precisely this first part, because we defined $X = B$. So by this definition of β, you have this identity, which you could write in SDE notation like this.

The whole point of all this, the thing to be emphasized, is you don't need any assumptions on b in terms of regularity. For instance, you don't need to assume b is Lipschitz in its variables. So the consequence of this application is that if you want to solve an SDE of this form, at the very least you can find a probability space and a BM and a process X which solves your SDE. And again, you don't need to assume b is Lipschitz or has any regularity properties. This is in sharp contrast to usual ODE properties — you might recall that usually to talk about existence for ODEs, b has to somehow depend continuously on your parameters, because you usually set up some fixed-point contraction mapping argument (you view an ODE as an integral equation and want to say the RHS defines a contracting map, and then you apply fixed point theorems to conclude; but in order for it to be a contraction map, you usually need b to have regularity). Girsanov gives you a different way to construct solutions to SDEs (which are like stochastic ODEs) using probability theory (change of measure) instead of more analytic considerations.

In the last week, we'll hopefully come back to how you could construct a solution to this SDE using fixed-point arguments. Maybe I give you a Brownian motion and want a solution to this SDE with this precise BM; that's more restrictive, and then you do need a fixed-point argument. This is more flexible — it's saying I can find a BM on *some* probability space and a process X satisfying the SDE, so your notion of 'solution' is looser.

§20.4 Cameron–Martin formula

Now let's talk about a special case of this example, called the Cameron–Martin formula.

Suppose now that actually this function b is just $b(t,x) = g(t)$ — so your function doesn't depend on space now, just time. And as before, suppose $g \in L^2(\mathbb{R}_+)$. For $t \geq 0$, let

$$h(t) = \int_0^t g(s)\,ds$$

(this is the $\int_0^t b(s, B_s)\,ds$ term from before, coming from $\langle B, L\rangle$).

Definition 20.8. The set of all functions h of this form is known as the Cameron–Martin space, and denoted by $\mathcal{H}$.

Remark 20.9. You're basically saying that for any such function, its derivative is in L^2 (since $\dot{h} = g$). So basically, the Cameron–Martin space is a sort of L^2-based Sobolev space of order 1, but restricted to the set of functions starting at 0. This is a side comment, but if you've taken analysis before you might have seen Sobolev spaces, and that's what this is — functions that have one derivative in L^2 (and start at 0).

From this previous discussion, we can define a new probability measure $\mathbb{Q}$; in this particular case, it will be given by

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = \exp\left(\int_0^\infty g(t)\,dB_t - \frac{1}{2}\int_0^\infty g(t)^2\,dt\right) = D_\infty.$$

And then as a special case of the previous result, under $\mathbb{Q}$, the process

$$\beta_t = b_t - h(t)$$

is a Brownian motion.

What's a consequence? If you think of β as a random variable taking values in $\mathcal{C}[0,\infty]$ (the space of continuous functions), you can think about measurable functions of this random variable — $\Phi : (\mathcal{C}([0,\infty)), \mathcal{C}) \to \mathbb{R}^+$, where $\mathcal{C}$ is the smallest σ-algebra making all your coordinate functions measurable (from the first homework; don't worry about it too much). The point is I'm viewing β as a single random variable, rather than a stochastic process.

And then we're saying if I compute

$$\mathbb{E}[D_\infty \Phi(B)]$$

(so I'm taking some expectation of some function of my Brownian path), the fact that I have D_∞ here means that this is

$$\mathbb{E}_{\mathbb{Q}}[\Phi(B)]$$

(by the way that $\mathbb{Q}$ is defined). But under $\mathbb{Q}$, I can write this as

$$\mathbb{E}_{\mathbb{Q}}[\Phi(\beta + h)]$$

(because that's what B is). And the claim was that under $\mathbb{Q}$, β is a BM. So it has the same law as B under $\mathbb{P}$. So I could actually have written this as

$$\mathbb{E}[\Phi(B+h)],$$

because the law of β under $\mathbb{Q}$ is the same as that of B under $\mathbb{P}$ (they're both Brownian motions).

That leads to the following result. Basically, a consequence of this is that the law of a shifted BM is absolutely continuous with the law of a BM. Let's write this a bit more precisely.

Theorem 20.10 (Cameron–Martin formula)

Let $W(dw)$ be the law of Brownian motion, i.e., the Wiener measure on $C([0,\infty),\mathcal{C})$. Let $h \in \mathcal{H}$ (this means it starts at 0 and its derivative is in L^2). Then for any nonnegative measurable Φ, we have

$$\int W(dw)\Phi(w+h) = \int W(dw)\Phi(w) \cdot \exp\left(\int_0^\infty h(t)dw(t) - \frac{1}{2}\int_0^\infty h(t)^2\,dt\right).$$

Here w is a continuous function; so the LHS is actually another way of writing $\mathbb{E}[\Phi(B+h)]$ (I'm just rewriting this identity in some other way). And for the LHS, again this is an expectation with respect to B, so I can write it as an integral with respect to this Wiener measure. There's a $\Phi(w)$, because we had $\Phi(B)$; and then D_∞ is what we wrote up there.

Note that $\int_0^\infty h(t)\,dw(t)$ is defined for W-almost every w (it's an Ito integral, defined by some Ito isometry; so it's defined for almost every w, but that's okay for this formula).

Another way to say this is that if I think about the law of this shifted BM, I can reduce any expectation of the shifted BM to an expectation with respect to the original. You might think that's useful because we understand BM very well, perhaps better than if you dd in some shift. So if you let W_h be the law of $B+h$, this measure W_h is again going to be a probability measure on $(\mathcal{C}([0,\infty)),\mathcal{C})$. And to integrate functions against W_h, you have

$$\int W_h(dw)\Phi(w) = \int W(dw)\Phi(w+h).$$

(This is some abstract thing about pushforwards.) But the point is that W_h is absolutely continuous with respect to W, and its RN derivative or density is precisely this extra thing

$$\frac{dW_h(w)}{dW(w)} = \exp\left(\int_0^\infty h(t)\,dw(t) - \frac{1}{2}\int_0^\infty \dot{h}(t)^2\,dt\right).$$

(This is just another way of writing the theorem, since the LHS of the theorem is $\int W_h(dw)\Phi(w)$.)

So this says the law of shifted BM is absolutely continuous with respect to the law of BM, and its RN derivative is given by this exponential local martingale. You can imagine if you take $h(t) = ct$ (so it's a linear function), you're basically shifting BM by a line — so now you have a straight line plus a BM. Then $\dot{h}(t) = c$. (This is technically not in L^2, but you can do a version of this for finite time, so that's fine.) Then you're basically saying the law of $B_t + ct$ relates to the law of B_t by something like

$$\exp\left(\int_0^T c\,dB_t - \frac{1}{2}c^2 t\right).$$

But if I integrate against a constant in a stochastic integral, I just get the constant times B; so this becomes

$$\exp\left(cB_t - \frac{1}{2}c^2 t\right).$$

(The point is you can do this same thing for a finite interval, where your laws are on finite times and ∞ is replaced by a finite time; then you can describe the law of shifted BM as some density times the law of BM.)

Remark 20.11. Next week Monday is a holiday. Sky will be gone on Wednesday, so there will be a substitute. We'll start with Markov processes, because they're needed for the more interesting stuff we'll talk about in the final weeks of class.

§21 April 23, 2025

Sky is not here, so Jeonghyun Ahn is giving the lecture.

Today we'll start a new topoic, called Markov processes. This will be the main theme of the rest of our lecture; and in the end, we'll see some connections between Markov processes and SDEs.

§21.1 Markov processes — intuition

You can intuitively think of a Markov process as a continuous generalization of Markov chains. Let's recall the definition of a Markov chain:

Definition 21.1. A Markov chain on a finite state space S is given by a transition matrix $p : S \times S \to [0,1]$, where $p(x,y) = \mathbb{P}[X_1 = y \mid X_0 = x]$.

Here the states are discrete, and time is also discrete — we're going from time 0 to time 1, time 2 to time 3, and so on.

For Markov processes, we want to generalize this in two extents. First, we want to make the state space into some general measurable space. Second, we want to make time continuous. These two generalizations will define something called a *Markov process*. So defining a Markov process will be one goal of today's lecture.

§21.2 Infinite state space

First, we want to generalize the state space to some measurable space $(E, \mathcal{E})$. To define a transition kernel (as with finite Markov chains), the inutition is that it should look like

$$p(x, A) = \mathbb{P}[X_1 \in A \mid X_0 = x] \quad \text{for } x \in E \text{ and } A \in \mathcal{E}.$$

This will be the motivation for defining a *Markov transition kernel*, which basically serves the same role as the Markov transition matrix for finite chains.

Definition 21.2. A function $\mathcal{Q} : E \times \mathcal{E} \to [0,1]$ is called a Markov transition kernel if it has the following two properties:

(1) For every $x \in E$, the map $\mathcal{E} \to [0,1]$ given by $A \mapsto \mathcal{Q}(x, A)$ is a probability measure on $(E, \mathcal{E})$.

(2) For every $A \in \mathcal{E}$, the map $E \to [0,1]$ given by $x \mapsto \mathcal{Q}(x, A)$ is measurable with respect to $\mathcal{E}$.

So this is a function which takes a state and a measurable set. The first condition says that if you fix any state x, then you get a measure on E.

Later we'll see that $\mathcal{Q}$ will correspond to the above object p we introduced.

Remark 21.3. One can see that if E is countable, then the $\mathcal{Q}$ we defined here is just the same as the transition kernel for a Markov chain (the kind we know pretty well) — then we can just track $(\mathcal{Q}(x, \{y\}))_{x,y \in E}$ (we track the probabilities of going from state x to state y). These will determine $\mathcal{Q}$; and this is the same thing as the transition kernel of the Markov chain.

We want to make one more definition — we also want to be able to apply $\mathcal{Q}$ to a function.

Definition 21.4. Let $f : E \to \mathbb{R}$ be bounded and measurable. Then we define $\mathcal{Q}f$ by

$$(\mathcal{Q}f)(x) = \int_E f(y)\mathcal{Q}(x, dy).$$

When we write $\mathcal{Q}(x, dy)$, we mean we're integrating with respect to the measure $\mathcal{Q}(x, \bullet)$ (where you fix x, and plug in your set $A \in \mathcal{E}$ — this gives a probability measure).

Intuitively what this function means is, you just start at x, and observe the state after one step and take the expectation of f at that new state — so we think of $(\mathcal{Q}f)(x)$ as $\mathbb{E}[f(X_1) \mid X_0 = x]$. This is just intuition, because we haven't yet made sense of what X_t is; but this is what you should think of $\mathcal{Q}f$ as.

Fact 21.5 — The function $\mathcal{Q}f$ is also bounded and measurable.

Proof. To see that it's measurable, you can first check this for indicator functions $f = \mathbf{1}_A$ — then we have

$$(\mathcal{Q}\mathbf{1}_A)(x) = \mathcal{Q}(x, A)$$

(applying $\mathcal{Q}$ is the same as plugging in A into your transition kernel), which is measurable by definition.

And since we have measurability for indicator functions, you can use simple function approximation to argue that $\mathcal{Q}f$ is measurable in general (because a limit of measurable functions is measurable).

For boundedness, you can just observe that $\mathcal{Q}(x, \bullet)$ is a probability measure — so we have

$$|(\mathcal{Q}f)(x)| \le \int |f(y)|\, \mathcal{Q}(x, dy) \le \|f\|_\infty$$

(because $\mathcal{Q}(x, \bullet)$ is a probability measure). □

Remark 21.6. In the above proof, we just bounded the norm of $\mathcal{Q}f$ by the norm of f. This means $\mathcal{Q}$ is a contraction on the space of bounded measurable functions $E \to \mathbb{R}$ (with the sup norm), which we call $B(E)$.

§21.3 Continuous time

Next, we want to generalize *time* to be continuous. How are we going to do this? Instead of considering the transition from tiem 0 to time 1, we want to consider all possible transitions from time 0 to time t. The intuition here is considering a *family* of the Markov transition kernel $(\mathcal{Q}_t)_t$, where t denotes the time. And $\mathcal{Q}_t$ should intuitively mean

$$\mathcal{Q}_t(x, A) = \mathbb{P}[X_t \in A \mid X_0 = x].$$

If we have a family of transition kernels $(\mathcal{Q}_t)$, then we can define a Markov process from it.

So let's make this formal by specifying what properties these $\mathcal{Q}_t$'s have to satisfy.

Definition 21.7. A family of transition kernels $(\mathcal{Q}_t)_{t \ge 0}$ is called a transition semigroup if it satisfies the following three properties:

(1) $\mathcal{Q}_0(x, dy) = \delta_x(dy)$ (i.e., when $t = 0$, $\mathcal{Q}_0(x, \bullet)$ is just a point mass at x).

(2) (Chapman–Kolmogorov equations) For all $s, t \ge 0$ and $A \in \mathcal{E}$, we have

$$\mathcal{Q}_{s+t}(x, A) = \int_E \mathcal{Q}_s(y, A)\mathcal{Q}_t(x, dy).$$

(3) (Measurability) For all $A \in \mathcal{E}$, the map $(t, x) \mapsto Q_t(x, A)$ is measurable with respect to $\mathcal{B}(\mathbb{R}) \otimes \mathcal{E}$.

In (2), $\mathcal{Q}(x, dy)$ is a probability measure, and you're integrating the function $\mathcal{Q}_s(y, A)$ over this measure.

Let's explain each of these conditions. First, (1) is obvious — at time 0 you should still be at state x. And (3) is just a measurability condition that makes the analyis work. For (2), the Chapman–Kolgomorov equation — which is the most nontrivial one —- can be thought of in the following way. Suppose you have a Markov process (which we haven't defined yet). We draw a graph with time on the x-axis and state on the y-axis. And we want to compute the probability, starting at x at time 0, that we end in A at time $s + t$. Then you can look at time t and see where you end up. If we're at y at time t, moving from x to y corresponds to $\mathcal{Q}_t(x, dy)$, and moving from y to A has probability $\mathcal{Q}_s(y, A)$. And if you integrate over all y, you get the right answer.

There's one more remark about this equation. The Chapman–Kolgomorov equation can be rewritten in the following form. The left-hand side is just $(\mathcal{Q}_{s+t}\mathbf{1}_A)(x)$. And the right-hand side can be written as

$$\int_E (\mathcal{Q}_s\mathbf{1}_A)(y)\, \mathcal{Q}_t(x, dy).$$

Recall the definition of $\mathcal{Q}f$ means that $\mathcal{Q}_s\mathbf{1}_A$ is a function of y, and you're integrating over this measure. YOu can write this as applying $\mathcal{Q}_t$ to this function — so this is simply $\mathcal{Q}_t(\mathcal{Q}_s\mathbf{1}_A)$. So what CK tells us is that

$$\mathcal{Q}_{s+t}\mathbf{1}_A(x) = \mathcal{Q}_t(\mathcal{Q}_s\mathbf{1}_A)(s),$$

which means that we have

$$\mathcal{Q}_{t+s} = \mathcal{Q}_t \cdot \mathcal{Q}_s$$

(if we view $\mathcal{Q}_t$ as an operator $B(E) \to B(E)$).

This property is where the name *semigroup* comes from —- a semigroup is a group without inverses, so you have a binary operation between two elements, and time adds up to give another element.

§21.4 Markov processes

NOw we're ready to define Markov processes (which we'll abbreviate as MP). First, we'll need to fix a probability space and filtration $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{P})$ (and your Markov space will be defined with respect to this).

Definition 21.8. Let $(\mathcal{Q}_t)$ be as above. A Markov process witih respect to $(\mathcal{F}_t)$ is an $\mathcal{F}_t$-adapted process taking values in E such that for all $s, t \geq 0$ and $f \in B(E)$, we have

$$\mathbb{E}[f(X_{s+t}) \mid \mathcal{F}_s] = (\mathcal{Q}_t f)(X_s).$$

Recall that the right-hand side is just $\int_E \mathcal{Q}_t(x, dy)$

This means if you observe X_{s+t} at time s, its expectation only depends on where you are at time s.

Remark 21.9. There's a better way to see this — let's take a conditional expectation over all of X_0, $\ldots$, X_s, i.e.,

$$\mathbb{E}[f(X_{s+t}) \mid \sigma(X_r)_{0 \leq r \leq s}] = (\mathcal{Q}_t f)(X_s).$$

Importantly, the left-hand side only has information on times $[0, s]$, and the right-hand side only depends on time S. So this tells you that the expectation only depends on where you are at at time s — it's about forgetting the past. This is the main philosophy of Markov processes.

§21.5 Finite-dimensional distributions

Next, we'll see that if you're given a Markov process with transition semigroup $\mathcal{Q}_t$, then you can determine the finite-dimensional distributions.

Proposition 21.10

$(\mathcal{Q}_t)$ determines the finite-dimensional distributions of (X_t).

Proof. Suppose you have an initial distribution $X_0 \sim \mu$. Then for any $0 < t_1 < \cdots < t_p$ and sets $A_0, \ldots, A_p$, we have that

$$\mathbb{P}[X_0 \in A_0, X_{t_1} \in A_1, \ldots, X_{t_p} \in A_p] = \int_{A_0} y(dx_0) \int_{A_2} \mathcal{Q}_{t_1}(x_0, dx_1) \cdots \int_{A_{p-1}} \mathcal{Q}_{t_{p-1}-t_{p-2}}(x_{p-2}, dx_{p-1}) \int_{A_p} \mathcal{Q}_{t_p - t_{p-1}}(x_{p-1}, dx_p).$$

□

This is a long integral, so let's parse it. The innermost thing is a function of x_p, so after integrating out with eespect to x_p, you get a function over x_0. Then the next thing is a function on x_{p-1}, and when you integrate that you get a function on x_{p-2}, and so on.

Proof. We'll use the fact that

$$\mathbb{E}[f(X_{s+t}) \mid \mathcal{F}_s] = (\mathcal{Q}_t f)(X_s) = \int \mathcal{Q}_t(X_s, dx) f(x).$$

Now we want to consider $\mathbb{E}[\mathbf{1}_{\{X_0 \in A_0, \ldots, X_{t_p} \in A_{t_p}\}}]$. The usual strategy is to condition on filtrations which become coarser and coarser.

First, if we condition up to time t_{p-1}, then we'll get that this is

$$\mathbb{E}[\mathbf{1}_{X_0 \in A_0, \ldots, X_{t_{p-1}} \in A_{t_{p-1}}} \mathbb{E}[\mathbf{1}_{X_{t_p} \in A_p} \mid \mathcal{F}_{t_{p-1}}]].$$

By applying CM, you can replace the inner expectation with $\mathcal{Q}_{t_p - t_{p-1}}(X_{t_{p-1}}, dx)$. Then you apply this again and get

$$\mathbb{E}[\mathbf{1}_{\ldots, X_{t_{p-2}} \in A_{p-2}} \mathbb{E}\mathbf{1}_{X_{t-1} \in A_{p-1}} \int_A \mathcal{Q}_{t_p - t_{p-1}}(X_{t_{p-1}}, dx_p) \mid \mathcal{F}_{t_{p-2}}].$$

Now you can apply $(*)$ in the same way as before, and you get

$$\mathbb{E}[\mathbf{1}_{\ldots} \int_{A_{p-1} \mathcal{Q}_{t_{p-1} - t_{p-2}}} (X_{t_{p-1}}, dX_{p-1}) \cdot \int_{A_p} \cdots].$$

□

Let's see one example of Markov processes, which is Brownian motion (our favorite example).

Example 21.11

A d-dimensional BM is a Markov process.

What should the transition kernel $\mathcal{Q}_t(x, A)$ be? This should be

$$\mathbb{P}[B_t \in A \mid B_0 = \frac{1}{\sqrt{2\pi t}^d} \exp\left(-\frac{(y-x)^2}{2t}\right).$$

To chedk that this is a Markov process, we need to check that ifyou condition on $\mathcal{F}_s$, then you have a transition kernel — i.e.,

$$\mathbb{P}[B_t \in A \mid \mathcal{F}_s] = Q_{t-s}(B_s, A).$$

And you h show this because $B_t - B_{t'}$ is the increment (whic is 0) and check this processholds.

In general, later you'll see that for some SDEs, the solution can be written as a MP; BM is one example of that.

§21.6 Resolvent

Now we'll go to the next topic, called the *resolvent*. This is an important object which we get from the transition semigroup.

Definition 21.12. Flr any $\lambda > 0$, the λ-resolvent of $(\mathcal{Q}_t)$ is a linear function $\mathcal{R}_\lambda : B(E) \to B(E)$ given by

$$\mathcal{R}_\lambda(f) = \int_0^\infty e^{-\lambda t}(\mathcal{Q}_t, f)(x)\, dt.$$

We need to check that this integral is well-defined, adn that it leads us to a bounded measurable function.

Proof of well-definedness. First, measurability is because by the definition of $\mathcal{Q}_t$ we have that $(t, x) \mapsto Q_t f(x)$ is measurable.

For boundedness, we can just use the fact that $\|\mathcal{Q}_t f\| \le \|f\|$. And you can similarl ycheck that it's bounded. □

Let's write some properties of the resolvent.

Proposition 21.13 (1) We have $\|\mathcal{R}_\lambda f\|_\infty \le \frac{1}{\lambda}\|f\|_\infty$.

(2) For $0 \le f \le 1$, we have $0 \le R_\lambda f \le \frac{1}{2}$.

(3) If $\lambda, \mu > 0$, then $R_\mu - R_\nu + (\lambda - \mu)R_\lambda R_\mu = 0$.

Proof. The proofs of (2) and (1) are direct; (3) is a long r computation, but we'll go through it.

We start wtih $(\mathcal{R}_\lambda \mathcal{R}_\mu)(\lambda)$ and expand out th definition, and we get

$$\int_0^\infty e^{-\lambda t}\mathcal{Q}_t \left(\int_0^\infty e^{\mu s}\, (\mathcal{Q}_s f)(dx) \right).$$

This means we're integrating $\mathcal{Q}_t$ over the measure — this is

$$\int_0^\infty e^{-\lambda t} \int_E \mathcal{Q}_t(x, y) \int_0^\infty e^{-\mu y}(\mathcal{Q}_s f)(y)\, ds\, dt.$$

Then we can exchange the order of summation by Fubini (everything is bounded by the sup-norm bound on f.) So we can swap s and q so that this becomes

$$\int_0^\infty e^{-\lambda t} \int_0^\infty e^{-\mu s} \int_E \mathcal{Q}_t(x, dy)(\mathcal{Q}_s f)(y)\, ds\, dt.$$

And this thing inside is $\mathcal{Q}_t(\mathcal{Q}_s(f))$. So now you have

$$\int_0^\infty e^{-\lambda t} \int_0^\infty e^{\mu s}\mathcal{Q}_{t+s} f(x).$$

Finally, the first term is $e^{-(\lambda-\mu)t}$ and the second is $e^{-\mu(s+t)}$.

Now in the second and third terms, you can substiteus $s + t$ into just s, so you get

$$\int_0^\infty \int_t^\infty e^{-\mu s}\mathcal{Q}_s f(x)\, ds\, dt.$$

Now we'll change the order of integration again; this gives

$$\int_0^\infty e^{-\mu s}(\mathcal{Q}_s f)(f) \int_0^s e^{-(\lambda-\mu)t}.$$

And now we can just compute this, and we get

$$\int_0^\infty e^{-\mu s}(\mathcal{Q}_s f)(x) \cdot \frac{1-e^{-(\lambda-\mu)t}}{\lambda-\mu}\, ds.$$

So we get

$$\mathcal{R}_t = \frac{1}{\lambda-\mu}(\mathcal{R}_\mu(f) - \mathcal{R}_\lambda f(x)),$$

which is exactly what we wanted. □

You may wonder how you think of that third equation. But there is some intuition where it comes from.

Later on, we'll see that this transition semigroup $\mathcal{Q}_t$ can be written as e^{tL} for some operator L, which will be called a *generator*. Then with this formal notation, we can compute the resolvent by thinking of this just as an exponential — as if it's a real number — then we have

$$\mathcal{R}_\lambda = \int_0^\infty e^{-\lambda t} \cdot e^{tL} = \int_0^\infty e^{t(L-\lambda)}\, dt = (\lambda - L)^{-1}$$

So the resolvent can be written as some sort of inverse operator. (This is informal right now, but we'll make sense of it.)

And we also have the matrix identity $M^{-1} - N^{-1} = M^{-1}(N-M)N^{-1}$. If you put in $M = R_\lambda^{-1}$ and $N \leftarrow R_\mu^{-1}$, then you get $R_\lambda - R_\mu = R_\lambda(\mu - \lambda)R_\mu$, which is exactly what we wanted. And we'll make sense of this at the end of the class.

§21.7 Feller process

The next thing we'll define is the Feller process, which is a class of Markov processes with better properties. The reason is we want to argue about continuity. In the definition of a Markov process, there's no notion of continuisy. So first we want to make our state spacs into a topological space.

Let $(E, \mathcal{E})$ be metrizable, locally compacat, σ-finite (i.e., you can cover it by a countable number of compact sets), and equipped with the Borel σ-algebra. Or you can just think of it as $\mathbb{R}^n$ and ignore all these fancy adjectives.

Definition 21.14. We say f tends to 0 at ∞ if for every $\varepsilon > 0$, there is a compact set $K \subseteq E$ such that $|f(x)| \le \varepsilon$ for all $x \in E \setminus K$.

And let $\mathcal{C}_0(E)$ be the space of continuous functions which tend to 0.

Then you can check that $\mathcal{C}_0(E)$ is a Banach space, equipped with the sup norm.

Definition 21.15. A Feller semigroup is a semigroup $(\mathcal{Q}_t)$ satisfying the following two properties:

(1) For any function $f \in \mathcal{C}_0(E)$, we have $\mathcal{Q}_t f \in \mathcal{C}_0(E)$ for all t.

(2) For any $f \in \mathcal{C}_0(E)$, we have $\|\mathcal{Q}_t f - f\|_\infty \to 0$ as $t \to 0$.

A Markov process is Feller If its associated semigroup $(\mathcal{Q}_t)$ is Feller.

You'll see that at frist, it looks like you're imposing many conditions; but actually this will cover a large enough class of Markov processes. For example, Bownian motion os a Feller process, and many processes arisig from SDEs will also be Feller.

Let's state some properties.

Fact 21.16 — If $f \in \mathcal{C}_0(E)$, then for all $\lambda > 0$, we have

$$\mathcal{R}_\lambda f = \int_0^\infty e^{-\lambda t} \mathcal{Q}_t f \, dt \in \mathcal{C}_0(E).$$

Proof. We need to show this function is continuous and vanishes at ∞. So we want to look at $R_\lambda f(y)$ taking $y \to x$. If you write down the definition, because $\mathcal{Q}_t y \to \mathcal{Q}_t x$ as $y \to \infty$, you want to apply dominated convergence to show you can put the limit inside the interval.

But we can see that $\left\| e^{-\lambda t} \mathcal{Q}_t f \right\|_\infty \le e^{-\lambda t} \left\| f \right\|_\infty$, so this will converge to $\mathcal{R}_\lambda(x)$.

Similarly you can argue that $\mathcal{R}_y \lambda f(y) \to 0$ as $y \to \infty$. □

Now let's consider the *space* of functions that can be represented as a resolvent. Let $\lambda > 0$, and let

$$\mathcal{R} = \{\mathcal{R}_lambda f \mid f \in \mathcal{C}_0(E)\}.$$

What we can show is that this set $\mathcal{R}$ doesn't actually depend on your choice of λ.

Proposition 21.17

The set $\mathcal{R}$ doesn't depend on λ. Moeover, it is a dense subset of $\mathcal{C}_0(E)$.

Proof. First we'll show the independence from λ. This can be done by recalling the resolvent identity that we just proved — we showed that

$$\mathcal{R}_\lambda(f + (\lambda - \mu)\mathcal{R}_\mu f) = \mathcal{R}_\mu f.$$

So any function in $\mathcal{R}_\mu$ can be interpreted as in the image of $\mathcal{R}_\lambda$; this shows $\mathrm{Im}(\mathcal{R}_\mu) \subseteq \mathrm{Im}(\mathcal{R}_\lambda)$, and you can swap their roles to get the reverse containment).

For (2), this is harder; we want to approximate any function in $\mathcal{C}_0(E)$. The idea here is — recall that at least intuitively, we have that $\mathcal{R}_\lambda = (\lambda - L)^{-1}$. This means $\lambda \mathcal{R}_\lambda = (\mathrm{Id} - \lambda^{-1} L)^{-1}$. If you take $\lambda \to \infty$, then the λ^{-1} term goes to 0, so you might hope this converges to Id as $\lambda \to \infty$.

With this in in mind, we might guess to consider

$$(\lambda \mathcal{R}_\lambda f)(x) = \int_0^\infty \lambda e^{-\lambda} (\mathcal{Q}_t f)(x) \, dt = \int_0^\infty e^{-\lambda t} (\mathcal{Q}_{t/\lambda})(f)(x) \, dt.$$

We want to show this uniformly converges, so we take a difference — we can bound

$$\left\| \lambda \mathcal{R}_\lambda f - f \right\| \le \int_0^\infty e^{-\lambda t} \left\| Q_{t/\lambda} f - f \right\|_\infty dt.$$

So you have a uniform obund. And you're taking $\lambda \to \infty$, so the $\mathcal{Q}_\lambda$ term should go to 0. Also, you can see this thing is bounded by $2 \left\| f \right\|_\infty$, so we can apply the dominated convergence theorem. □

§21.8 Generators

We'll finish just by introducing what a generator is.

Let's recall the intuition from earlier about the Markov semigroup — we wanted to interpret this kernel $\mathcal{Q}_t$ as some exponential $\mathcal{Q}_t f = e^{tL} f$ (where L is an operator); this L would be a generator. How could you cecover L from t? You could take a derivative at time 0 — so you consider

$$\lim_{t \to 0} \frac{\mathcal{Q}_t f - f}{t}.$$

Unfortunately, however, this limit doesn't exist in general for arbitrary f.

So we want to restrict the domain $\mathcal{C}_0$ into some smaller domain so that this lmiit does exist, and show that subset is dense in $\mathcal{C}_0$. And that subset will turn out to be the $\mathcal{R}$ we just defined.

Goal 21.18. Show that $Lf = \lim_{t \to 0} \frac{\mathcal{Q}_t f - f}{t}$ is well-defined for $f \in \mathcal{R}$, and that $\mathcal{R}$ is dense in $\mathcal{C}_0(E)$.

§22 April 28, 2025

§22.1 Review

Last class, we introduced Markov transition groups and started discussing the basics of general Markov processes in continuous state space and continuous time. This is a kind of basic language that if you want to continue doing this kind of probability theory, it's crucial to see (if you want to do continuous-time continuous-state processes). Also, last time we saw the notion of a Feller semigroup, which is basically some sort of regularity property (which will be satisfied for all Markov processes that arise for us):

Definition 22.1 (Feller semigroup). Let $(\mathcal{Q}_t)$ be a transition semigroup on E. We say $(\mathcal{Q}_t)$ is a Feller semigroup if:

(i) For all $f \in C_0(E)$, we have $\mathcal{Q}_t f \in C_0(E)$.

(ii) For all $f \in C_0(E)$, we have $\|\mathcal{Q}_t f - f\| \to 0$ as $t \to 0$.

Usually we take $E = \mathbb{R}^d$ (at least, that'll suffice for us). And you want your semigroup to map functions decaying at ∞ to functions decaying ∞. That's kind of saying that as a probability measure, $\mathcal{Q}_t$ should kind of not put too much mass at ∞ — you kind of want your probability mass to be controlled in some sense.

There's also the definition of a resolvent:

Definition 22.2. Let $\lambda > 0$. The resolvent is the map $\mathcal{R}_\lambda : B(E) \to B(E)$ defined by

$$(\mathcal{R}_\lambda f)(x) = \int_0^\infty e^{-\lambda t} (\mathcal{Q}_t f)(x)\, dt.$$

We won't ever use this directly; the main reason we introduced it is to talk about this thing called a *generator*, which is going to characterize your semigroup (and which comes up over and over again — it's maybe the most important concept for continuous-time Markov processes).

§22.2 Generators

Definition 22.3. Let $D(L) = \{f \in C_0(E) \mid \lim_{t \downarrow 0} \frac{\mathcal{Q}_t f - f}{t} \text{ exists in } C_0(E)\}$. For $f \in D(L)$, we define

$$Lf = \lim_{t \downarrow 0} \frac{\mathcal{Q}_t f - f}{t}.$$

Remark 22.4. The set $D(L)$ is actually a linear subspace of $C_0(E)$. We call L the generator of the transition semigroup $(\mathcal{Q}_t)$, and $D(L)$ is called the domain of L.

Example 22.5

For Brownian motion, the generator will basically be the Laplacian Δ. The Laplacian isn't defined for all continuous functions (because you have to take two derivatives). So you have to take some space on which you can define the Laplacian; for example, you can consider the space of twice continuously differentiable functions with compact support — we have

$$C_c^2(\mathbb{R}^d) \subseteq D(\Delta) \subseteq C_0(\mathbb{R}^d).$$

This is a definition; now we'll begin to develop some foundational properties of this generator. The first is that it actually commutes with your semigroup.

Proposition 22.6

Let $f \in D(L)$ and $s \geq 0$. Then $\mathcal{Q}_s f \in D(L)$ and $L\mathcal{Q}_s f = \mathcal{Q}_s L f$.

Proof. If $s = 0$, then $\mathcal{Q}_s f$ is f itself, and this is a trivial identity; so we just need to prove it when $s > 0$. For this, we look at the difference quotient

$$\frac{1}{t}(\mathcal{Q}_t(\mathcal{Q}_s f) - \mathcal{Q}_s f).$$

We want to verify that if we fix s and take the $t \to 0$ limit, this will actually exist. To see why, we want to use the semigroup property of $\mathcal{Q}$ — $\mathcal{Q}$ is a semigroup, so you can actually take the $\mathcal{Q}_s$ outside and think of this as

$$\mathcal{Q}_s\left(\frac{1}{t}(\mathcal{Q}_t f - f)\right)$$

(this is because $\mathcal{Q}_t \mathcal{Q}_s = \mathcal{Q}_{t+s} = \mathcal{Q}_{s+t} = \mathcal{Q}_s \mathcal{Q}_t$, so these commute). And by assumption $f \in D(L)$, so as you take $t \to 0$, the inside converges to Lf. And $\mathcal{Q}_s$ is a contraction (so in particular it's continuous); so this converges to $\mathcal{Q}_s L f$ as $t \downarrow 0$. □

Another fundamental property of the generator is the 'heat equation.'

Proposition 22.7

For $f \in D(L)$, for all $t \geq 0$, we can write

$$\mathcal{Q}_t f = f + \int_0^t \mathcal{Q}_s(Lf)\, ds = f + \int_0^t L(\mathcal{Q}_s f)\, ds.$$

This is the integral form of the identity, but if you take derivatives in t on both sides, you're basically saying that the function $t \mapsto \mathcal{Q}_t f$ is C^1 (it takes values in a space of functions, but you can define continuously differentiable functions which are function-valued), and moreover, its time derivative is precisely

$$\partial_t \mathcal{Q}_t f = L\mathcal{Q}_t f = \mathcal{Q}_t L f.$$

You see that if you integrate both sides of this identity, you precisely get the integral form.

Before we talk about the proof:

Remark 22.8. If you view $\partial_t \mathcal{Q}_t f = L\mathcal{Q}_t f$ as an equation on the operator $\mathcal{Q}_t$, you're saying that $\mathcal{Q}_t$ basically satisfies the differential equation $\partial_t \mathcal{Q}_t = L\mathcal{Q}_t$. If you're on a *finite* state space, then $\mathcal{Q}_t$ is just going to be a (finite square) matrix. And if you were to try solving this matrix ODE (where L is some other matrix), there's really just one solution — you'd get $\mathcal{Q}_t = e^{tL}$. (If everything was a real number, this would be the solution; and the same is true if you have matrices.) This matrix exponential can be defined by the usual Taylor series

$$\mathcal{Q}_t = e^{tL} = \sum_{k=0}^{\infty} \frac{t^k}{k!} L^k.$$

And you can make sense of $\partial_t \mathcal{Q}_t = L\mathcal{Q}_t$ even in infinite dimensions. This is why you can expect the generator to determine the semigroup — you can expect you can write $\mathcal{Q}_t$ in terms of L in a similar manner (but that won't be super important for us, so we won't really go into it).

Proof. Let's directly try computing the time derivative of $\mathcal{Q}_t f$ — fix some t, and consider a difference quotient

$$\frac{1}{\varepsilon}(\mathcal{Q}_{t+\varepsilon} f - \mathcal{Q}_t f).$$

We want to take the $\varepsilon \to 0$ limit (if it exists, that gives us the time derivative).

And this looks very similar to the generator — for instance, you can write this as

$$\mathcal{Q}_t \left(\frac{\mathcal{Q}_\varepsilon f - f}{\varepsilon} \right).$$

As $\varepsilon \to 0$, this converges to $\mathcal{Q}_t L f$, which by the previous proposition is $L\mathcal{Q}_t f$.

So you've shown $\mathcal{Q}_t f$ is C^1 (as a function of t), with $\partial_t \mathcal{Q}_t f = L\mathcal{Q}_t f = \mathcal{Q}_t L f$. And being C^1 means that the derivative has to be continuous in time; but Lf by definition is a continuous function decaying at ∞, and by your Feller assumption $\mathcal{Q}_t L f$ is going to be continuous in time. (This should come out of the second identity — you have $\|\mathcal{Q}_{t+\varepsilon} f - \mathcal{Q}_t f\| \le \|\mathcal{Q}_\varepsilon f - f\|$, and if you send $\varepsilon \to 0$, then this goes to 0. Here we're using the fact that $\mathcal{Q}_t$ is a contraction.) □

Also, we have this formal viewpoint of the semigroup in terms of the generator, you can also think about the resolvent in the following way: we have

$$\mathcal{R}_\lambda f = \int_0^\infty e^{-\lambda t} \mathcal{Q}_t f \, dt.$$

Now if you formally substitute in $\mathcal{Q}_t = e^{tL}$, you get

$$\mathcal{R}_\lambda f = \int_0^\infty e^{-\lambda t} e^{Lt} \, dt = \int_0^\infty e^{t(L-\lambda)} \, dt$$

(by $L - \lambda$ we really mean L minus λ times the identity operator). If you formally believe that the usual rules of integration apply with these things, then you should get that this is

$$-(L-\lambda)^{-1}$$

(this is becaus $\int_0^\infty e^{-t\alpha} \, dt = \alpha^{-1}$ for real numbers α). So you should really think of the resolvent as some sort of inverse of your generator (really plus some constant). In the notes, this is why you have the resolvent identity — taking the difference of two resolvents is really the difference of inverses of two matrices.

Remark 22.9. You can make all this precise if L is a finite matrix (assuming it's diagonalizable) — then you can write an eigendecomposition of e^{tL} in terms of the eigenbasis (where you act on each eigenbasis element by an exponential of the corresponding eigenvector). Then you're really just reducing, on each eigenvector, to computing a one-dimensional real integral $\int_0^\infty e^{-t\alpha}\,dt$. So this all makes sense when L is a finite-dimensional diagonalizable square matrix. And you can make this all make sense even in infinite dimensions if e.g. L is the Laplacian; that gives you relations to PDE.

But at least because of this formal relation, you expect that

$$\mathcal{R}_\lambda(\lambda - L)f = f \quad \text{and} \quad (\lambda - L)\mathcal{R}_\lambda f = f$$

(because of this formal inverse relation). This is what we'll prove next.

Remark 22.10. We're kind of proving all these propositions to eventually prove that the generator actually determines the semigroup, because that's maybe the main conceptual thing to understand — if you want to give the semigroup, you can just talk about the generator.

Remark 22.11. The reason you want $\lambda > 0$ is that in principle, L could have a 0 eigenvalue, and that would be really bad when you integrate (if $\lambda = 0$), because then you'd be integrating 1 from 0 to ∞, which would blow up.

That's actually the case for the Laplacian — it has a 0 eigenvalue, because it's 0 on any constant (or even linear) function. Having a 0 eigenvalue basically means having a nonzero kernel.

That's why you want $\lambda > 0$ — then actually $\lambda - L$ will never have a 0 eigenvalue.

Proposition 22.12

Let $\lambda > 0$. Then:

(i) For every $g \in C_0(E)$, we have that $\mathcal{R}_\lambda g \in D(L)$ and

$$(\lambda - L)\mathcal{R}_\lambda g = g.$$

(ii) For $f \in D(L)$, we have $\mathcal{R}_\lambda(\lambda - L)f = f$.

For (i), the point is that if you know $\mathcal{R}_\lambda g \in D(L)$, then you can apply L to it. The reason why you expect this to hold is precisely because of the above formal interpretation — $\mathcal{R}_\lambda$ is kind of like taking an inverse of L. The whole issue with the domain thing is because coming back to the Laplacian, in general you can't take two derivatives of a function. But $\mathcal{R}_\lambda$ is like taking the *inverse* of a Laplacian, which actually maps a continuous function to a twice differentiable function; and then you *can* apply two derivatives. So heuristically, the point of why you have this statement is that $\mathcal{R}_\lambda$ is heuristically like applying an inverse, so then you should be able to apply the generator (since you just applied its inverse).

Proof of (i). The proof is basically playing around with these definitions. For (i), let's again write out the definition of the generator — we want to compute

$$\varepsilon^{-1}(\mathcal{Q}_\varepsilon \mathcal{R}_\lambda g - \mathcal{R}_\lambda g).$$

Eventually we want to show this has a limit as $\varepsilon \to 0$, and we want to show that limit is whatever is going to be imposed by this identity — we basically want to have that

$$L\mathcal{R}_\lambda g = \lambda \mathcal{R}_\lambda g - g.$$

So basically we want to show that as $\varepsilon \to 0$, this thing converges to the right-hand side.

To do that, let's write out the definition of the resolvent; then this becomes

$$\varepsilon^{-1}\left(\mathcal{Q}_\varepsilon \int_0^\infty e^{-\lambda t}\mathcal{Q}_t g\,dt - \int_0^\infty e^{-\lambda t}\mathcal{Q}_t g\,dt\right).$$

The first thing you have to say is that you can move the operator inside them integral to get

$$\varepsilon^{-1}\left(\int_0^\infty e^{-\lambda t}\mathcal{Q}_{t+\varepsilon} g\,dt - \int_0^\infty e^{-\lambda t}\mathcal{Q}_t g\,dt\right).$$

Basically the reason why you can do this is $\mathcal{Q}_\varepsilon$ is a contraction, so it's continuous. You can approximate both integrals by taking a limit of discrete sums; and on the discrete level you have this commutation, so you just need to say why you can take this limit in the right way, and this can be justified using the fact that $\mathcal{Q}_\varepsilon$ is a continuous operator.

Now we want to figure out how to take the $\varepsilon \to 0$ limit. After playing around with it, you realize you want to decompose the integral in the following way: We write the second integral as a 0 to ε part and an ε to ∞ part. So we get

$$-\varepsilon^{-1}\left(\int_0^\varepsilon e^{-\lambda t}\mathcal{Q}_t g\,dt\right) + \varepsilon^{-1}\left(\int_0^\infty e^{-\lambda t}\mathcal{Q}_{t+\varepsilon} g\,dt - \int_\varepsilon^\infty e^{-\lambda t}\mathcal{Q}_t g\,dt\right).$$

The whole point of doing this was that you can group the last two integrals — we'll call this $I_1 + I_2$, where I_1 is the first term and I_2 is the second (these depend on ε, but we won't write that out).

Let's first look at I_1, which is easier to handle. Since $\mathcal{Q}_t g \to g$ as $t \to 0$, we'll have that $I_1 \to -g$ as $\varepsilon \to 0$. This is because you kind of see that you have this integral from 0 to ε, but you also have this ε^{-1}, so this is really behaving like an average of your integrand at very very small times. And at very small times $e^{-\lambda t}$ is basically 1, and $\mathcal{Q}_t g$ converges to g (in a uniform sense — in the Banach space $C_0(E)$). So this whole thing should converge to $-g$ (the $-$ is because of the $-$ sign in front). So the first term is relatively direct, just using the Feller assumption (specifically the second part of it).

Now let's talk about I_2. Now we want to group these two integrals together, which means we want to change variables to rewrite the second as an integral from 0 to ∞. Once you do that, you can write

$$I_2 = \varepsilon^{-1}\int_0^\infty (e^{-\lambda t} - e^{-\lambda(t+\varepsilon)})\mathcal{Q}_{t+\varepsilon} g\,dt.$$

(In the second term, when we change variables, we replace t by $t+\varepsilon$; then you can group the $\mathcal{Q}_{t+\varepsilon}$'s and just get this difference.)

Now as $\varepsilon \to 0$, you want to interpret the ε^{-1} and this grouping of exponentials as converging to something (we're basically taking the derivative of these exponentials) — pointwise in t, we have that

$$\varepsilon^{-1}(e^{-\lambda t} - e^{-\lambda(t+\varepsilon)}) \to \lambda e^{-\lambda t}$$

(the derivative of $e^{-\lambda t}$ is $-\lambda e^{-\lambda t}$).

Now we just need to say why we can commute the limit and integral. If we can do that, we'll be done — if we substitute $\lambda e^{-\lambda t}$ and $\varepsilon \to 0$, we precisely get $\lambda \mathcal{R}_\lambda g$, which is what we want I_2 to be.

To do that, we need to do some sort of dominated convergence. We want a bound on this thing which is uniform in ε (because we want dominated convergence, so we need to bound this by some common integrable function). For this, we can use the bound

$$\varepsilon^{-1}\left|e^{-\lambda t} - e^{-\lambda(t-\varepsilon)}\right| = e^{-\lambda}\frac{\left|1 - e^{-\lambda\varepsilon}\right|}{\varepsilon}.$$

Now we can bound this by

$$e^{-\lambda t} \cdot \frac{\lambda \varepsilon}{\varepsilon}$$

(using the fact that $1 - e^{-x} \leq x$). This is a uniform bound in ε. And that's basically what you need to get dominated convergence (you also need to say $\mathcal{Q}_{t+\varepsilon} g$ is uniform, but g is bounded because it's in $C_0(E)$, and $\mathcal{Q}$ is a contraction, so you can always bound this by some constant that doesn't depend on ε).

So this allows you to apply the dominated convergence theorem, to say that

$$I_2 \to \int_0^\infty \lambda e^{-\lambda t} \mathcal{Q}_t g \, dt,$$

which you recognize is $\lambda \mathcal{R}_\lambda g$.

Now if you put these two things together, we get precisely that

$$L\mathcal{R}_\lambda g = \lambda \mathcal{R}_\lambda g - g.$$

We're saying that as $\varepsilon \to 0$, this difference quotient converges, so $\mathcal{R}_\lambda g \in D(L)$; and the value of L on this thing is basically the limit, where we showed $I_2 \to \lambda R_\lambda g$ and $I_1 \to -g$, giving what we want. □

Proof of (ii). To prove (ii), we need to apply $\mathcal{R}_\lambda$ to the generator. So let's write out the definition of the resolvent; then this becomes

$$\mathcal{R}_\lambda L f = \int_0^\infty e^{-\lambda t} \mathcal{Q}_t L f \, dt.$$

Then using the fact that L and $\mathcal{Q}_t$ commute, this becomes

$$\int_0^\infty e^{-\lambda} L \mathcal{Q}_t f \, dt.$$

What you want to do is basically be able to pull L out of the integral. Let's assume for a moment that you can do that, so this becomes

$$L \int_0^\infty e^{-\lambda t} \, \mathcal{Q}_t f \, dt$$

(we'll call this identity $(*)$; we're going to prove it at the end). Assuming this, we're basically done — then you get that this is $L\mathcal{R}_\lambda f$. And we computed $L\mathcal{R}_\lambda f$ in part (i) — this is precisely $\lambda \mathcal{R}_\lambda f - f$. And now if you rearrange this identity, you should get what you want — we've shown

$$\mathcal{R}_\lambda L f = \lambda \mathcal{R}_\lambda f - f,$$

and if you rearrange that identity you get $\mathcal{R}_\lambda(\lambda - L) f = f$.

So it all becomes about justifying why $(*)$ is true — you want to be able to commute the generator with the integral. And why is that going to be true? Well, f by assumption is in $D(L)$; we're going to have to use that. Let's start with the right-hand side of $(*)$, i.e., the expression

$$L \int_0^\infty e^{-\lambda t} \mathcal{Q}_t f \, dt.$$

And let's write out the definition of L: so this is

$$\lim_{\varepsilon \downarrow 0} \varepsilon^{-1} \left(\mathcal{Q}_\varepsilon \int_0^\infty e^{-\lambda t} \mathcal{Q}_t f \, dt - \int_0^\infty e^{-\lambda t} \mathcal{Q}_t f \, dt \right).$$

And in the first part, we saw that you can commute $\mathcal{Q}_\varepsilon$ inside the first integral, because $\mathcal{Q}_\varepsilon$ is a bounded operator; so we can write this as

$$\lim_{\varepsilon \downarrow 0} \int_0^\infty e^{-\lambda t} \varepsilon^{-1} (\mathcal{Q}_{t+\varepsilon} f - \mathcal{Q}_t f) \, dt.$$

So the identity is true for any $\varepsilon > 0$, before taking the limit. Now you take the limit; you see that $\varepsilon^{-1}(\mathcal{Q}_{t+\varepsilon}f - \mathcal{Q}_t f) \to \mathcal{Q}_t L f$. Then you have to justify some sort of dominated convergence, but that's doable (using that $\mathcal{Q}_t$ is a contraction). In the end, you're going to get that this limit is

$$\int_0^\infty e^{-\lambda t}\mathcal{Q}_t L f\, dt.$$

And now you can use the fact that $\mathcal{Q}_t$ and L commute to get that this is $\int_0^\infty e^{-\lambda t} L\mathcal{Q}_t f\, dt$, which was the other side of $(*)$. □

Remark 22.13. A consequence of this is that the set $\mathcal{R} = \mathrm{Im}(\mathcal{R}_\lambda)$ (the image of the resolvent — last time it was proven that the image of $\mathcal{R}_\lambda$ doesn't depend on λ, using the resolvent identity) is precisely $D(L)$. Moreover, $\mathcal{R}_\lambda$ and $\lambda - L$ are inverses of each other.

§22.3 The generator determines the semigroup

Now we can prove this corollary we mentioned earlier that the generator determines the semigroup. And then after we prove that (which is a kind of foundational fact — we might not ever apply it, but it's good to know), we'll talk about how you compute the generator in explicit examples.

Corollary 22.14

The (Feller) semigroup $(\mathcal{Q}_t)$ is determined by the pair $(L, D(L))$.

Proof. Let $f \in C_0(E)$, and for convenience assume $f \geq 0$. Then first, we can note that L and $D(L)$ characterize $\mathcal{R}_\lambda f$ — it's the unique element of $D(L)$ such that $(\lambda - L)\mathcal{R}_\lambda f = f$. We just proved that this identity is true, and the fact that it's unique comes from the fact that if you have g_1 and g_2 such that $(\lambda - L)g_1 = (\lambda - L)g_2$ (so g_1 and g_2 are both possible values of $\mathcal{R}_\lambda f$, and we want to show they're equal), then you can apply the other identity to say that

$$\mathcal{R}_\lambda(\lambda - L)g_1 = \mathcal{R}_\lambda(\lambda - L)g_2,$$

and since $\mathcal{R}_\lambda(\lambda - L)$ is the identity, this implies $g_1 = g_2$. (So uniqueness is basically because you have inverses.)

So L and $D(L)$ characterize the value of $\mathcal{R}_\lambda f$ for any λ.

On the other hand, the values of $\mathcal{R}_\lambda f$ for all $\lambda > 0$, defined by

$$(\mathcal{R}_\lambda f)(x) = \int_0^\infty e^{-\lambda t}(\mathcal{Q}_t f)(x)\, dt,$$

uniquely determine the function $t \mapsto (\mathcal{Q}_t f)(x)$. (Here we're fixing some x in our space and viewing this as a function of t. The point is that if I know this integral for every λ, then I actually know the function of t. This is some statement about inverting Laplace transforms. You can basically view this as the Laplace transform of some function in t, and you want to say why that uniquely determines the function. You can just Google this; Le Gall doesn't provide a reference.)

Because of that, you get that $\mathcal{Q}_t f$ is determined by the pair $(L, D(L))$ for every f which is nonnegative. But then knowing $\mathcal{Q}_t f$ for every nonnegative f allows you to obtain $\mathcal{Q}_t f$ for general f (you can split a continuous function into positive and negative parts). □

Remark 22.15. This is a very soft proof — the main step, the fact that you can invert Laplace transforms, we're sort of assuming.

§22.4 Computing the generator

Now let's talk about how you can compute the generator. If you had a formula for the semigroup — for example, last time we saw it for BM — then yolu can try to directly compute the limit using the definition of the generator. So that's the direct way. But usually there's maybe an easier way to compute the generator, and that's what we'll talk about.

We need some setup: Suppose that on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we have, for all $x \in E$, a stochastic process (X_t^x) (the superscript x denotes the starting point, as we'll see soon) with continuou sample paths, which is actually Markov with semigroup $(\mathcal{Q}_t)$. Here we're always assuming $(\mathcal{Q}_t)$ is Feller (we're talking about a generator, which requires you to assume this), with respect to the filtration $(\mathcal{F}_t)$, and such that

$$\mathbb{P}[X_0^x = x] = 1$$

(i.e., the process X^x starts at x).

This is some setup — suppose that given $(\mathcal{Q}_t)$, we were able to construct all these stochastic processes on a common probability space. For BM, this is very easy — to get a BM started at x, you just take a standard BM and add x.

Theorem 22.16

Let $h, g \in C_0(E)$. Then the following are equivalent:

(i) $h \in D(L)$ and $Lh = g$.

(ii) For all $x \in E$, the process given by

$$h(X_t^x) - \int_0^t g(X_u^x)\, du \quad \text{is a martingale with respect to } (\mathcal{F}_t).$$

The way we'll usually use this is that we'll verify (ii), and then you can get the value of L on a function h (it'll be g). And as we'll see, the way we verify (ii) is typically by Ito's formula.

Proof $(i) \implies (ii)$. For this, recall we have the integral form

$$\mathcal{Q}_t h = h + \int_0^t L\mathcal{Q}_s h\, ds = h + \int_0^t \mathcal{Q}_s Lh\, ds.$$

And the assumption in (i) is that $Lh = g$, so we can write this as

$$\mathcal{Q}_t h = h + \int_0^\infty \mathcal{Q}_s g\, ds.$$

Now we want to verify (ii). So let's compute the conditional expectation of $h(X_t^x)$ (we eventually want to show something is a martingale) — so we want to compute

$$\mathbb{E}[h(X_t^x) \mid \mathcal{F}_s].$$

First, you have a Markov process, so you can write this as

$$(\mathcal{Q}_{t-s}h)(X_s^x)$$

(since X is a Markov process with semigroup $\mathcal{Q}$). Now you want to use your identity (which we wrote for time t) at time $t - s$. If you do so, you get that this is

$$h(X_s^x) + \int_0^{t-s} (\mathcal{Q}_u g)(X_s^x)\, du.$$

Now you can see where the martingale comes from — you're taking h minus an integral, and you want to get h at time s minus an integral. So to get it, you just need to reinterpret the second term. For this, we again use the fact that you have a Markov process, so you can write this as

$$h(X_s^x) + \int_0^t \mathbb{E}[g(X_{s+u}^x) \mid \mathcal{F}_s]\, du.$$

Now you rewrite this integral as one from s to t, ygiving

$$h(X_s^x) + \int_s^t \mathbb{E}[g(X_u^*) \mid \mathcal{F}_s]\, du.$$

That implies this thing is a martingale — if you write out hte martingale identity for this process, it'll precisely be a rarrangement of the above, because you subtract out the integral from 0 to t.

We actually need one more step — we want to take the conditional expectation outside and write this as

$$h(X_s^x) + \mathbb{E}\left[\int_0^t g(X_u^x)\, du \;\middle|\; \mathcal{F}_s\right].$$

This takes some justification, but it's true; and *now* this is precisely the martingale identity. □

Proof $(ii) \implies (i)$. (This is the direction we're most interested in.)

Let's start with the definition of a generator — we want to compute

$$\varepsilon^{-1}(\mathcal{Q}_\varepsilon h(x) - h(x))$$

(as $\varepsilon \to 0$). You want to interpret this in terms of your Markov process; we have

$$\mathcal{Q}_\varepsilon h(x) = \varepsilon^{-1}(\mathbb{E}\left[h(X_\varepsilon^x)\right] - \mathbb{E}[h(X_0^x)])$$

(because $X_0^x = x$). (The last term is just $h(x)$.)

This is what we're trying to understand. Now we want to use the fact that we have a martingale, so when we take an expectation at time t, that's equal to the expectation at time 0. So we get

$$\mathbb{E}[h(X_t)] - \mathbb{E}\int_0^t g(X_u)\, du = h(x).$$

Now we can rearrange this identity to rewrite our difference as just an expectation of the ingegral of g — we get

$$\varepsilon^{-1}\int_0^\varepsilon \mathbb{E}[g(X_u)^2]\, du.$$

Now it's kind of clear wha tyou want to do — we want to say that as $\varepsilon \to 0$, this converges to g. And this is going to be true (though there's some justification required).

As a first step, let's write this in terms of the semigroup as

$$\varepsilon^{-1}\int_0^\varepsilon (\mathcal{Q}_u g)(x)|, du$$

(since X^x starts ar 0).

Now we want to basically say that

$$\sup_x \left\| \varepsilon^{-1}\int_0^\varepsilon (\mathcal{Q}_u g)(x)\, dx - g(x) \right\| \to 0.$$

Because we want to show that $\varepsilon^{-1}(\mathcal{Q}_\varepsilon h(x) - h(x))$ converges (as a function) oiin $C_0(E)$, where convergence is given by the sp norm. For the left-hand side, we can write it in terms of this norm as

$$\left\| \varepsilon^{-1} \int_0^\varepsilon (\mathcal{Q}_u g - g)\, du \right\|.$$

Then you use the triangle inequalit yto put the norm inside the integral; so you get

$$\varepsilon^{-1} \int_0^\varepsilon \|\mathcal{Q}_u g - g\|\, du.$$

And the inside goes to 0 as $u \to 0$; so you're taking an average of a function which goes to 0, which means this also goes to 0.

This shows $h \in D(L)$ and $Lh = g$. □

Student Question. *How do you justify exchanging the integral and expectation in* $\int_s^t \mathbb{E}[g(X_u^x)\ mid \mathcal{F}_s]$*?*

Answer. Let's say you have something like

$$\int_0^t \mathbb{E}[X_u \mid \mathcal{G}]\, du,$$

where X is a process in u, and you want to say why this equals

$$\mathbb{E}\left[\int_0^t X_u\, du \,\middle|\, \mathcal{G} \right]$$

(where $\mathcal{G}$ is some σ-algebra). So you want to take $A \in \mathcal{G}$ and show that

$$\mathbb{E}\left[\int_0^t \mathbb{E}[X_u \mid \mathcal{G}]\, du \mathbf{1}_A \right] = \mathbb{E}\left[\int_0^t X_u\, du \mathbf{1}_A \right]$$

(by the definitions of conditional expectation). Now on the left-hand side, we can take the indicator inside the integral (you need some kind of regularity, but suppose $\mathbb{E}[X_u \mid \mathcal{G}]$ is measurable as a function in n, which should be true), we can write this as

$$\int_0^t \mathbb{E}[\mathbb{E}[X_u \mid \mathcal{G}]\mathbf{1}_A]\, du.$$

(You can justify the exchange of regular expectation and integration by assuming X is bounded, which will be fine.) And now you juse hte definition of conditional expectation to get that this is equal to

$$\int_0^t \mathbb{E}[X_u \mathbf{1}_A]\, du.$$

And you can again take the expectation outside the integral to get this identity.

Let's see an example of the theorem.

Example 22.17

Let (B_t) be a d-dimensional $(\mathcal{F}_t)$-Brownian motion starting at 0. Define $B_t^x = x + B_t$. Then for all x, (B_t^x) is a Markov process with semigroup

$$(\mathcal{Q}_t f)(x) = \mathbb{E}[f(x + \sqrt{t}Z)]$$

for $Z \sim \mathcal{N}(0, I)$ (because $x + \sqrt{t}Z$ is the law of B_t^x). And you can write this as an explicit integral against the Gaussian density, but we'll skip that.

The semigroup is going to be Feller (that's left as an exercise), so it has a generator. How do you compute it?

Well, by Ito's formula, say we take $f \in C_b^2(\mathbb{R}^2)$ (a twice-differentiable function where f and its first and second derivatives are bounded and continuous). Then we can compute $df(B_t^x)$, and we'll get a martingale part and a FV term. If you subtract the FV term, then you'll get a martingale, so you can apply this theorem.

Ito's formula gives you that

$$df(B_t^x) = \nabla f(B_t^x)\, dB_t^x + \frac{1}{2}\Delta f(B_t^x)\, dt$$

(the fact that we assumed boundedness means that the first term is a martingale, not just a local martingale). So you can rewrite Ito's formula another way, and you get that

$$f(B_t^x) - \frac{1}{2}\int_0^t \Delta f(B_s^x)\, ds$$

is a martingale. (This is the integral formulation of Ito's formula.) Thus by the theorem, you get that $f \in D(L)$, and moreover

$$Lf = \frac{1}{2}\Delta f.$$

This shows $C_b^2(\mathbb{R}^d) \subseteq D(L)$.

So this theorem allows you to explicitly compute the generator, and for BM it's this explicit expression — just the Laplacian.

Remark 22.18. The identity $\partial_t \mathcal{Q}_t f = L\mathcal{Q}_t f$ in the case of Brownian motion becomes

$$\partial_t \mathcal{Q}_t f = \frac{1}{2}\Delta \mathcal{Q}_t f.$$

Basically you're saying you can solve the heat equation by running a Brownian motion. If you call $\mathcal{Q}_t f = u_t$ (it's a function of t and also x — we're not writing the spatial variable), you get $\partial_t u_t = \frac{1}{2}\Delta u_t$. So BM is intimiately related to the heat equation, because if you specialize this equation to BM you get exactly the heat equation.

The Markov property stuff we're going to skip, because next time we'll go into some more explicit examples of Markov processes and talk about some of their properties. The Markov property is in the notes, but we won't really use it. But it's also good to know, so if you want to you can read the notes.

§22.5 Invariant measure

We'll end today by defining the notion of an invariant measure.

Definition 22.19. Let $(\mathcal{Q}_t)$ be a Markov semigroup, and μ be a probability measure on E. We say μ is an invariant measure for $(\mathcal{Q}_t)$ if for all bounded functions $f \in B(E)$ and all times $t \geq 0$, we have that

$$\int \mu(dx)\,(\mathcal{Q}_t f)(x) = \int \mu(dx)\,f(x).$$

Usually E is some topological space, so you take E and its Borel σ-algebra; but we can just think of E as $\mathbb{R}^d$.

In terms of an actual process, basically if you have a Markov process X with semigroup $\mathcal{Q}$ and you let its initial distribution be $X_0 \sim \mu$, then you're just saying that $X_t \sim \mu$ for all $t \geq 0$. That's why you call it an invariant measure — the law of your Markov process doesn't change if you start from this measure.

Next class we'll see a process whose invariant measure is a standard Gaussian; so you should think of it as the SDE version of a standard Gaussian. It has many important properties that we'll begin discussing next time.

§23 April 30, 2025

Today we'll begin discussing the Ornstein–Uhlenbeck process, which is one of the most explicit examples of a SDE. (Well, the most explicit would just be $dX = dB$, but this is not very interesting; this is the second-most.)

§23.1 Ornstein–Uhlenbeck process

To set the stage, throughout B will be a d-dimensional $(\mathcal{F}_t)$-Brownian motion, where we assume $(\mathcal{F}_t)$ is complete (we always assume completeness, because we needed it to e.g. define the quadratic variation).

Definition 23.1. An Ornstein–Uhlenbeck (OU) process (X_t) is an adapted stochastic process with continuous sample paths satisfying the SDE

$$dX_t = -X_t + \sqrt{2}\,dB_t. \tag{23.1}$$

This is an SDE. We'll comment on the factor of $\sqrt{2}$ later (more generally you can let it be any positive constant, and we'd probably still call it an OU process). Even more generally, you can also put a positive constant factor in front of the $-X_t$ — you could consider

$$dX_t = -\alpha X_t + \beta\,dB_t,$$

where $\alpha, \beta > 0$. But if you understand the standard case, you basically understand the more general one as well. So we'll just focus on the standard one (and we'll see later why we want the $\sqrt{2}$).

By definition, this means X satisfies the integral equation

$$X_t = X_0 - \int_0^t X_s\,ds + B_s$$

for all t. We haven't specified the interval on which (X_t) is defined; usually we think of it as an infinite interval, but if it's a finite interval then you just ask this on that finite interval.

§23.2 Existence

Given an SDE, the first thing you ask is:

> **Question 23.2.** Why do solutions to this SDE even exist?

One reason this SDE is very explicit is it's aactually linear. If you ignore the Brownian term, it's exactly just a linear ODE. And if I had a linear ODE like $\dot{x}(t) = -x(t)$, I know how to solve this — we'd just have $x(t) = e^{-t}x(0)$.

More generally, there's this thing called Duhamel's principle where if you have a linear *inhomogeneous* ODE

$$\dot{x}(t) = -x(t) + g(t),$$

it turns out that if you can solve the corresponding *homogeneous* ODE, you can also write out a solution to the inhomogeneous one. What you basically think of is this term $g(t)$ is inserting new initial data at every time t. And if our initial data at time 0 is $x(0)$ we can solve the homogeneous ODE; that kind of leads us to guess that the explicit formula

$$x(t) = e^{-t}x(0) + \int_0^t e^{-(t-s)}g(s)\,ds$$

should work. The idea is the thing is linear, so the solution should be given by adding up a bunch of solutions. The first term corresponds to what you'd have if g didn't exist; and the rest corresponds to inputting a bunch of information from g at different times. If g enters at time s, then you have to solve up to $t - s$, which is why you get the $e^{-(t-s)}g(s)$ term.

And you can check by explicit computation that indeed this explicit formula does solve your ODE. We're not going to do that, because we're going to check it for the SDE, and the proof is very similar.

But the point is if you have a linear inhomogeneous ODE, you can just write down the solution.

Now we have a SDE — g is kind of this Brownian forcing. So you kind of can guess a format for the solution:

> **Proposition 23.3**
> Let $X_0 \in \mathcal{F}_0$. Then the process
>
> $$X_t = e^{-t}X_0 + \sqrt{2}\int_0^t e^{-(t-s)}\,dB_s$$
>
> is an OU process.

(You want your process to be adapted, so you certainly want X_0 to be $\mathcal{F}_0$-measurable.) We basically define things in exactly the above form, except htat we replace g by $\sqrt{2}\,dB_t$; and this is now a stochastic integral.

Another way to put it is that if I specify my initial data and view this as an evolution equation, I can write down an explicit solution, which is just this. This is why this kind of SDE is very special — it's linear, so we can just write down an explicit solution.

Proof. We basically just want to apply Ito's formula — we want to compute dX_t, and writing out the definition, this is

$$dX_t = d\left(e^{-t}X_0 + \sqrt{2}\int_0^t e^{-(t-s)}\,dB_s\right).$$

The first term is FV, so it behaves as ordinary calculus, and we get $-e^{-t}X_0$.

The way to think about the second term is really that it's the product of two terms — it's e^{-t} times a stochastic integral. So we can write that term as

$$\sqrt{2}d\left(e^{-t}\int_0^t e^s\,dB_s\right).$$

And now we see this is a product of two semimartingales (really a FV process and a martingale). And the way to think about it is when you have a product, you put the d on either term; so by Ito, we get

$$-e^{-t}X_0+\sqrt{2}(de^{-t})\int_0^t e^s\,dB_s+\sqrt{2}e^{-t}d\left(\int_0^t e^s\,dB_s\right)$$

(in the general case we have a term where we put a d on both factors, but here the first is a FV process, so we don't need to — you only need that when both have a martingale part). We can group the first two terms to get

$$-\left(e^{-t}X_0+\sqrt{2}\int_- 0^t e^{-(t-s)}\,dB_s\right).$$

And the last term is d of a stochastic integral, so that just gives the integrand

$$\sqrt{2}e^{-t}e^t\,dB_t.$$

And then you kind of explicitly see that this is what I defined X_t as — the first term is $-X_t$, and the second term is $\sqrt{2}\,dB_t$. And that's precisely the definition of an OU process. □

So indeed, this specific formula for X does solve the SDE. (This is also how you check the ODE example with Duhamel's principle — the Ito correction term doesn't show up in this case, so you're actually behaving like ordinary calculus.)

As a consequence of this, we can define X_t for any time, so actually global solutions always exist. That's again a consequence of the fact that the SDE is linear — linear SDEs will always have global solutions (they don't blow up in finite time).

§23.3 Uniqueness

So we've at least shown solutions exist. But it's also a natural question whether this is unique:

Question 23.4. If I specify the initial data, is this the only solution I have?

That's the next claim — that this is the only solution you can have. For a general SDE, usually you have to take advantage of some sort of Lipschitz condition on the right-hand side, and you want to set up some sort of contraction mapping argument. But linear equations are kind of special and you can just explicitly verify things.

Proposition 23.5

Suppose that (X_t) is an OU process. Then

$$X_t=e^{-t}X_0+\sqrt{2}\int_0^t e^{-(t-s)}\,dB_s.$$

So that's the claim — actually the solution is unique.

Proof. To prove this, you can do something slightly clever — again you use Ito's formula, and you compute the evolution of $e^t X_t$. Let's apply Ito's formula to this; then you get

$$d(e^t X_t) = d(e^t) \cdot X_t + e^t \cdot dX_t = e^t(X_t + dX_t).$$

And then plugging in the definition of an OU process, we get

$$d(e^t X_t) = e^t X_t - e^t X_t + \sqrt{2} e^t \, dB_t.$$

And conveniently, the first two terms cancel. So that tells you actually $e^t X_t$ is going to be given by a stochastic integral — this implies

$$e^t X_t = e^0 X_0 + \sqrt{2} \int_0^t e^s \, dB_s$$

for all t. Now you move the e^t to the RHS and get precisely what we claimed. □

So this formula is the only solution you have to this SDE.

§23.4 Distributions

Now in what follows:

Notation 23.6. Let (X_t^x) be an OU process starting at x. In other words,

$$X_t^x = e^{-x} x + \int_0^t e^{-(t-s)} \, dB_s.$$

A corollary of this explicit formula is that:

Corollary 23.7
We have $X_t^x \sim \mathcal{N}(e^{-t}x, 1 - e^{-2t})$.

Proof. Here you use the fact that whenever your integrand is deterministic, your stochastic integral is just a normal random variable — so we'll have

$$\int_0^t e^{-(t-s)} \, dB_s \sim \mathcal{N}(0, \sigma^2),$$

where the variance is just given by the variance of the LHS, which we can compute by Ito's isometry —

$$\sigma^2 = 2\mathbb{E}\left[\left(\int_0^t e^{-(t-s)} \, dB_s\right)^2\right] = 2\mathbb{E}\left[\int_0^t e^{-2(t-s)} \, ds\right].$$

And now there's no more expectation. (In general you need the expectation because the QV term could be random; but the QV of BM is just s, which is deterministic.) And so you can remove the expectation and evaluate this integral, and it will be $1 - e^{-2t}$.

(We forgot to write some factors of $\sqrt{2}$, but this is one of the reasons for it — having it makes this variance be $1 - e^{-2t}$.) □

The fact that this is Gaussian — there's a problem on the HW saying that if you have a deterministic function, then this integral is the same as your white noise evaluated on that function, which is a Gaussian. Another way to see this is to use the discrete approximation to the Ito integral — you have the convergence of approximations. And at the discrete level, this is a linear combination of independent Gaussians, which is always Gaussian. And the limit in distribution of a sequence of Gaussians has to be Gaussian, so you just have to compute its variance.

That's one corollary. And as a result, you see that:

Corollary 23.8

For any x, we have $X_t^x \xrightarrow{\text{dist}} \mathcal{N}_d(0, \text{id})$.

Note that B is a d-dimensional BM, so X is also d-dimensional (in general it's a vector-valued process). And in the limit in distribution, you just get some iid standard Gaussians. That's why we want the $\sqrt{2}$ factor — without it you'd converge to $\mathcal{N}(0, \frac{1}{2})$ instead of $\mathcal{N}(0, 1)$.

Remark 23.9. The other way you can get $\mathcal{N}(0, 1)$ is if you put a 2 in front of the X (and make the other $\sqrt{2}$ disappear); then you can do the same computations and you should get $\mathcal{N}(0, 1)$.

§23.5 Relation to Markov processes

The meta-thing to take away is that whenever you have a solution to a SDE, that's going to be a Markov process, for very good reasons.

Proposition 23.10

Let (X_t) be an OU process. Then (X_t) is a Markov process whose semigroup is given by

$$(\mathcal{P}_t f)(x) = \mathbb{E}[f(e^{-t}x + \sqrt{1 - e^{-2t}}Z)]$$

where the expectation is with respect to $Z \sim \mathcal{N}_d(0, \text{id})$.

You should think of $(\mathcal{P}_t f)(x)$ as $\mathbb{E}[f(X_t)]$ when X starts at x. And we just saw that when X starts at x, you have this explicit formula for the law of X_t. So you would guess that the semigroup should be given by the above thing.

If we want, we can also be more explicit and write this out as an integral, by writing out what is the Gaussian density for Z (the reason you multiply by the square root is so that the whole thing has variance $1 - e^{-2t}$). So if we write this out, we get

$$\frac{1}{\sqrt{2\pi(1 - e^{-2t})}^d} \int dy\, f(y) \exp\left(\frac{|y - ee^{-t}x|^2}{1 - e^{-2t}}\right).$$

It's maybe more immediate to write it as

$$\frac{1}{\sqrt{2\pi(1 - e^{-2t})}^d} \int dy\, f(e^{-t}x + y) \exp\left(-\frac{|y|^2}{1 - e^{-2t}}\right).$$

Proof. We use the fact that you solve the SDE, and you can write the integral form of the SDE starting at s instead of 0 — we have

$$X_t = e^{-(t-s)}X_s + \sqrt{2}\int_s^t e^{-(t-u)}\, dB_u.$$

(We're rewriting the above general formula, you just kind of replace time 0 with time s — if you view time s as the initial time, you still solve the same exact SDE, so we have the same formula just with these shifted times.) And the extra Brownian part, by independence of Brownian increments, is independent of $\mathcal{F}_s$; while the first term $e^{-(t-s)}X_s$ is adapted to $\mathcal{F}_s$. This allows us to compute $\mathbb{E}[X_t \mid \mathcal{F}_s]$ — you basically think of it as adding the extra term.

Note that $\sqrt{2}\int_s^t e^{-(t-u)}\,dB_u \sim \mathcal{N}_d(0, e^{-2(t-s)}\,\mathrm{id})$, because you're looking at the difference of your endpoints to get this $t-s$. This means

$$\mathbb{E}\left[f(X_t) \mid \mathcal{F}_s\right],$$

when you condition on $\mathcal{F}_s$ you can treat the $e^{-(t-s)}X_s$ term in the beginning as constant. So you're going to get exactly

$$(\mathcal{P}_{t-s}f)(X_s).$$

Why? Basically you're fixing X_s, so you kind of start with this constant thing $e^{-(t-s)}X_s$, and then you add this Gaussian noise; and that's exactly how the semigroup was defined. So this verifies X is a Markov process whose semigroup is this $\mathcal{P}_t$. □

Again, we'll emphasize there's many ways you can get a Markov process, but one of the main ways is by solving SDEs; that's a source of many examples of Markov processes. If you look into the history, this might be why Ito developed Ito calculus — he was trying to understand these Markov processes (or something like that).

And you see if you solved a different SDE, you still expect you'd have a Markov process, but maybe the semigroup is not as explicit — maybe for a different SDE you don't have this very explicit way of writing down your solution. But it's also clear that a Markov process just says if you're trying to understand where to go next, you just need to look at your current time. And that's exactly what a SDE is — you specify the evolution of X based on where it is at the current time, plus some Brownian noise. So that's intuitively why you expect SDE solutions to be MPs. (You probably have to make some assumptions to make this a formal statement, e.g., that the SDE has global solutions.)

One exercise that will be on the homework:

> **Exercise 23.11.** $(\mathcal{P}_t)$ is Feller.

Recall this means that if $f \in C_0(\mathbb{R}^d)$ (meaning it's continuous and decays to 0 at ∞) then so is $\mathcal{P}_t f$, and secondly that you have uniform convergence at small times

$$\|f - \mathcal{P}_t f\| \to 0$$

as $t \to 0$ (where the norm is the sup norm).

You have this very explicit formula for the semigroup, so this is just a technical argument. Continuity is quite direct — basically, your explicit formula is continuous in X_0 (if you let X_0 vary, that thing varies continuously — the right-hand term doesn't depend on the initial data at all). That'll imply $\mathcal{P}_t f$ is continuous. You also have to say why it decays to 0 at ∞; Sky didn't feel like covering it in class, so it's on the homework. The second thing says no matter where you start, at small times you expect to be close to where you started (uniformly in the starting point). That's also maybe quite direct, but we won't go into it.

But we have a Feller semigroup, so now we can talk about things like the generator. This is the main point for why we talked about the Markov process stuff — it comes up when you consider these SDEs.

§23.5.1 The invariant measure

Before that, let's talk about the invariant measure — it turns out that's basically the standard Gaussian.

Proposition 23.12

The standard Gaussian $\mathcal{N}_d(0, \mathrm{id})$ is the invariant measure for $(\mathcal{P}_t)$.

Recall this means that if your initial data is distributed as a standard Gaussian, then at all times you're distributed as a standard Gaussian.

Proof. Suppose $X_0 \sim \mathcal{N}_d(0, \mathrm{id})$. Then from the explicit formula — or applying the corollary — we see that

$$X_t \stackrel{d}{=} e^{-t} X_0 + \sqrt{1 - e^{-2t}} Z,$$

where $Z \sim \mathcal{N}_d(0, \mathrm{id})$ is independent of X_0. (This comes because these Brownian increments are independent of $\mathcal{F}_0$.) And the variances add up — if I want to compute a variance of a sum of independent Gaussians, I have to square the variances and add them. So this is equal in distribution to Z.

Another way to say this is that if we let γ_d be the probability density of the standard Gaussian, if we consider

$$\mathbb{E}[f(X_t)] = \int dx\, \gamma_d(x)(\mathcal{P}_t f)(x)$$

(you think of $\mathcal{P}_t f$ as $\mathbb{E}[f(X_t) \mid X_0 = x]$; and now if I'm integrating against γ_d, I'm basically integrating over all possible values of x, where X_0 has law given by γ_d). And the statement is that this is exactly equal to

$$\int \gamma_d(x) f(x) = \mathbb{E}[f(X_0)].$$

So there's two different ways you can think about invariance — either in terms of the process, or the measure. □

Like mentioned before, you should think of the OU process as the SDE version of a standard Gaussian; they're linked because the standard Gaussian is the invariant measure for the corresponding MP. And because we know everything about standard Gaussians, we know everything about the OU process — for instance, we can write down a solution.

We'll see this satisfies very nice properties, like rapid convergence to equilibrium. If you've taken a class on discrete time discrete state-space Markov chains, the point of that theory is to say under certain conditions, your MC converges in distribution to the equilibrium/invariant measure (e.g., if it's aperiodic and irreducible). You should think of this as continuous-time continuous-state space MC theory. Usually you want to understand how fast you converge to equilibrium. In discrete time maybe there are some conditions about rapid convergence. Something special about OU is that it actually converges exponentially quickly to the standard Gaussian.

But of course, in terms of actual applications, you don't care about sampling standard Gaussians using a Markov chain (you can just do that directly), but rapid convergence is at least theoretically quite important.

§23.6 The generator

Now let's talk about the generator of this Markov process. Recall last time we talked about how you would compute the generator; basically you want to find this martingale. So let's apply this to the OU process.

Suppose $f \in C_b^2(\mathbb{R}^d)$ is twice continuously differentiable, and it and its first and second derivatives are bounded. Last time we did this for BM, now we'll do it for the OU process. Given X, you apply Ito's formula to get

$$df(X_t) = \nabla f(X_t)\, dX_t + \frac{1}{2} \partial_{ij} f(X_t)\, \langle dX_t^i, dX_t^j \rangle.$$

And then if we write otu the SDE for X, X is an OU process so dX is specified. We get

$$\sqrt{2}\nabla f(X_t)\,dB_t - X_t\nabla f(X_t)\,dt + \frac{1}{2}\partial_{ij}f(X_t)\langle\sqrt{2}dB_t^i, \sqrt{2}\,dB_t^j\rangle$$

(the first term is the martingale part; the second is one FV thing; and to compute the QV terms in the end, you only care about the martingale part of X, which corresponds to the dB part — because when you input the FV part into the QV you just get 0).

And these QVs are 0 unless $i = j$, and the $\sqrt{2}$'s cancel out the $1/2$ factor, and in the end when we sum in i and j and use the fact that this is nonzero only when $i = j$, we get the Laplacian; so this becomes

$$\sqrt{2}\nabla f(X_t)\,dB_t + (-X_t\nabla f(X_t) + \Delta f(X_t))\,dt.$$

And this holds for any OU process — any of the X_t^xs for example.

So we can apply the result from last time (since f and its derivatives are bounded, this martingale part is actually a martingale, since the integrand is bounded) — f minus its FV part gives me a martingale

$$f(X_t) - \int_0^t (\Delta f(X_s) - X_s\nabla f(X_s))\,ds.$$

This means $f \in D(L)$, and the generator evaluated at f is going to be the right function g — so

$$(Lf)(x) = (\Delta f)(x) - x\nabla f(x).$$

So that's how you compute the generator of a SDE giving you a Markov process — you just apply Ito's formula.

Here's one proposition that appears quite often in other settings; let's first see it in the explicit OU setting.

We showed that

$$Lf = \Delta f - x\cdot\nabla f,$$

and from now on, when we write L, we mean this operator.

Proposition 23.13

As an operator, L is a symmetric, negative-semidefinite operator on $C_b^2(\mathbb{R}^d) \subseteq L^2(\gamma_d)$.

What does this mean? This means if I compute the L^2 inner product with respect to the standard Gaussian measure (which we denote by γ_d — the measure of the standard d-dimensional Gaussian) — that gives me a Hilbert space, and this says

$$\langle Lf, g\rangle_{L^2(\gamma_d)} = \langle f, Lg\rangle_{L^2(\gamma_d)}.$$

That's what symmetric means; and negative semidefinite means

$$\langle Lf, f\rangle_{L^2(\gamma_d)} \le 0.$$

This is somehow the continuous version of what it means to be a reversible Markov process. In discrete time discrete state space MCs, if you satisfy the detailed balance equations you get a reversible chain; this first statement is somehow the continuous analog of detailed balance and reversibility.

Proof. Let's first try to show it's symmetric, so we're computing this inner product

$$(Lf, g)_{L^2(\gamma_d)} = \int dx\,\gamma_d(x)(Lf)(x)g(x).$$

And we substitute in the explicit formula we had for Lf and get

$$\int dx\, \gamma_d(x)(\Delta f(x) - x\nabla f(x))g(x).$$

And now you want to integrate by parts. If I integrate by parts the Laplacian term, I'm going to get

$$-\int dx \nabla f(x) \cdot \nabla(\gamma_d(x)g(x))$$

(this term comes when I integrate by parts $\Delta f(x) \cdot \gamma_d \cdot g$ — I take one derivative on the Laplacian and move it to the other two functions). And then I have the other two terms

$$-\int dx \gamma_d(x)(x \cdot \nabla f(x))g(x).$$

(With this integration by parts, we're applying the general formula $\int \Delta hg = -\int \nabla h \cdot \nabla g$, where these integrals are against the Lebesgue measure). Now we have to compute $\nabla(\gamma_d g)$, which we can do using the product rule; we have

$$\nabla(\gamma_d g) = (\nabla \gamma_d)g + \gamma_d \nabla g.$$

And now we use the fact that we can explicitly compute $\nabla\gamma_d$, and it's going to be precisely $-x \cdot \gamma_d$ (this is all evaluated at a point x). So then you see that actually when the gradiant falls on the Gaussian density, it's going to cancel out the third term — this first term comes with a $-$, which cancels out this $-$, and then you get exactly the last term $\int dx\gamma_d(x)(x \cdot \nabla f(x))\, gx$. So if you take into account this cancellation, you're just left with whatever you get from $\gamma_d \nabla g$. And that's going to be

$$-\int dx \gamma_d(x) \nabla f(x) \cdot \nabla g(x).$$

So what we actually managed to show is that

$$(Lf, g)_{L^2(\gamma_d)} = -(\nabla f, \nabla g)_{L^2(\gamma_d)}.$$

And you see that's why you get symmetry — the thing on the right is manifestly symmetric in f and g. So if you swap the roles of f and g, you get the symmetry statement.

You also immediately get the negative-semidefinite statement — if $g = f$, this thing is $\nabla f \cdot \nabla f$, which is always nonnegative. □

Remark 23.14. All these computations hold in much more generality, but we're doing it explicitly in this Gaussian case first; if we have time we'll later talk about more general measures and more general SDEs.

Remark 23.15. The way you compute $\nabla\gamma_d$ is that up to constants this density is $\exp(-|x|^2/2)$; the constants don't matter for evaluating the gradient, so you just have to evaluate the gradient of this guy, and that's what we want.

§23.7 Symmetry and invariance

This symmetry gives you another way to prove invariance of the Gaussian measure with respect to the OU semigroup (which generalizes) — it's a general fact that if you have symmetry of your operator with respect to the L^2 space of a given measure, then the semigroup L generates is going to leave this measure invariant

(recall that the operator determines the semigroup). So if some operator L is symmetric with respect to the L^2 space of a given measure, then that measure has to be invariant.

In the Gaussian case we did this by an explicit computation; but for more general SDEs you don't have explicit formulas, so you have to use this more general statement. So let's talk about that.

So we'll give another proof of invariance using only that

$$(lf, g)_{L^2(\gamma_d)} = (f, Lg)_{L^2(\gamma_d)}.$$

We'll also have to use some other general facts about Markov processes — in particular, last time we discussed that

$$\partial_t \mathcal{P}_t f = L\mathcal{P}_t f,$$

at least for $f \in D(L)$. In integral form, this says

$$\mathcal{P}_t f(x) = f(x) + \int_0^t L\mathcal{P}_s f(x)\, ds.$$

Now let's integrate with respect to the measure γ_d; we have

$$\int dx \gamma_d(x)(\mathcal{P}_t f)(x).$$

Eventually we want to show that this is just $\int \gamma_d(x) f(x)$ — we want to take out the $\mathcal{P}_t$. For that, if we insert this formula, we get

$$\int dx \gamma_d(x) f(x) + \int dx \gamma_d(x) \int_0^t (L\mathcal{P}_s f)(x)\, ds.$$

The first term is exactly what we want, so we have to somehow say the second term gives 0. Why? The first thing you do is swap the order of integration (over space and time) and write this as

$$\int_0^t ds \int dx \gamma_d(x)(L\mathcal{P}_s f)(x)\, ds.$$

And you can interpret this as the L^2 inner product of $L\mathcal{P}_s f$ against the constant function 1, so this is

$$\int_0^t ds (L\mathcal{P}_s f, \mathbf{1})_{L^2(\gamma_d)}.$$

And now you use Gaussian integration by parts. So using the symmetry (maybe we need slightly more than symmetry; you need to know what the generator looks like, that it's a differential operator). Using the symmetr, you can move the L to the other side to get

$$\int_0^t ds (\mathcal{P}_s f, L\mathbf{1})_{L^2(\gamma_d)}.$$

And we know L just involves taking derivatives of your function, and $\mathbf{1}$ is a constant, so its derivatives are 0; so this whole term is 0, and you get the claimed invariance.

Remark 23.16. In general your operator *is* going to be differential (you'll always be taking derivatives), so this statement about $L\mathbf{1} = 0$ is going to be true.

§23.8 Reversibility

Now let's talk about reversibility. Here's one way you can say reversibility. All these things will be needed later whne we talk about exponential convergence to equilibrium — this fact about generators, this integration by parts formula that we proved ($(Lf, g)_{L^2(\gamma_d)} = -(\nabla f, \nabla g)_{L^2(\gamma_d)}$).

Proposition 23.17

Let $X_0 \sim \gamma_d$. Then for all t, we have

$$(X_0, X_t) \stackrel{d}{=} (X_t, X_0).$$

This says the joint law of X_0 and X_t is the same as X_t and X_0. One main consequence of this is that $\mathcal{P}_t$ itself is symmetric —

$$(\mathcal{P}_t f, g)_{L^2(\gamma_d)} = (f, \mathcal{P}_t g)_{L^2(\gamma_d)}.$$

Remark 23.18. We proved the earlier identity for everything in the domain of L, but you want it for all bounded continuoous f. This is a general fact that the domain of L is going to be dense in the space of bounded continuous functions. Here at least we know it contains $C_b^2(\mathbb{R}^d)$. And this definitely would be enough to characterize the law. You just need some large enough space of functions so that if you know the expectation of every one of those functions against your measure, then you know the law of the measure; but this is usually going to be true.

The way we're going to prove this is a bit special to the Gaussian case (the way you'd prove it in general is by Girsanov, and if we have time we'll get to that, but for now we want to get to this rapid convergence to equilibrium thing, so we'll just prove it directly using some special facts about Gaussians and the OU process).

Student Question. *Proving it in general means for a different process with different invariant measure?*

Answer. Yes. The OU process is a concrete example of a Longevin dynamic, where basically given any measure you can write down an SDE which leaves it invariant, just like if I give you a measure on a finite set I can find a Markov chain that leaves it invariant; in the continuous case you can find a Markov process that does it. Such dynamics will also be reversible, and to prove that you have to use Girsanov's theorem. This thing, even the OU process, is something of modern interest, even in *diffusion models* in CS (maybe).

Proof of proposition. From the explicit formula for $\mathcal{P}_t$ — let's suppose we're in one dimension (otherwise it's a bunch of iid things, so it extends directly; but in one dimension the notation is slightly simpler). First of all, (X_0, X_t) is jointly Gaussian. This is kind of a consequence of your explicit representation of

$$X_t = \bullet \cdot X_0 + \text{independent Gaussian noise}.$$

Any time you have that, your pair is going to be jointly Gaussian.

And this is in two dimensions, so if we compute the law of this joint Gaussian, the mean is just 0 because they're both standard normals. And the covariacne matrix; on the diagonals it's 1, because they're standard normals. And what is the covariance? YOu see that

$$X_t = e^{-t} X_0 + \text{independent Gaussian},$$

so the covariance is just e^{-t}. But then very manifestly, the joint law

$$\mathcal{N}_2\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & e^{-t}\\ e^{-t} & 1\end{bmatrix}\right)$$

is exchangable, i.e., if (Y, Z) is distributed according to this thing, then (Y, Z) and (Z, Y) have the same distribution.

In d dimensions, it's just a bunch of i.i.d. collections of this. □

Now let's prove the symmetry of the semigroup:

Proposition 23.19
We have $(\mathcal{P}_t f, g)_{L^2(\gamma_d)} = (f, \mathcal{P}_t g)_{L^2(\gamma_d)}$.

Proof. Writing out what this means,

$$(\mathcal{P}_t f, g)_{L^2(\gamma_d)} = \int dx \gamma_d(x) \mathbb{E}[f(X_t) \mid X_0 = x] g(x)$$

(just writing out what $\mathcal{P}_t$ is by definition). And you can write this as

$$\mathbb{E}[f(X_t)g(X_0)],$$

because x means we're taking X_0. But then using this reversibility of X_0 and X_t, we can swap their roles and say this is

$$\mathbb{E}[f(X_0)g(X_t)].$$

And by the same identity here, this is going to be equal to $(f, \mathcal{P}_t g)_{L^2(\gamma_d)}$. So this proves the symmetry of this semigroup. □

At a very formal level, if the generator is symmetric in this sense, you expect the semigroup to be — last time we said that at least formally you should think of $\mathcal{P}_t$ as e^{tL}. And at least in finite dimensions, if you have a symmetric matrix, then its matrix exponential is also going to be symmetric. So formally that's why you believe $\mathcal{P}_t$ should be symmetric if the generator is.

§23.9 Rates of convergence

That concludes the preliminaries, so now we can begin discussing rates of convergence.

The main result that we're leading towards, at least one of the main intermediate ones (we're going to use this to prove rapid convergence to equilibrium, but it's the main intermediate step and also of its own interest):

Theorem 23.20 (Gaussian Poincaré inequality)
For any f such that $\nabla f \in L^2(\gamma_d)$, we have

$$\mathrm{Var}_{\gamma_d}[f] \le (\nabla f, \nabla f)_{L^2(\gamma_d)}.$$

By $\mathrm{Var}_{\gamma_d}[f]$, we mean $\mathrm{Var}[f(X)]$ for $X \sim \gamma_d$.

So you're bounding the second moment of a function by its gradient. You can also think of this as a spectral gap statement (it should be equivalent to the MP having a spectral gap, and this should be why you get exponentially fast convergence, though we won't really talk about that).

So this is what we're leading towards. We'll probably prove it next time. This is some functional inequality. For Sky the main reason it's interesting is it will lead to rapid convergence to equilibrium, but there are also other applications of this kind of thing.

To begin, one first step is the following commutation identity.

Lemma 23.21 (Commutation identity)
Let $f \in C_b^1(\mathbb{R}^d)$. Then

$$\nabla \mathcal{P}_t f = e^{-t} \mathcal{P}_t \nabla f.$$

This means f and its first derivative are continuous and bounded.

So you can commute the gradient and your semigroup, and when you do that, you pickk up this exponentially decreasing term. This is the source of your exponential convergence, as we will see.

The proof — and this inequality — is quite special to the Gaussian distribution.

Remark 23.22. Often measures which satisf this kind of inequality will have nice convexity properties. For example, if the density of your measure is log-concave, that's very good for sampling. More generally, if you have log-concave density, you'd expect to have Poincare and rapid mixing. This is because you can efficiently minimize a convex function; but if it's not convex and has local minima that might be hard. And this is related to sampling from densities with good convexity properties.

Proof. We can write out explicitly what $(\mathcal{P}_t f)$ is; we have

$$(\mathcal{P}_t f)(x) = \mathbb{E}[f(e^{-t} + \sqrt{1 - e^{-2t}} Z)]$$

where $Z \sim \mathcal{N}_d(0, \mathrm{id})$. Now if we take a gradient in x, at least formally — and this is maybe the source of the regularity assumption — you should be able to put the gradient inside the expectation, so you get

$$(\nabla \mathcal{P}_t f)(x) = \mathbb{E}\left[\nabla f(e^{-t} x + \sqrt{1 - e^{-2t}} Z)\right].$$

But this is a function of x where you first multiply it by e^{-t}, so by the chain rule this is

$$e^{-t} \mathbb{E}[(\nabla f)(e^{-t} + \sqrt{1 - e^{-2t}} Z)] = e^{-t} \mathcal{P}_t \nabla f.$$

The point is you're differentiating in x, but x gets hit by this e^{-t}, and when you differentiate you get this extra e^{-t} factor. (We're not going to worry about the exchanging the expectation and gradient; we'll just assume that's fine.) □

Using this you can immediately prove the Gaussian Poincaré inequality. Next time we'll do that and use it to show rapid convergence to equilibrium with respect to a certain distance on measures. It's the χ^2 distance. It actually controls the TV distance, so this is quite strong. After we discuss that, we'll talk about entropy. There's a notion of relative entropy which is also a distance between probability measures in some sense (though it's not symmetric). Surprisingly you actually get exponential decay in relative entropy with respect to the standard Gaussian. This actually relates to this thing called log Sobolev inequalities, which is also a topic of modern interest.

§24 May 5, 2025

Last time we began discussing the OU process

$$dX_t = -X_t + \sqrt{2}\, dB_t,$$

a very explicit SDE. The point is it's very nicely behaved. One reason is the fact that it's linear; but also, you have this $-X_t$ damping term. This basically says you have an autoregressive process — any time it gets big, it tends to come back down to $O(1)$ magnitude (if X is ever big, the $-X_t$ term dominates your dynamics). So the tendency is always to push you back to something $O(1)$. That's why this process not only doesn't explode in finite time, but is also quite nicely behaved.

There's a meta thing about how any time you have a SDE solution, this should define a MP — infinitesimally this tells you where I am next should only depend on where I am now, plus some noise. So we can calculate its generator, using Itô's formula, as

$$(Lf)(x) = \Delta f(x) - x \cdot \nabla f(x).$$

Then we computed an integration by parts relation — L is symmetric with respect to the L^2 inner product (with respect to the standard Gaussian measure), i.e.,

$$(Lf, g)_{L^2(\gamma_d)} = -(\nabla f, \nabla g)_{L^2(\gamma_d)} = (f, Lg)_{L^2(\gamma_d)}.$$

And finally, we ended by proving the commutation identity

$$\nabla \mathcal{P}_t f = e^{-t} \mathcal{P}_t \nabla f.$$

The point is you can explicitly write down how $\mathcal{P}_t$ acts — you're adding in some explicit Gaussian noise — so you can do this computation. This is also why you get rapid convergence to equilibrium — suppose f is bounded and has bounded first derivative. Then $\mathcal{P}_t f$ is also going to be bounded. And you're basically saying its gradient decays exponentially in time. So as you apply this semigroup $\mathcal{P}_t$, basically it converges to a constant function, because the gradient decays exponentially (and if the gradient is 0, you're a constant). And also we know, what's the constant you should be converging to? It's

$$\mathcal{P}_t f \to \mathbb{E}_{\gamma_d} f$$

(because last class, we proved that as $t \to \infty$, the process converges to the standard Gaussian). This is the reason why you get rapid convergence — it's basically saying as you apply this semigroup $\mathcal{P}_t$ to a function, that function becomes basically constant exponentially quickly, in the sense that you'll have an exponentially decaying bound on its gradient.

So this is the key thing; and today we'll see how we actually use it to get convergence to equilibrium, with respect to two notions of distance.

§24.1 Poincare inequality

Theorem 24.1 (Poincare)
We have $\mathrm{Var}_{\gamma_d}(f) \le (\nabla f, \nabla f)_{L^2(\gamma_d)}$.

(There are various conditions on f — at the very least, both sides have to be finite — but at least for nice enough functions, this is the statement.)

This is actually just a statement about the standard Gaussian density — there's no dynamical thing in here. But the way we'll prove this is by using the dynamics of the OU semigroup. The relation is that γ_d is the

invariant measure for the dynamics. So we'll introduce the dynamics, which will somehow let us prove this inequality for the invariant measure.

Why introduce the dynamics? It's very nicely behaved; in particular, rapid convergence in the form of this commutation identity will be the key step.

Before we prove this, we'll need a lemma.

Lemma 24.2 (Variance decay)

For nice enough (e.g., smooth) f, we have

$$\partial_t(\mathcal{P}_t f, \mathcal{P}_t f)_{L^2(\gamma_d)} = -2(\nabla \mathcal{P}_t f, \mathcal{P}_t f)_{L^2(\gamma_d)}.$$

So if you compute the time-derivative of the second moment of f, you get something that's always decaying.

(We can always extend things by density arguments, so we might as well assume f is smooth and bounded and so on.)

This is some sort of monotonicity formula — the second moment of $\mathcal{P}_t f$ with respect to γ_d is actually *decaying*.

Proof. The proof is a calculation. We won't keep writing $L^2(\gamma_d)$ in the notation; we'll just make that implicit. How do you compute

$$\partial_t(\mathcal{P}_t f, \mathcal{P}_t f)?$$

We use the product rule. The time derivative could fall on either of the $\mathcal{P}_t$'s, so we get

$$(\partial_t \mathcal{P}_t f, \mathcal{P}_t f) + (\mathcal{P}_t f, \partial_t \mathcal{P}_t f) = 2(\partial_t \mathcal{P}_t f, \mathcal{P}_t f)$$

(using the symmetry of the inner product). And we have this equation that $\mathcal{P}_t f$ satisfies in t — when you differentiate in t, you just get L. So this is actually

$$2(L\mathcal{P}_t f, \mathcal{P}_t f)$$

(where L is the generator). And then you use integration by parts, and this becomes $-2(\nabla \mathcal{P}_t f, \mathcal{P}_t f)$, which is what we wanted. □

Actually if you use this, you can prove the Poincare inequality. The proof Sky is presenting here is slightly different from in the notes (there's a supplement he'll upload soon with this one), because this one is more conceptual (and similar to how we'll prove a later inequality called log Sobolev).

Proof. You use this monotonicity formula. First we can assume $\mathbb{E}_{\gamma_d} f = 0$ (the formula doesn't change if you shift f by a constant, so we're free to do that), so that $\mathrm{Var}_{\gamma_d}(f) = (f, f)$. (Again, we're not writing $L_2(\gamma_d)$, but that's what we mean by this inner product.) Now we can integrate this monotonicity formula to get

$$(\mathcal{P}_0 f, \mathcal{P}_0 f) - (\mathcal{P}_t f, \mathcal{P}_t f) = -\int_0^t \partial_s(\mathcal{P}_s f, \mathcal{P}_s f)\, ds.$$

So we view $(\mathcal{P}_t f, \mathcal{P}_t f)$ as a function in t, and we just write this as an integral of its time derivative. And the point is I have a formula for this time derivative, and if I substitute in this formula, I get

$$2\int_0^t (\nabla \mathcal{P}_s f, \nabla \mathcal{P}_s f)\, ds.$$

But you also have that if you send $t \to \infty$, then $\mathcal{P}_t f \to \mathbb{E}_{\gamma_d} f$, which we assumed is 0. So actually $\mathcal{P}_t f \to 0$ pointwise as $t \to \infty$. So if you take the limit as $t \to \infty$ of both sides, you're basically going to get that

$$(\mathcal{P}_0 f, \mathcal{P}_0 f) = 2 \int_0^\infty (\nabla \mathcal{P}_s f, \nabla \mathcal{P}_s f)\, ds.$$

And the left-hand side is just (f, f), because $\mathcal{P}_0$ is the identity operator. (You kind of have to justify why if $\mathcal{P}_t f \to 0$ pointwise, its second moment goes to 0; but that's part of the regularity assumptions — if you assume f is C^1 and bounded, then this should just be bounded convergence, so that's fine.)

Now here's the key step. You use the special property satisfied by the OU semigroup — that $\mathcal{P}_t f$ is converging to a constant function exponentially quickly (in terms of the gradient), and that constant is going to be 0 (because we assumed $\mathbb{E}_{\gamma_d} f = 0$). So we use the commutation formula (swapping $\nabla \mathcal{P}_s f$ to $\mathcal{P}_s \nabla f$ and picking up an exponential) to write this as

$$2 \int_0^\infty e^{-2s} e^{-2s} (\mathcal{P}_s \nabla f, \mathcal{P}_s \nabla f)\, ds.$$

Now you're almost done; you just need to apply Cauchy–Schwarz in the right way. For fixed s, we want to say that

$$(\mathcal{P}_s \nabla f, \mathcal{P}_s \nabla f) = \|\nabla \mathcal{P}_s \nabla f\|^2_{L^2(\gamma_d)} \le \|\nabla f\|^2_{L^2(\gamma_d)}.$$

What we want to claim is that $\mathcal{P}_s$ is a contraction on $L^2(\gamma_d)$ — it can't increase the L^2 norm. And why is that true? This is somehow supposed to be Hölder's inequality, in a way — if I take $|(\mathcal{P}_s g)(x)|^2$, well, this is least confusing if you think about it in terms of probability theory. (We'll say $g = \nabla f$, but we'll prove it in general.) The way to think about

$$\int |(\mathcal{P}_s g)(x)|^2 \, \gamma_d(x)\, dx$$

is that it's

$$\mathbb{E}[\mathbb{E}[g(X_s) \mid X_0]^2].$$

Because the way you think of this semigroup $\mathcal{P}_s$ is in terms of the stochastic process

$$(\mathcal{P}_s g)(x) = \mathbb{E}[g(X_s) \mid X_0 = x]$$

(that's the interpretation of this semigroup in terms of the process). And now we're also averaging over your initial position, and we're saying X_0 has the standard Gaussian as its law; that's why you can interpret the L^2 norm of $\mathcal{P}_s g$ as the expectation of the square of the conditional expectation. And now you use Jensen's (for instance) to take the square inside the conditional expectation and get

$$\mathbb{E}[\mathbb{E}[g(X_s) \mid X_0]^2] \le \mathbb{E}[\mathbb{E}[g(X_s)^2 \mid X_0]].$$

And by tower, this is just $\mathbb{E}[g(X_s)^2]$. But then here we started in stationarity — we were saying X_0 has standard Gaussian density — so X_s is also distributed as X_0, which means we could also write this as $(g, g)_{L^2(\gamma_d)}$.

This may have been good to make a lemma — the key point is that as an operator, $\mathcal{P}_s$ is always a contraction on L^2. So if we use this contraction property, we'll basically be done. The point is that ∇f on the right-hand side doesn't depend on s — we basically used stationarity to remove the s-dependence. So now we're just left with an explicit integral — we can bound everything by

$$2 \int_0^\infty e^{-2s}\, ds \cdot (\nabla f, \nabla f)_{L^2(\gamma_d)}.$$

And $2 \int_0^\infty e^{-2s}\, ds = 1$, so that concludes the proof. □

Remark 24.3. Actually the only place you had an inequality in all this was basically when we used this contraction. The commutation identity is actually an equality. More generally, you see you don't actually need an equality — if your semigroup satisfies the *inequality*

$$\|\nabla \mathcal{P}_t f\|_{L^2(\gamma_d)} \le e^{-t} \|\mathcal{P}_t \nabla f\|_{L^2(\gamma_d)},$$

then you can prove this in the same way. For OU it turns out to be an equality; but if for some other semigroup you're somehow able to verify this inequality, you'd also have Poincare for its invariant measure. So if you had som eother general semigroup, in general its invariant measure wouldn't be Gaussian; but you're basically saying if the dynamics mixes rapidly, then the measure itself also satisfies a nice functional inequality (namely Poincare).

(You would replace all instances of γ_d with the invariant measure.)

§24.2 Evolution of densities

Now let's talk about why this implies rapid mixing. Suppose that you start your OU process at an initial distribution which has a density with respect to γ_d; let f_0 be that density, which we'll write as

$$f_0 = \frac{d\mathsf{Law}(X_0)}{d\gamma_d}$$

(the Radon–Nikodym derivative of $\mathsf{Law}(X_0)$ with respect to the standard Gaussian).

Claim 24.4 — The law of X_t also has a density with respect to γ_d (for all t), and moreover, this density f_t is given by

$$f_t = \frac{d\mathsf{Law}(X_t)}{d\gamma_t} = \mathcal{P}_t f_0.$$

You can put some sort of distances on probability measures (e.g., total variation distance; the one we'll use is χ^2, which actually controls TV). We'll show that under some notions, this distance converges exponentially quickly under $\mathcal{P}_t$. The first step is saying that if you have a density at time 0, you'll also have a density at later times (because the distance will be defined out of the densities).

Proof. We need to be able to compute $\mathbb{E}[g(X_t)]$ for some bounded measurable g; and if we're saying we have a density with respect to γ_d, I should be able to express this as an integral with respect to γ_d, times this density. So how are we going to do that?

The first thing to note is that you can write this in integral form as

$$\int \gamma_d(x) f_0(x) (\mathcal{P}_t g)(x)$$

(when we have $\gamma_d(x) f_0(x)$, we're basically integrating with respect to the original law of X_0).

(An intermediate step would have been to first write $\mathbb{E}[g(X_t)] = \mathbb{E}[\mathbb{E}[g(X_t) \mid X_0]]$, which is then equal to this integral.)

And the reason you write it llike this is now we can view this as

$$(f_0, \mathcal{P}_t g)_{L^2(\gamma_d)}.$$

And last time we proved a statement about the symmetry of $\mathcal{P}_t$ as an operator — you can actualy move $\mathcal{P}_t$ to the LHS, so this becomes

$$(\mathcal{P}_t f_0, g)_{L^2(\gamma_d)},$$

which you can now interpret as

$$\int \gamma_d(x) g(x) (\mathcal{P}_t f_0)(x).$$

So we've been able to express expectations with respect to X_t as integrals with respect to γ_d times this extra thing. This is a way of showing the law of X_t is absolutely continuous with respect to γ_d, and the density is precisely this extra term $\mathcal{P}_t f_0$. □

All of this is just so that we can put this into the framework we currently have. We're going to be using the variance decay proposition, which tells us that if you apply $\mathcal{P}_t$ to some f, its variance decays. In our setup, we're eventually going to take f to be f_0; that's what we're working towards.

§24.3 More about variance decay

This is kind of going slightly out of order, but maybe we should've said this before: A corollary of variance decay and the Poincare inequality is that actually the variance decays exponentially.

Corollary 24.5
If $\mathbb{E}_{\gamma_d} f = 0$, we have $(\mathcal{P}_t f, \mathcal{P}_t f)_{L^2(\gamma_d)} \leq e^{-2t} (f, f)_{L^2(\gamma_d)}$.

If you combine the variance decay inequality with the Poincare inequality, you get this.

Proof. Combining variance decay and Poincare, we have

$$\partial_t (\mathcal{P}_t f, \mathcal{P}_t f) = -2(-\mathcal{P}_t f, \nabla \mathcal{P}_t f)_{L^2(\gamma_d)} \leq -2(\mathcal{P}_t f, \mathcal{P}_t f)_{L^2(\gamma_d)}.$$

And this exactly implies exponential decay — for example, assuming everything is well-defined, this implies that

$$\partial_t \log(\mathcal{P}_t f, \mathcal{P}_t f) \leq -2.$$

But then if you integrate this and then exponentiate, you get precisely what we wanted. Again, this is kind of under the assumption that $(\mathcal{P}_t f, \mathcal{P}_t f)$ never becomes 0; but this is an instance of Gronwall's lemma more generally, that if you have a function in time satisfying this ordinary differential inequality, then you can integrate it to get this sort of exponential decay. □

Again, the key point about this variance decay is — and we'll see this again when talking about entropy and log-Sobolev — is that variance decay is an actual identity, so to get exponential decay you only need a lower bound on the right-hand side by the thing you're differentiating. And such a lower bound is precisely the statement of Poincare.

§24.4 Chi squared distance between distributions

Now we'll come back to our question:

Question 24.6. How close in distribution are you getting to the standard Gaussian?

So let's define a notion of distance.

Definition 24.7. Given probability measures μ and ν on $\mathbb{R}^d$ such that ν is absolutely continuous with respect to μ (written $\nu \ll \mu$), we define

$$\chi^2(\nu \mid \mu) = \operatorname{Var}_\mu\left(\frac{d\nu}{d\mu}\right).$$

You can write this more explicitly as

$$\int d\mu(x)\left(\frac{d\nu}{d\mu}(x) - 1\right)^2.$$

Why? The variance of a random variable is $\mathbb{E}[(X - \mathbb{E}[X])^2]$. And we have

$$\mathbb{E}_\mu\left[\frac{d\nu}{d\mu}\right] = \int d\mu \cdot \frac{d\nu}{d\mu} = \int d\nu(x) = 1$$

(by the definition of density). So that's why you have this formula.

And with this formula, you can kind of see that this is actually going to control total variation: we have

$$d_{\mathsf{TV}}(\nu, \mu) = \int \left|\frac{d\nu}{d\mu} - 1\right| d\mu(x)$$

(this is one way to define it; the more common way is to assume both μ and ν have a density with respect to Lebesgue).

The main point is not to talk about the details of TV; but it's a common notion of distance between two probability measures which is quite strong (if TV is small, the two distributions are quite close in a quantitative sense). And if you just apply Cauchy–Schwarz, this is bounded by

$$\left(\int \left|\frac{d\nu}{d\mu} - 1\right|^2 d\mu(x)\right)^{1/2} = \chi^2(\nu \mid \mu)^{1/2}.$$

So the point is that the χ^2 distance (which you may or may not have seen before) controls TV distance. (But that's the main point of this computation; otherwise we won't directly use it.)

§24.5 Convergence in chi squared

But now the point is that if we combine all this discussion, we get that

$$\chi^2(\mathsf{Law}(X_t) \mid \gamma_d) = \chi^2(\mathcal{P}_t f_0 \mid \gamma_d),$$

since we're in the setting where we assumed X_0 has a density, and derived that the density exactly evolves as $\mathcal{P}_t$ so the density of X_t is $\mathcal{P}_t$ of the original.

And by the definition of the χ^2 distance, this is exactly

$$\operatorname{Var}_{\gamma_d}(\mathcal{P}_t f_0).$$

And now you can apply Corollary 24.5, though we need to be a bit careful. In the end, what we want to say is that this is at most $e^{-2t}\operatorname{Var}_{\gamma_d}(f_0) = e^{-2t}\chi^2(\mathsf{Law}(X_0) \mid \gamma_d)$. And the reason why this will be true is, we want to reduce to Corollary 24.5. The variance is not the second moment if you don't have zero mean, but the point is we could have written

$$\operatorname{Var}_{\gamma_d}(f_0) = (f_0, f_0) - 1$$

(the variance is the second moment minus the first; and you have a density, so its first ,oment will always be 1). Similarly, we always have

$$\mathrm{Var}_{\gamma_d}(\mathcal{P}_t f_0) = (\mathcal{P}_t f_0, \mathcal{P}_t f_0) - 1.$$

And the identity we want to claim follows from these identities (the fact that this -1 is the same in both cases) and Corollary 24.5.

The main point is that Poincare basically leads to rapid decay of this type of distance between probability measures — in a very quantiative form, you converge to stationarity exponentially quickly. To summarize, this really follows from Poincare, which really follows from this commutation identity, which is why the OU process is quite nice.

Student Question. *Doesn't Corollary 24.5 only hold if $\mathbb{E}f = 0$? How are we applying it to f_0?*

Answer. We want to apply it to $\widetilde{f_0} = f_0 - 1$, which does have mean 0. Then you get

$$\mathrm{Var}_{\gamma_d}(\mathcal{P}_t f_0) = \mathrm{Var}_{\gamma_d}(\mathcal{P}_t \widetilde{f_0}) \le e^{-2t}(\widetilde{f_0}, \widetilde{f_0}) = e^{-2t}\chi^2(\mathsf{Law}(X_0) \mid \gamma_d)$$

(in the definition of the χ^2 distance, you always subtract 1).

§24.6 Entropy and log-Sobolev

Now we'll talk about another notion of distance for which you have this exponential decay.

This notion of entropy is quite important — it shows up in many areas of math. For instance, anything related to physics (it came from thermodynamics). In particular, it shows up in analysis, if you talk about PDEs which model the evolution of fluids, basically there are natural notions of entropy which come out. It also shows up in statistics and probability because it's a notion of a measure of randomness. Interestingly, it actually shows up in geometry as well — if you look at Perelman's proof of the Poincare conjecture, he had a sequence of papers, and in the first his main breakthrough was that he identifies a new controlled quantity. Ricci flow is some nonlinear evolution equation in geometry, and because it's nonlinear, it's very complicated. But his first breakthrough is that even though this is super nonlinear, you can identify certain controlled quantities — along this flow, there are certain statistics that are decreasing. And this controlled quantity he identified has to do with entropy. So he identified something in terms of entropy that got monotonicity and allowed him to prove the conjecture.

So entropy is quite important in many areas of math. Let's define it.

Definition 24.8 (Relative entropy)**.** Let μ and ν be probability measures on $\mathbb{R}^d$ such that ν is absolutely continuous with respect to μ. Define the relative entropy

$$\mathcal{D}(\nu \mid \mu) = \int f \log f \, d\mu,$$

where $f = \frac{d\nu}{d\mu}$.

Remark 24.9. This is also called Kullblack–Leibler divergence.

Remark 24.10. At a discrete level, if μ is a discrete measure, this can also be called Shannon entropy, which shows up in information theory.

Notation 24.11. For nonnegative $f \geq 0$, define

$$\mathsf{Ent}_{\gamma_d}(f) = \mathbb{E}_{\gamma_d}[f \log f] - \mathbb{E}_{\gamma_d} f \log \mathbb{E}_{\gamma_d} f.$$

The point is in general f doesn't have to be a density — it doesn't have to integrate to 1 — which is why you have the extra term. If f does integrate to 1 (i.e., $\mathbb{E}_{\gamma_d} f = 1$), then $\log 1 = 0$, so the second term drops out and

$$\mathsf{Ent}_{\gamma_d}(f) = \mathcal{D}(f\gamma_d \mid \gamma_d).$$

(By $f\gamma_d$, we mean the measure which has density f with respect to γ_d.)

Goal 24.12. Understand how the entropy evolves along the OU process — how the function $t \mapsto \mathsf{Ent}_{\gamma_d}(\mathcal{P}_t f_0)$ behaves.

The previous stuff was basically understanding the evolution of the variance — variance decay and Poincare let us understand the evolution of the variance. Now we have a different statistic, the entropy; and we want to understand its evolution as well.

The punchline is that this also decays exponentially as well (with a statement very similar to the one we had for variance).

To prove that exponential decay, we proved two things — a variance decay lemma, and Poincare. To get this, we need an entropy decay lemma, and some replacement for Poincare, which is called *log Sobolev*.

§24.7 Preliminaries

Before we get into this, we'll make a preliminary scaling observation about entropy — if you scale f by a constant, how does that change things? This should just scale the entropy by a constant, i.e.,

$$\mathbb{E}_{\gamma_d}(cf) = c\mathsf{Ent}_{\gamma_d}(f).$$

THe reason is if we compute this out, we get

$$\mathbb{E}_{\gamma_d}(cf \log(cf)) - \mathbb{E}_{\gamma_d}(cf) \log \mathbb{E}_{\gamma_d(cf)}$$

(just writing out the definition of the entropy of the scaled function). We can expand out log of a product as a sum of logs, so the first term becomes the two terms

$$c\mathbb{E}_{\gamma_d}[f \log f] + c\mathbb{E}_{\gamma_d}[f \log c].$$

Similarly, the second term expands out to

$$-(c\mathbb{E}_{\gamma_d} f \log \mathbb{E}_{\gamma_d} f + c\mathbb{E}_{\gamma_d f} \log c).$$

Two of these cancel, and we just get $c\mathbb{E}_{\gamma_d} f$.

This means we can without loss of generality assume that f has mean 1 (scaling it to have mean 1 just picks up a constant, which you can track).

§24.8 An entropy evolution identity

We want to understand how the entropy behaves as a function of t.

Lemma 24.13

We have

$$\partial_t \mathsf{Ent}(\mathcal{P}_t f) = -\mathbb{E}_{\gamma_d}\left[\frac{|\nabla \mathcal{P}_t f|^2}{\mathcal{P}_t f}\right].$$

Remark 24.14. (We only ever apply entropy to nonnegative functions, or else log doesn't make sense.)

This maybe looks complicated, but note that the right-hand side (without the $-$) is always nonnegative. So $\mathsf{Ent}(\mathcal{P}_t f)$ does decay along $\mathcal{P}_t$! This is not obvious at all.

The next question is how rapidly it decays, but that's for later. The fact that the time derivative is always nonpositive tells you that the entropy is always non-increasing — so if you measure distance from stationarity using entropy, it never increases (only decreases). That's not obvious at all, but this calculation will tell you that indeed it's decaying.

Remark 24.15. If $\mathbb{E}_{\gamma_d} f = 1$, then $\mathcal{P}_t f \to 1$ as $t \to \infty$, and thus we actually get

$$\mathsf{Ent}(\mathcal{P}_t f) \to 0$$

as $t \to \infty$. Again you have to justify this using some sort of convergence theorem, but $\mathsf{Ent}(1) = 0$ (because you're comparing the standard Gaussian density to itself; and if the entropy is any reasonable notion of distance and you compare something to itself, you should just get 0).

So this is telling you (modulo some sort of convergence theorem) that entropy converges to 0 as time goes to ∞. This is reasonable because we expect to converge to stationarity. The more surprising thing is the first thing — that it's monotone (it never increases, only decreases).

Proof. The proof is actually not that involved — you again use integration by parts. In this example, we can assume $\mathbb{E}_{\gamma_d} f = 1$, because if we scale f so that it has mean 1, we pick up a constant on the LHS and the same constant on the RHS (if we scale f by c, we get a c^2 in the numerator and a c on the bottom). So we can without loss of generality scale so that f has mean 1 (and the general case follows by multiplying by a constant). This is convenient because in $\mathsf{Ent}_{\gamma_d}(f)$ you lose the second term (which is 0).

Mow we can compute

$$\partial_t \mathsf{Ent}_{\gamma_d}(\mathcal{P}_t f) = \partial_t \mathbb{E}_{\gamma_d}(\mathcal{P}_t f \log \mathcal{P}_t f).$$

Now we again use the product formula — the time-derivative can hit either $\mathcal{P}_t f$ (in which case you get L, since that's the time derivative of $\mathcal{P}_t$) or $\log \mathcal{P}_t f$, in which case you have to differentiate log; so you get

$$\mathbb{E}_{\gamma_d}[L\mathcal{P}_t f \log \mathcal{P}_t f] + \mathbb{E}_{\gamma_d}\left[\mathcal{P}_t f \cdot \frac{L\mathcal{P}_t f}{\mathcal{P}_t f}\right]$$

(because log gives you a $\frac{1}{x}$, and then the time derivative of $\mathcal{P}_t f$ is again $L\mathcal{P}_t f$).

But we claim the second term is just 0. Why? The $\mathcal{P}_t f$'s cancel, so this is

$$\mathbb{E}_{\gamma_d}[L\mathcal{P}_t f] = (1, L\mathcal{P}_t f)_{L^2(\gamma_d)}.$$

But you can integrate by parts to get that this is

$$-(\nabla 1, \nabla \mathcal{P}_t f)_{L^2(\gamma_d)} = 0$$

(the gradient of 1 is 0).

Now with the first term, we want to view this as an inner product and integrate by parts; so we write this as

$$(L\mathcal{P}_t f, \log \mathcal{P}_t f) = -\left(\nabla \mathcal{P}_t f, \frac{\nabla \mathcal{P}_t f}{\mathcal{P}_t f}\right)$$

(again using the gradient of log).

And this is another way to write the expression we wanted on the RHS, so we're done. □

Remark 24.16. You see that ignoring questions of interchanging derivatives and integrals, the computation is quite simple — you basically apply the same thing (integration by parts) over and over again.

Now how are we going to get exponential decay of the entropy? You want to lower-bound the expectation on the right (without the $-$ sign) by the thing on the left:

Goal 24.17. To get exponential decay of $\mathsf{Ent}_{\gamma_d}(\mathcal{P}_t f)$, we want an inequality that looks like

$$\mathbb{E}_{\gamma_d}\left[\frac{|\nabla \mathcal{P}_t f|^2}{\mathcal{P}_t f}\right] \leq C\mathsf{Ent}_{\gamma_d}(\mathcal{P}_t f).$$

If you had an inequality like this for all f (where C is a constant not dependingon f), you could substitute this in and get

$$\partial_t \mathsf{Ent}(\mathcal{P}_t f) \leq -C\mathsf{Ent}_{\gamma_d}(\mathcal{P}_t f),$$

and as soon as you get this, you get exponential decay.

So the only question now is, do you have an inequality of this form? This is precisely log-Sobolev, which basically says indeed you do. So this is one way you can motivate log-Sobolev (and it's how Sky prefers to think about it).

Remark 24.18. In statistics, the RHS of the lemma is often called the *Fisher information*.

§24.9 Log-Sobolev inequality

So this log-Sobolev is what we're going to proceed to prove next.

Theorem 24.19 (Log-Sobolev inequality)

For any $f \geq 0$, we have

$$\mathsf{Ent}_{\gamma_d}(f) \leq \frac{1}{2}\mathbb{E}_{\gamma_d}\left[\frac{|\nabla f|^2}{f}\right] \leq$$

Remark 24.20. Why is it called log-Sobolev? In Sobolev inequalities, you're bounding L^p norms of your function by L^p norms of the gradient (or derivative). Here you're bounding the L^1 norm of $f \log f$ by the gradient.

Remark 24.21. A key thing is that the constant does not depend on your dimension — it's just $\frac{1}{2}$ (though everything is in $\mathbb{R}^d$). This is in sharp contrast to the usual Sobolev inequalities. Usually their precise statements heavily depend on the dimension. For instance, in two dimensions you might have

$$\|f\|_{L^p(\mathbb{R}^2)} \lesssim \|f\|_{H^1(\mathbb{R}^2)},$$

where $\|f\|_{H^1} = \|f\|_{L^2} + \|\nabla f\|_{L^2}$. So you can bound higher L^p norms by the L^2 norm and the L^2 norm of a gradient, for any $p < \infty$. But this is only true in 2 dimensions — in 3 dimensions this only works for $p \leq 6$, and in 4 dimensions for $p \leq 4$.

But for log-Sobolev, this is true in *any* dimension. This is important — log-Sobolev originally arose in quantum field theory, which in some sense is analysis in infinite dimensions. So while Sobolev inequalities are about functions on finite-dimensional space, QFT arises out of physics and is supposed to describe protons and neutrons and electrons and quarks and somehow when you try to make this mathematically rigorous, you need some sort of analysis; and log-Sobolev was originally discovered in that context.

And it's very important because this $\frac{1}{2}$ doesn't depend on dimension.

In particular, Perelman's first paper cites Len Gross (1970's); this may be the first place log-Sobolev inequalities were introduced, and he motivated them because he was motivated by QFT. Perelman cited this because found some entropy-like quantity that's decaying.

(Len Gross is a mathematician at Cornell.)

To prove this, you kind of just want to mimic the proof of Poincare.

Proof. By the entropy decay lemma, you can write

$$\mathsf{Ent}_{\gamma_d}(f) - \mathsf{Ent}_{\gamma_d}(\mathcal{P}_\infty f)$$

(we know the latter is 0, because at time ∞ you converge to 0) as an integral of the time-derivative, which we computed; so this is

$$\int_0^\infty dt \mathbb{E}_{\gamma_d}\left[\frac{|\nabla \mathcal{P}_t f|^2}{\mathcal{P}_t f}\right]$$

(in the proof of Poincare we first said this for finite t and took $t \to \infty$; here we're just skipping that step). So again you're using special properties of your dynamics of the OU process to prove a statement about the measure, which has nothing to do with OU — you're just introducing this process to help you prove the inequality. And the reason it's helpful is that the process satisfies this commutation relation; we can commute the gradient and semigroup and pick up an e^{-2t}. So this becomes

$$\int_0^\infty dt e^{-2t} \mathbb{E}_{\gamma_d}\left[\frac{|\mathcal{P}_t \nabla f|^2}{\mathcal{P}_t f}\right].$$

Like in Poincare, we want to find some inequality which completely removes the dependence on t, using a clever application of Cauchy–Schwarz. For this, we use that

$$|\mathcal{P}_t \nabla f| = \left|\mathcal{P}_t\left(\frac{\nabla f}{\sqrt{f}}, \sqrt{f}\right)\right|.$$

The point of all of this is that $\mathcal{P}_t$ is an expectation with respect to a probability measure, so you can apply Cauchy–Schwarz at that level to get that this is at most

$$\left|\mathcal{P}_t\left(\frac{|\nabla f|^2}{f}\right)\right|^{1/2} (\mathcal{P}_t f)^{1/2}.$$

A more expliict way to write this is that

$$|\mathcal{P}_t \nabla f| = \mathbb{E}[\nabla f(X_t) \mid X_0 = x] = \mathbb{E}\left[\frac{\nabla f(X_t)}{\sqrt{f(X_t)}}\sqrt{f(X_t)} \mid X_0 = x\right].$$

And now I use Cauchy–Schwarz for this conditional expectation, so this becomes

$$\mathbb{E}\left[\frac{|\nabla f(X_t)|^2}{f(X_t)} \mid X_0 = x\right]^{1/2} \cdot \mathbb{E}[f(X_t) \mid X_0 = x]^{1/2}.$$

So actually this inequality holds for all x.

Now if you use this in our integral, the $\mathcal{P}_t f$'s cancel out, and our integral is bounded by

$$\int_0^\infty dt\, e^{-2t} \mathbb{E}_{\gamma_d}\left[\left|\mathcal{P}_t\left(\frac{|\nabla f|^2}{f}\right)\right|\right].$$

And you don't need the absolute value because everything is positive.

Now you use stationarity — $\mathcal{P}_t$ is stationary with respect to γ_d, so the expectation of $\mathcal{P}_t$ is actually just the expectation, and this is

$$\mathbb{E}_{\gamma_d}\left[\frac{|\nabla f|^2}{f}\right].$$

And $\int_0^\infty e^{-2t}\, dt = \frac{1}{2}$. □

Remark 24.22. The original proof of log-Sobolev was not this; later on people found clever proofs like this.

Next time we will wrap up log-Sobolev, and then discuss a different consequence — 'hypercontractivity' — and then we'll begin discussing more general dynamics.

§25 May 7, 2025

§25.1 Review

Last time, we defined this notion of entropy with respect to the standard Gaussian, as

$$\mathsf{Ent}_{\gamma_d}(f) = \mathbb{E}_{\gamma_d}[f \log f] - \mathbb{E}_{\gamma_d} f \log \mathbb{E}_{\gamma_d} f.$$

We computed that along the OU semigroup, the entropy is monotone decreasing, with

$$\partial_t \mathsf{Ent}_{\gamma_d}(\mathcal{P}_t f) = -\mathbb{E}_{\gamma_d}\left[\frac{|\nabla \mathcal{P}_t f|^2}{\mathcal{P}_t f}\right]^2.$$

Then we wanted to say it decays *exponentially* quickly. To conclude that, we want a bound on the RHS in terms of the entropy itself. That's what log-Sobolev is — we proved that

$$\mathsf{Ent}_{\gamma_d}(f) \le \frac{1}{2}\mathbb{E}_{\gamma_d}\left[\frac{|\nabla f|^2}{f}\right].$$

This was by combining the above identity with the commutation identity $\nabla \mathcal{P}_t f = e^{-t}\mathcal{P}_t \nabla f$. This is really what makes the OU process special — when you apply your semigroup to a Lipschitz function, its Lipschitz constant decreases exponentially, which is what lets you get all these rapid mixing results.

As a consequence, we get that

$$\partial_t \mathsf{Ent}_{\gamma_d}(\mathcal{P}_t f) \le -2\mathsf{Ent}_{\gamma_d}(\mathcal{P}_t f).$$

And this is precisely the differential inequality which gives you exponential decay

$$\mathsf{Ent}_{\gamma_d}(\mathcal{P}_t f) \le e^{-2t}\mathsf{Ent}_{\gamma_d}(f).$$

And now that we have this exponential decay, if you start the OU process at $X_0 \sim f_0$, we saw that $X_t \sim \mathcal{P}_t f_0$. So then you can apply this entropy decay to say that the law of X_t converges to γ_d exponentially quickly (at least, when measured in terms of this relative entropy distance). That's something very nice about the OU process.

Let's also stress that interestingly, this log-Sobolev inequality is just a fact about the Gaussian measure, but the way you proved it was introducing a dynamic process (the OU flow). The reason you want to use the dynamics is because you're able to directly prove it's rapidly mixing, basically, through the gradient commutation identity.

§25.2 Hypercontractivity

That kind of wraps up the foray into log-Sobolev inequalities. Well... we won't say wraps up. There's another consequence of log-Sobolev we'll talk about today, called *hypercontractivity*.

To set the stage, recall that $(\mathcal{P}_t)$ is a semigroup of *contractions* on $L^p(\gamma_d)$. We stated it for L^2 (we needed this somewhere in the proof of the log-Sobolev inequality), but more generally it's true for all p. The way you see this is it ultimately comes from the fact that $\mathcal{P}_t$ applied to a function can be interpreted as averaging that function with respect to a probability measure.

Fact 25.1 — For any t, $\mathcal{P}_t$ is a contraction on L^p (for any $1 \le p \le \infty$).

Proof. We can write

$$\|\mathcal{P}_t f\|_{L^p(\gamma_d)}^p = \int \gamma_d(x)\,|\mathcal{P}_t f(x)|^p\,dx = \int \gamma_d(x)\mathbb{E}[f(X_t) \mid X_0 = x]\,dx.$$

And if you use Hölder for expectations, you can get an upper bound on this by taking the pth power inside, getting that this is at most

$$\int \gamma_d(x)\mathcal{P}_t(|f|^p)(x)\,dx.$$

And then you use invariance — $(\mathcal{P}_t)$ is the OU semigroup and γ_d is its invariant measure, so this is the same as just $\int \gamma_d\,|f(x)|^p\,dx = \|f\|_{L^p(\gamma_d)}^p$. □

This is true not just of the OU semigroup, but any Markov semigroup — a Markov semigroup is basically a thing where applying the semigroup corresponds to integrating f along some probability measure.

The interesting thing about OU is that it satisfies something even stronger, called *hyper*contractivity — it turns out you can bound a *higher* L^p norm of $\mathcal{P}_t f$ in terms of the L^p norm of f. Specifically, you can show

$$\|\mathcal{P}_t f\|_{L^q(\gamma_d)} \lesssim \|f\|_{L^p(\gamma_d)}$$

for $q > p$. This is very non-obvious, and not true in general — it's really using something special about the OU semigroup (and really, the thing you use is rapid decay, in the sense of log-Sobolev).

The precise statement of this result:

Theorem 25.2 (Gaussian hypercontractivity)

For all $1 \le p < \infty$, for all $t > 0$, there exists $q_t(p) > p$ such that

$$\|\mathcal{P}_t f\|_{L^{q_t(p)}(\gamma_d)} \le \|f\|_{L^p(\gamma_d)}.$$

(If $p = \infty$, then you can bound everything by L^∞, so that case is not interesting.)

We'll define $q_t(p)$ in the proof; it'll be greater than p for all $t > 0$ (for $t = 0$, it'll be exactly p). You'll see it increases with t — if you run the OU longer and longer times, you get better and better integrability. You should think this because the semigroup has a regularizing effect — both interms of derivatives (you see the derivative decays exponentially quickly) and also integrability (the integrability part is what this is saying — when you smooth out using the semigroup, you gain more additional integrability than you started with).

Remark 25.3. We won't see many applications of hypercontractivity in this class, but it's also of modern interest and comes up in many contexts.

Proof. Define $h(t) = \|\mathcal{P}_t f\|^{q_t(p)}_{L^{q_t(p)}(\gamma_d)}$. As we've done many times by now, we want to understand the evolution of this quantity, which we'll write as $\dot{h}(t)$. So let's compute that. Eventually, we'll show that — once we understand the evolution of $\dot{h}(t)$, we'll really care more about

$$\frac{d}{dt} h(t)^{1/q_t(p)}.$$

The magical thing is that this will be negative (so that this is monotone), which will automatically give the thing we want.

What is $\dot{h}(t)$? Let's first write out the q-norm of $\mathcal{P}_t f$; so this is

$$\dot{h}(t) = \frac{d}{dt} \int dx \gamma_d(x) \left|\mathcal{P}_t f(x)\right|^{q_t(p)}.$$

Now the derivative can hit two parts — the $\mathcal{P}_t f$ or the exponent — so you get two terms. Also, let's just assume f is nonnegative in this proof; we might use that at certain points. You can always upper-bound $\|\mathcal{P}_t f\|$ by $\|\mathcal{P}_t |f|\|$, while $\|f\|$ doesn't change if you replace f by $|f|$, so we can do this.

When it hits $|\mathcal{P}_t f|$, we have to bring one exponent down; so we get

$$q_t(p) \int dx \gamma_d(x) \mathcal{P}_t f(x)^{q_t(p)-1} \partial_t \mathcal{P}_t f(x).$$

This is the first term, when you differentiate the $\mathcal{P}_t f$ part.

The second term is when the derivative hits the exponent. Here the way you should think about this thing is as $\exp((\log \mathcal{P}_t f(x)) q_t(p))$; and right now we're thinking about $\log \mathcal{P}_t f(x)$ as a constant and differentiating the following term. (Also we'll stop writing x.) When you differentiate, you get

$$\int dx \gamma_d \log \mathcal{P}_t f \cdot (\mathcal{P}_t f)^{q_t} \cdot \dot{q}_t$$

(we're also going to stop writing p). Let's call these terms $I_1 + I_2$.

The first thing, which is a priori not obvious, is that — you somehow want to use log-Sobolev, so you want to find some term that looks like the entropy. And that'll come from this second term, since it has a log. As a side calculation, let's compute

$$\mathsf{Ent}_{\gamma_d}((\mathcal{P}_t f)^{q_t}).$$

By definition, this is

$$\int dx \gamma_d (\mathcal{P}_t f)^{q_t} \log (\mathcal{P}_t f)^{q_t} - \int dx \gamma_d (\mathcal{P}_t f)^{q_t} \log \int dx \gamma_d (\mathcal{P}_t f)^{q_t}.$$

For the first term, we can put the q_t in the front. The first term basically looks like the one we wanted, except for the $\dot{q_t}$ vs. q_t. And the second term you recognize — it's actually just $h \log h$, based on how we defined h. So what does this tell us? We want to write the second term in terms of the entropy and $h \log h$; so if we do the algebra correctly, we should get that

$$I_2 = \frac{\dot{q_t}}{q_t} \cdot (\mathsf{Ent}_{\gamma_d}((\mathcal{P}_t f)^{q_t}) + h \log h).$$

So we've successfully introduced this entropy term (eventually we're going to apply log-Sobolev). When we look at log-Sobolev, we're eventually going to upper-bound the entropy by the RHS of log-Sobolev, so we also have to identify that. And it turns out you can actually identify I_1 as basically something like this. Why is that? Well, we have to integrate by parts. As usual, this thing $\partial_t \mathcal{P}_t f(x)$ is going to be $L\mathcal{P}_t f$. And as usual, it's better in these things, whenever you see an L you want to integrate by parts. So you could have thought of I_1 as

$$q_t \left(\mathcal{P}_t f^{q_t - 1}, L\mathcal{P}_t f \right)$$

(we're writing it in this inner product way to apply integration by parts; the inner product is with respect to $L^2(\gamma_d)$). After integration by parts, we get that this is

$$-q_t \left(\nabla (\mathcal{P}_t f)^{q_t - 1}, \nabla \mathcal{P}_t f \right).$$

When you compute the gradient of this power, you get

$$-q_t(q_t - 1)((\mathcal{P}_t f)^{q_t - 2} \nabla \mathcal{P}_t f, \nabla \mathcal{P}_t f)$$

(using the chain rule). Eventually, we're going to want to think of this as an expectation with respect to the standard Gaussian. So we get that this is

$$-q_t(q_t - 1)\mathbb{E}_{\gamma_d} \left[(\mathcal{P}_t f)^{q_t - 2} \cdot |\nabla \mathcal{P}_t f|^2 \right].$$

So that's what we get for I_1; we're probably going to need this later.

So that's I_1. But we wanted to identify something of the form on the RHS of log-Sobolev. We have $(\mathcal{P}_t f)^{q_t}$, so let's try to compute

$$\mathbb{E}_{\gamma_d} \left[\frac{|\nabla (\mathcal{P}_t f)^{q_t}|^2}{(\mathcal{P}_t f)^{q_t}} \right]$$

(this is what will appear in log-Sobolev). When we use the chain rule, you get

$$q_t^2 \mathbb{E}_{\gamma_d} \left[\frac{\left| (\mathcal{P}_t f)^{q_t - 1} \nabla \mathcal{P}_t f \right|^2}{(\mathcal{P}_t f)^{q_t}} \right].$$

Now I'm going to have $2(q_t - 1)$ powers of $\mathcal{P}_t f$ on the top, and q_t on the bottom; so in the end, I get

$$q_t^2 \mathbb{E}_{\gamma_d} \left[(\mathcal{P}_t f)^{q_t - 2} |\nabla \mathcal{P}_t f|^2 \right].$$

This looks exactly like what we got for I_1, ignoring the pre-factors.

So now if we apply log-Sobolev to the entropy term in I_2, we get an upper bound

$$I_2 \leq \frac{\dot{q_t}}{q_t} \left(\frac{1}{2} q_t^2 \mathbb{E}_{\gamma_d} [(\mathcal{P}_t f)^{q_t - 2} |\nabla \mathcal{P}_t f|^2] + h \log h \right).$$

So we've upper-bounded I_2 and introduced this term here, and naturally we want to match it up with the term coming from I_1.

So in summary, what have we found so far? We've found that the time evolution of h is upper-bounded by $I_1 + I_2$, which we have this bound on. Now if I group this first term in the bound with I_1, I should get

$$\dot{h}(g) \leq \left(\frac{\dot{q}_t q_t}{2} - q_t(q_t - 1) \right) \mathbb{E}\left[(\mathcal{P}_t f)^{q_t - 2} |\nabla \mathcal{P}_t f|^2 \right] + \frac{\dot{q}_t}{q_t} h(t) \log h(t).$$

Now let's look at the first term

$$\frac{\dot{q}_t q_t}{2} - q_t(q_t - 1).$$

We're going to choose q so that this thing is 0 — that's where the choice of q comes in. We can write this as

$$q_t^2 \left(\frac{\dot{q}_t}{2q_t} - \frac{q_t - 1}{q_t} \right).$$

And we want q_t such that this yellow thing is just 0; that will help us greatly in understanding the evolution of h.

So we have to solve this kind of ODE — you basically want $\dot{q}_t = 2(q_t - 1)$. It's also natural to impose that $q_0 = p$ — because when $t = 0$, you just have f on the LHS, and you cannot hope to control a higher L^q norm of f by the L^p norm of f. So it's natural to impose that $q_0 = p$, so that initially you're controlling the L^p norm of f by itself, which is of course true.

So you can solve this ODE, and it turns out the explicit form of q is

$$q_t(p) = e^{2t}(p - 1) + 1.$$

You see it is true that this is strictly greater than p (for $t > 0$), because $e^{2t} > 1$.

And if you plug in this choice for q_t and compute the time-derivative, we get

$$\dot{q}_t = 2e^{2t}(p - 1) = 2(e^{2t}(p - 1) + 1) - 2 = 2(q_t - 1).$$

So indeed, with this choice of q_t, this thing is 0.

The point is that you want a nice formula for the evolution of h; so it's natural to choose q so that this complicated-looking term is 0. The other reason is we want to control the RHS in terms of h itself, but it's unclear how the first term relates to h (the one with $\mathcal{P}_t f$). But the second term clearly relates to h — you want to control the evolution of h in terms of h itself when you do these things.

So in summary, we have that

$$\dot{h}(t) \leq \frac{\dot{q}}{q} \cdot h(t) \log h(t).$$

So this is the thing we've arrived at. And now we can go about understanding the evolution of $h^{1/q}$, which is the L^q norm of $\mathcal{P}_t f$ — for

$$\frac{d}{dt} h(t)^{1/q_t},$$

you again use the product rule. Either the derivative hits h, or it hits the exponent; so you get

$$\frac{1}{q_t} h(t)^{1/q_t - 1} \dot{h}(t) + \log h(t) h(t)^{1/q_t} \left(-\frac{\dot{q}_t}{q_t^2} \right)$$

(again, you think about this as $\exp(\frac{1}{q_t} \log h(t))$, and you're thinking of $\log h(t)$ as a constant for the second term).

And now if we plug in what we found for $\dot{h}(t)$, we get

$$\frac{\dot{q}_t}{q_t^2} h(t)^{1/q_t} \log h(t) - \frac{\dot{q}_t}{q_t^2} h(t)^{1/q_t} \log h(t).$$

But this is just 0! And thus we've shown what we wanted — that the evolution of the q-norm of $\mathcal{P}_t f$ is actually monotone decreasing! So integrating this gives the desiredresult. □

Student Question. *Do we have to assume $p > 1$?*

Answer. Yeah, otherwise $q = 1$. The result is true if $p = 1$, but it's not really telling you anything; you only really gain if $p > 1$.

This computation is magical — somehow you use log-Sobolev in the right place and the rest is a bunch of computation, but the computation is quite simple (you just compute a bunch of derivatives and magically a bunch cancel out, partly due to your choice of q, but finally you get this cancellation which is quite nice and gives you monotonicity).

What did we use about the Gaussian measure? You can try to generalize the statement of hypercontractivity to cases where you have a measure and the associated semigroup, for which you satisfy this log-Sobolev inequality. As long as you have that pair, you're going to have that the semigroup satisfies some sort of hypercontractivity condition — because really all you used was log-Sobolev.

So that's that for hypercontractivity. We're probably not going to talk about any applications of this, but if you continue with this part of probability, eventually you will run into it; so it is nice to now know about it.

Remark 25.4. Actually, hypercontractivity and log-Sobolev are equivalent — if you assume you have a hypercontractivity statement, that'll actually imply a log-Sobolev inequality for your measure. This is again not super obvious.

§25.3 Fokker–Planck equation

We'll have one more topic on OU processes, and then we'll move on to Langevin dynamics, which are more general versions of OU. This is again something we won't directly use in this class, but it's good to know.

Recall that if X_0 has some density f_0 with respect to γ_d (we write this as '$X_0 \sim f_0$ with respect to γ_d'), then $X_t \sim \mathcal{P}_t f_0$ with respect to γ_d. The way you derive this is using the reversibility of $\mathcal{P}_t$ — you want to compute the expectation of some function of X_t, and use the fact that $\partial_t \mathcal{P}_t f = L\mathcal{P}_t f$.

This means $\partial_t f_t = L f_t = \Delta f_t - x \cdot \nabla f_t$ — basically, the density of x solves this equation. But if you look on Wikipedia for the evolution equation of the probability density, it's not this one. Why? Here we're looking at the density with respect to the standard Gaussian; usually you talk about the density with respect to Lebesgue. If X has a density with respect to the standard Gaussian then it also has one with respect to Lebesgue; for us it was convenient to use the standard Gaussian (so we could use integration by parts and stuff), but if you actually want to understand the evolution with respect to Lebesgue, that's a different equation; and that'll be Fokker–Planck.

Now let $X_0 \sim g_0$ with respect to the Lebesgue measure. Then X_t will also have a density g_t with respect to the Lebesgue measure (X_t is obtained from X_0 by adding some Gaussian noise, so you can explicitly write down the density given the initial one — so the *existence* of density is not an issue).

Question 25.5. How does g_t evolve in time?

It has to satisfy some sort of equation, so let's derive that.

First, what will the form of g_t look like? We have

$$X_t = e^{-t} X_0 + \sqrt{1 - e^{-2t}} Z$$

(where Z is some independent Gaussian noise). When you add two independent random variables, the density of their sum is the convolution of their densities; so you get

$$g_t = g_0 * \rho_{1-e^{-2t}},$$

where the second thing is the Gaussian density with variance $1 - e^{-2t}$.

So we can explicitly write down the density in this case. In principle you can start differentiating in t and seeing what kind of equation it'll satisfy. But there's a more general way you can go about finding the evolution of g_t, which is — well, we know that if we integrate some function times $g_t(x)$ with respect to the Lebesgue measure, by the definition of g_t as the density, we should get $\mathbb{E}[f(X_t)]$, i.e.,

$$\int g_t(x)\, f(x)\, dx = \mathbb{E}[f(X_t)]$$

(where $X_0 \sim g_0$). The RHS is just $(\mathcal{P}_t f)$ — or more precisely, it's

$$\mathbb{E}[(\mathcal{P}_t f)(X_0)].$$

So thus if we differentiate in time on both sides, we get

$$\partial_t \int g_t(x) f(x)\, dx = \partial_t \int g_0(x)(\mathcal{P}_t f)(x)\, dx$$

(writing out $\mathbb{E}[(\mathcal{P}_t f)(X_0)]$ more explicitly).

I'm doing this because I want to use the fact that I know the time-derivative of $\mathcal{P}_t f$. This is in short going to be a duality argument — somehow the evolution of the density is going to be dual to that of $\mathcal{P}_t f$.

Now if we continue and compute, the RHS is

$$\int g_0(x)(L\mathcal{P}_t f)(x)\, dx.$$

Now, it's actually better in this case to not write it as $L\mathcal{P}_t f$, but as $\mathcal{P}_t L f$ (we proved much earlier that these are the same — you can commute $\mathcal{P}_t$ and L). The point of this is you could've written this as

$$\mathbb{E}\left[(\mathcal{P}_t L f)(X_0)\right]$$

(the same reason we could go from $\mathbb{E}[(\mathcal{P}_t f)(X_0)]$ to an integral). And now I can go back to thinking about this as an expectation

$$\mathbb{E}[Lf(X_t)]$$

(because $\mathcal{P}_t$ is a Markov process, so $\mathcal{P}_t L f$ has an interpretation as a conditional expectation given X_0).

And I did all this so that I get an integral with respect to g_t — this is

$$\mathbb{E}[Lf(X_t)] = \int g_t(x)(\Delta f - x \cdot \nabla f(x))\, dx.$$

So in the end, we get a long chain of equalities saying

$$\int \partial_t g_t(x)\, f(x)\, dx = \int g_t(x)(\nabla f - x \cdot \Delta f(x))\, dx.$$

Now, we want to integrate by parts to move all the derivatives onto g_t — we want to write the RHS as

$$\int (\cdots g_t) f(x)\, dx$$

(where $\cdots$ is some differential operator). The point is once you have this identity that testing $\partial_t g_t$ against f is equal to some differential operator tested against f, by some density argument you're going to get that $\partial_t g_t$ has to be whatever's in the $\cdots g_t$ term. So that's what we're working towards.

Okay, so what is this going to look like? The Laplacian we can move over with no problem — when you take $(g, \Delta f)$, that's just $(\Delta g, f)$. So you get a Δg_t term.

What about the second? We claim that in the end you should get

$$\int (\Delta g_t + \operatorname{div}(x g_t)) f(x)\, dx,$$

i.e., when you integrate by parts the gradient term, you get this $\operatorname{div}(x g_t)$ term.

So let's write this out; we get

$$\int g_t(x) x_j \partial^j f(x)\, dx$$

(using the notation where repeated indices are summed over). Nwo you can move the derivative over and get that this is

$$\int \partial^j (g_t(x) x_j) f(x)\, dx$$

(we don't worry about boundary terms because we're assuming f is compactly supported). And this is precisely

$$\int \operatorname{div}(g_t(x) x) f(x)\, dx.$$

So the final identity we get is

$$\int (\partial_t g_t) f(x)\, dx = \int (\Delta g_t + \operatorname{div}(x g_t)) f dx.$$

There's an additional step where you can write this as a pure divergence term, since the Laplacian is the divergence of the gradient; so you could write this as

$$\int \operatorname{div}(\nabla g_t + x g_t)\, dx$$

(this is just a cosmetic thing, but often Fokker–Planck will be written in this form).

So assuming you have some sort of density argument or pairing argument, you should basically be getting that

$$\partial_t g_t = \Delta g_t + \nabla(x g_t) = \operatorname{div}(\nabla g_t + x g_t).$$

This is the Fokker–Planck equation, and this is the one that's usually used when you talk about the evolution of the density of the OU process (or a more general solution to a SDE) — they usually talk about it as the evolution of the density with respect to Lebesgue, and in that case you get this type of divergence equation.

For calculations it is more convenient to use the density with respect to your invariant measure, because then you can do integration by parts.

There's a very good reason why the divergence has to be here. Why should the evolution of a probability density be given by something with a divergence? If you take $f = 1$, well, you know that

$$0 = \partial_t \int g_t(x)\, dx.$$

Now, if you use the equation satisfied by g, you get that this is

$$\int \operatorname{div}(\nabla g_t + x g_t)\, dt.$$

But this is a pure divergence; you can think of it as $1 \cdot \operatorname{div}(\nabla g_t + x g_t)$, and you can imagine integrating by parts so that this becomes $(\nabla 1) \cdot (\nabla g_t + x g_t)$, which is just 0. So that's basically why it has to be some divergence thing — you have to satisfy this condition that when you test against 1, you just get 1.

§25.4 Langevin dynamics

We've finished the OU section, so we'll now start on a more general topic of Langevin dynamics, which are a generalization of the OU process. They're given by SDEs of the form

$$dX_t = -\nabla V(X_t)\,dt + \sqrt{2}\,dB_t.$$

As usual, we keep the $\sqrt{2}$ normalization, but you can also take it out if you want; it just affects what exactly is the invariant measure, as we saw in the OU case.

Associated to SDEs of this form, you should have some measure, which is basically supposed to be given as follows. We'll define it as a density with respect to the Lebesgue measure, and

$$d\mu_V(x) = \frac{1}{Z_V} e^{-V(x)}\,dx,$$

where $Z_V = \int e^{-V(x)}\,dx$ is just some normalizing constant so that this is a probability measure. (You need some conditions on V so that this is finite, but let's assume you have those.)

The point is that this measure should be invariant for the dynamics, at least under some regularity assumptions on V. So the point is you write down this measure, and then you write down the dynamics which leave this measure invariant, and that's where the Langevin dynamics come from.

Remark 25.6. They're named after Paul Langevin; if you're in Paris you can go to the place where all the famous French people are buried (like Marie Curie), and he's also buried there.

You should think of this as the continuous time continuous state space analog of Markov chains. There, you have a measure on some finite discrete state space, and you want to write down a dynamic that samples from that measure — a Markov chain with that as your invariant measure — and then run ten billion steps of your MC to get samples. This is the continuous analog — you start with a measure and you can try to study it using the dynamics. If you can prove rapid mixing about the dynamics, you can then prove things about the measure. We saw this with OU — we easily showed the dynamics are rapidly mixing, and that got us Poincare and log-Sobolev for the measure.

The main example we've seen is when $V(x) = \frac{|x|^2}{2}$. When you do that, μ_V is precisely γ_d, and the Langevin dynamics are the OU process (because ∇V becomes x).

Remark 25.7. We're usually always working in $\mathbb{R}^d$; when you replace $\mathbb{R}^d$ by an infinite dimensional space, such as a function space — you can try to make all this work on an infinite-dimensional space of functions — usually this measure μ_V is called a Euclidean quantum field theory. This arises in mathematical physics, which is a topic of contemporary interest — basically you want to define measures of this form on infinite dimensional spaces. Even defining such a measure is a huge problem, because the space is infinite.

Then the associated Langevin dynamics (LD) is a stochastic PDE. This is also a quite active field of research (it's actually one Sky works in). Even showing why solutions to these dynamics exist is highly nontrivial. Actually, until even 11 years ago, there were many situations in infinite dimensions where you have some Langevin dynamics, and you can't even show you have local-in-time existence. Often you have initial data x_0 and want to say I can at least run my dynamics for a short time starting from x_0. But we didn't really know how to do this in many interesting examples until 11 or 12 years ago, and the breakthrough won a Fields medal — he was studying stochastic PDEs and gave a very general way to prove local existence to these types of Langevin dynamics in infinite dimensions, and it was a huge breakthrough.

All this is to say is that instances of Langevin dynamics are quite active fields of research, especially when you go to infinite dimensions. In finite dimensions it also appears in other areas, more to do with applications.

§25.5 Existence of solutions

The first thing we hinted at wanting to be able to say is why you have local solutions to this SDE:

> **Question 25.8.** Why do solutions exist?

In the case of OU, you were just able to write down a solution. That's very special to that case, because the SDE was linear, and if it's linear in X you can just write down an explicit solution and show it's unique. But in general, that's not the case.

Let's start in a simplified setting.

> **Proposition 25.9**
>
> Let $b : \mathbb{R}^d \to \mathbb{R}^d$ be Lipschitz. Then for all Y which are $\mathcal{F}_0$-measurable (and $|Y| < \infty$), there exists a unique solution to the SDE
>
> $$dX_t = b(X_t) + \sqrt{2}\, dB_t \text{ and } X_0 = Y,$$
>
> on the entire time-interval $[0, \infty)$. Also, X is adapted.

It is nice to have adapted processes; that'll come from our assumption that Y is $\mathcal{F}_0$-measurable. If you didn't assume that, you could still have done all this, but then X would not be adapted.

Here you think of b as $-\nabla V$; so if that's Lipschitz, you get a global solution. In the OU case it was $-\nabla V(x) = x$, which is Lipschitz with constant 1; so you can apply it to the OU case if you wanted.

We'll get started on (but may not finish) the proof.

Proof. As usual, what does it mean to solve a SDE? It means you satisfy the corresponding integral equation — so we want a stochastic process which satisfies

$$X_t = Y + \int_0^t b(X_s)\, ds + \sqrt{2} B_t.$$

The way you usually do this is you define a map $\mathcal{G}$ which takes in a continuous function on some time interval $[0, T]$ (here T is a fixed time we'll choose later, not a stopping time) and outputs another — so we define $G : C([0,T]) \to C([0,T])$ by the following formula, inspired by this integral equation —

$$(Gf)(t) = Y + \int_0^t b(f(s))\, ds + \sqrt{2} B_t.$$

So the above integral equation is the same as finding a fixed point of G — we want $X = \mathcal{G}(X)$.

You want to view these integral equations as fixed-point equations — if you took ODE theory, the way you solve an *ordinary* differential equation locally in time is to set up a fixed point argument.

> **Remark 25.10.** Here we're not going to use any distributional properties of B, only that it's a continuous function. In some sense there's not any probability in this statement; it's really an analytic thing.

How do you find fixed points to functions of this form? That's where the contraction mapping principle comes in. We'll state it slightly abstractly:

Theorem 25.11 (Contraction mapping principle)

Suppose (M, d) is a complete metric space and $G : M \to M$ is a strict contraction, i.e., there exists some $0 \leq \alpha < 1$ such that

$$d(G(x), G(y)) \leq \alpha d(x, y)$$

for all $x, y \in M$. Then there exists a unique $x_* \in M$ such that $G(x_*) = x_*$. Moreover, $x_* = \lim_{n\to\infty} G^{(n)}(x)$ for any $x \in M$.

So there is a unique fixed point, and the way you find it is by starting at any point in your metric space and applying G a bunch of times.

This gives you a way to find fixed points. We won't prove this, but the main idea is, why should you expect $\lim_{n\to\infty} G^{(n)}(x)$ to be a fixed point? At least formally, we have

$$G(G^{(\infty)}(x)) = G^{(\infty+1)}(x) = G^{(\infty)}(x),$$

which is why you'd intuitively expect this limit to be a fixed point. So what we want to show is if you start at any $x \in X$ and apply G many times, you get a Cauchy sequence; and then you can use completeness to show it converges. And then that limit is going to be a fixed point. And to show uniqueness, you just use the contraction map — if you had x_* and y_* which were fixed points, then you're saying

$$d(x_*, y_*) \leq \alpha d(x_*, y_*),$$

which means this distance has to be 0.

So let's try to set up a situation where we can apply this — let's try to prove the contraction inequality. Looking at the sup norm on $C([0, T])$, we have

$$\|Gf - Gg\|_{C[0,T]} \leq \sup_{t\in[0,T]} \left| \int_0^t b(f(s)) - b(g(s))\, ds \right| \leq \int_0^T |b(f(s)) - b(g(s))|\, ds.$$

(This integral definitely bounds each of the above integrals for any $t \in [0, T]$, using the triangle inequality.) And the point is you want to say this is controlled by $|f - g|$. To do that, we need a Lipschitz assumption on b. But that's exactly what we assumed. So if b is Lipschitz with constant k, you get the bound

$$K \int_0^T |f(s) - g(s)|\, ds.$$

I wanted to bound this by a sup norm, but this is a L^1 norm; but I can just bound this by

$$KT \|f - g\|_{C[0,T]}.$$

And we're free to choose T on our own, so we can just choose it so that KT is much smaller than 1 (e.g., $\frac{1}{2}$).

So if we set $T = 1/2K$, then indeed G is going to be a contraction, with

$$\|Gf - Gg\|_{C([0,1])} \leq \frac{1}{2} \|f - g\|_{C([0,1])}.$$

So that's the first step.

But we're not done yet, because we need to identify a metric space M for which G maps M to itself. So basically, what we need to do is show there exists some $R \geq 0$ such that if I look at the radius-R ball in the space of continuous functions — i.e.,

$$B_R = \{\|f\|_{C([0,T])} \leq R \mid f \in C([0, T])\}$$

(note that the way we defined T is completely uniform in f, g, and the initial data Y), then G maps this ball to itself. If we can show this, then we can take $M = B_R$.

So this is the last thing we need to show; once we have that, we can apply the contraction mapping principle to get a unique fixed point in this ball. And that fixed point, by construction, will satisfy this integral equation, which is the thing we're trying to construct.

Probably we don't have enough time for this (though it's not too long), so we'll leave that for next time. And once you've shown this, you'll be able to make the statement that for any value of Y, a solution exists on the interval $[0, 1/2K]$. But that was uniform in your initial data, so you can keep doing this to ∞. That's why in this case you'll automatically get global existence — the time of existence you get from the local theory doesn't depend on the initial data (at all). That's heavily due to the Lipschitz assumption. Later we're going to remove the Lipschitz assumption and just assume you're locally Lipschitz. This will let us deal with a much wider class of V (assuming ∇V is Lipschitz is very strong; but if we assume V is C^2, then ∇V will be continuous and therefore at least locally Lipschitz). □

Student Question. *What is the exam going to be like?*

Answer. Sky doesn't intend it to be super tricky; more like just if you know the concepts. We can discuss more on Monday. He'll think about whether it's open/closed book.

§26 May 12, 2025 — Langevin dynamics

Today is the last day of class; we'll talk about some properties of the Langevin dynamics, which are SDEs of the general form

$$dX_t = -\nabla V(X_t) + \sqrt{2}\, dB_t.$$

You should think of it as a stochastic gradient flow — if you take out the Brownian term, it's really the (negative) gradient flow of V. So without that, the ODE would be trying to minimize V. (You might converge to a local minimum if V has multiple wells. But generally if you have some function V you want to minimize, you could run gradient flow to drive you to at least a local minimum.) You start with a gradient flow and then add in this stochastic Brownian forcing term.

We'll see today that LDs are naturally associated to measures of the form

$$d\mu_V(x) = Z_V^{-1} \cdot \exp(-V(x))\, dx$$

(where Z is just a normalizing constant); this is the invariant measure. After we talk about existence questions, we'll talk about what the generator of the LD is; and after we compute it, we'll be able to show that indeed the invariant measure of this dynamic will be given precisely by μ_V.

You can imagine you're doing Markov chain Monte Carlo where you start with a measure and want to take samples from it, so you want to write down a Markov process for it; and this is the natural one.

§26.1 Existence

But before we get to this, we want to talk about existence of solutions to this SDE. We can't write down an explicit one in general (like we did with OU), so we have to prove existence.

§26.1.1 The Lipschitz case

Lsat time we started talking about existence when your drift term b (which you should think about as $-\Delta V$) is globally Lipschitz — we said that then you have global existence (and it'll be adapted if your initial data is $\mathcal{F}_0$-measurable).

Proposition 26.1

Let $b : \mathbb{R}^d \to \mathbb{R}^d$ be Lipschitz. For $Y \in \mathcal{F}_0$ with $|Y| < \infty$, there is a unique solution to

$$dX_t = b(X_t) + \sqrt{2}\, dB_t$$

with $X_0 = Y$ on $[0, \infty)$.

The way you prove existence of solutions to SDEs is by following the proof for ODEs. You want to write this as an integral equation, so you define $G : C([0,T]) \to C([0,T])$ by

$$(Gf)(t) = Y + \int_{-} 0^t b(f(s))\, ds + \sqrt{2}B_t.$$

We want to find a fixed point $GX = X$, and that'll be your solution X.

To do so, we need to find $T > 0$ and $R > 0$ such that writing $B_R = \{f \in C([0,T]) \mid \|f\| \le R\}$ (here the norm is the norm with respect to $C([0,T])$, which is the sup norm) — to find a fixed point, you want to show it's a contraction on a suitable space. So the suitable space is going to be some B_R. So we just need to show G maps B_R to itself, and for two functions f and g in this ball, we have the strict contraction property

$$d(G(f), G(g)) \le \frac{1}{2} d(f, g).$$

(The $\frac{1}{2}$ is arbitrary; it just has to be a constant less than 1.)

When we do this, by the contraction mapping principle we'll get a fixed point at least to the time T.

Last time we showed that for all $f, g \in C([0,T])$ (we don't even need the assumption that $f, g \in B_R$), we have

$$d(Gf, Gg) \le KT d(f, g),$$

where K is the Lipschitz constant of b. This gives a natural choice of what T should be, e.g., $T = 1/2K$. Then definitely the second property (contraction) is true.

Now we just need to find some radius R for which the first property is true. For this, if I want to bound $\|Gf\|$ on $[0,T]$, I can bound it by

$$\|Gf\| \le \sup_{t \in [0,T]} \left(|Y| + \int_0^t |b(f(s)) - b(0)|\ ds + \int_0^t |b(0)|\ ds + \sqrt{2}\, |B_t| \right)$$

(this is just the definition of G, where we added and subtracted $b(0)$). Why did we insert $b(0)$? I want to use the fact that b is Lipschitz, so we'll essentially compare $f(s)$ to 0. Now we can bound this whole thing by

$$|Y| + K \int_0^t |f(s)|\ ds + T\, |b(0)| + \sqrt{2}\, \|B\|$$

(these are all integrals of nonnegative things, so I can bound them by the endpoint integral; I use the Lipschitz assumption to bound $|b(f(s)) - b(0)|$, and $\|B\|$ is again on $[0,T]$).

And I'm trying to say if I assume $\|f\|$ is controlled, so is $\|Gf\|$; so I want to bound the RHS by $\|f\|$ plus my inputs (which are $|Y|$, $\|B\|$, and whatever $|b(0)|$ is). So in the end, I should be able to bound this by

$$\|Gf\| \le |Y| + KT\, \|f\| + T\, |b(0)| + \sqrt{2}\, \|B\|\,.$$

This is true for any continuous function f on $[0,T]$. And now, if you kind of play around with what R should be — it definitely has to be greater than $|Y|$ and all the other terms. So let's say we define

$$R = 2\left(|Y| + \sqrt{2}\,\|B\|_{C([0,T])} + \frac{|b(0)|}{2K}\right).$$

Where does this come from? If I define my radius like this, then I'll see that if $\|f\| \le R$ (where the norm is in $C([0,T])$), then

$$\|Gf\|_{C([0,T])} \le \frac{1}{2}\|f\|_{C([0,T])} + \text{the contribution from the other three terms},$$

and by how I chose R, those contributions are at most $\frac{1}{2}R$. And so if I impose $\|f\|_{C([0,T])} \le R$, I get $\frac{1}{2}R + \frac{1}{2}R = R$.

So with this choice of R — which again, only depends on the input data — you can find a radius on which G is a self-map on the ball of that radius.

So in summary, we've succeeded at finding a time T and radius R for which these two properties are true (so that G is a contraction on this complete metric space). This means there exists a unique $X \in B_R$ such that $GX = X$.

This gives *some* fixed point. What does unique mean? If you had another element in this ball which was a fixed point, that element has to be X. But there's a slightly stronger notion of uniqueness:

Exercise 26.2. Show that the fixed point is unique even in $C([0,T])$.

Here the statement is just that if I have a continuous function on this interval which solves the integral equation — not assuming *a priori* that it's bounded by R — it still has to be X.

The way you do this is that in the contraction mapping principle, if you're willing to adjust your existence time T, you can take your value of R to be *anything* strictly greater than the initial data; but then you have to take T to depend on R (and be sufficiently small). If it's a continuous function, on a tiny time-interval, its norm cannot be too far from Y — so on some small-enough time, you'll have a bound on its norm by some R, which means you can make this contraction argument work; and you show on that tiny time-interval it has to be X. And then you can iterate and patch all these together.

And then also, you have that — we need to say why X is going to be adapted. Right now it's just some abstract fixed point thing. But — how do you find the fixed point X in the contraction mapping theorem? You just start from an arbitrary point in the ball and repeatedly find G. So you could have written

$$X = \lim_{n\to\infty} G^{(n)}(0)$$

(since the 0 function is definitely an elemnet of B_R; you could've started from anything in the ball). This convergence is in the L^∞ norm; so that implies it also has to be true for any fixed time t. And one checks inductively that for all n and all t, the thing on the right-hand side is measurable with respect to $\mathcal{F}_t$, i.e.,

$$(\mathcal{G}^{(n)}(0))(t) \in \mathcal{F}_t.$$

Looking at the definition of G, if f is adapted to the filtration, why is Gf adapted? Well, $Y \in \mathcal{F}_0$ so that's fine; $\int_0^t b(f(s))\,ds$ is adapted (it's the integral of a continuous function, so it's $\mathcal{F}_t$-measurable); and you're assuming you have a $\mathcal{F}_t$-Brownian motion, so that's fine. So by induction you get this adaptedness statement.

Finally, we showed existence on this time-interval $\frac{1}{2K}$; but we're claiming it actually exists on the entire real line. So for global existence, you just use that $T = \frac{1}{2K}$ does not depend on your initial data $|Y|$ — no matter how large your initial data is, you can take a step of size $1/2K$. Then you patch these local solutions to ∞; and you never run into trouble because this time doesn't depend on how large you are.

Remark 26.3. In general nonlinear problems you're worried that this time might depend like $1/|Y|^2$, for example, which would be very problematic for showing global existence — if

$$T \sim \frac{1}{|Y|^2}$$

(think of Y as very large), then you cannot just do this thing, because you might be worried that as you get super large your time-interval gets really tiny, so that in the limit your time of existence just converges to a finite time.

Student Question. *Why do you need the ball of radius R? Why can't we just apply the contraction mapping principle on $C([0,T])$ itself, since that's complete?*

Answer. Actually you're right, we probably don't; if you assume globally Lipschitz, R can be ∞.

§26.1.2 Local existence

Now we'll state a result that's slightly more general than this — the globally Lipschitz assumption is kind of restrictive. So let's state a result that works for *locally* Lipschitz functions b (i.e., on any bounded subset of $\mathbb{R}^d$, you're Lipschitz).

Theorem 26.4

Let $b : \mathbb{R}^d \to \mathbb{R}^d$ be locally Lipschitz, and let B be an $(\mathcal{F}_t)$-Brownian motion. Then for all $Y \in \mathcal{F}_0$ with $|Y| < \infty$, the SDE

$$dX_t = b(X_t) + \sqrt{2}\, dB_t, \quad X_0 = Y$$

has a unique continuous adapted solution X defined on some maximal time-interval $[0, T]$, where T is a stopping time. Moreover, if $T < \infty$, then $\lim_{t \uparrow T} |X_t| = \infty$.

So you're looking at the maximal time for which you have a solution to this equation. And you're saying if that time is finite, that has to mean your solution blows up as you approach that time.

So now if b is locally Lipschitz you may not have global solutions — basically because of this problem with $T \sim 1/|Y|^2$ where it could be that your time of existence depends quite badly on the initial data, so you can't iterate to ∞. But at least you have this characterization of what this maximal time should be — precisely the time you blow up.

We won't spend the time going into all the details on this, so we'll give a proof sketch.

Proof sketch. At least for local existence — forget about the maximal time of existence, how do you show you just have a solution on some tiny time-interval T? You can for instance take $R > |Y|$ (you again want to make a contraction mapping argument by finding some radius and time; here you probably do actually need the finite radius, you can't just do this on $C([0,T])$ because you're not globally Lipschitz). Then you're locally Lipschitz, so you can at least hope to make the contraction mapping principle work on small time, if you impose some bound on the norm. If your function is in B_R, then you can act as if b is Lipschitz (because if f is bounded by R, then you're applying b to some bounded subset, and you have some local Lipschitz constant there). So for the question of local existence on some time-interval, you can basically pretend you're globally Lipschitz.

Then to characterize the maximal time of existence, we can imagine drawing a picture with your x-axis representing time and y-axis $|X_t|$. You start at Y, and you can imagine you're tracking the size of your solution as a function in time. If it happens that at some later time T_0 you converge to some finite value,

or at least you don't blow up, then you can always improve your time-interval and get a solution on this expanded time interval to $T_0 + \varepsilon$. So if for some T_0 you don't blow up as you converge to that time, you can always expand a little bit. That's by applying the local result. So if you consider the left limit approaching T_0, if it doesn't blow up you can extend. The only way you can not extend past a time is if you blow up at that time.

That's in words what you want to write in this argument. But that's why we're sketching it; you kind of have to write out all the details, which we don't want to do today. □

§26.1.3 Conditions for global existence

Let's assume we now have this result where for any locally Lipschitz b, you get a local solution whose maximal time of existence is characterized by this kind of stopping time. Let's investigate conditions on V for which you actually get global existence. For us V will always be such that b is locally Lipschitz — it'll always be smooth, so ∇V is also smooth.

Before stating this, we'll use a lemma:

Lemma 26.5 (Gronwall)

Suppose $\alpha(t)$, $\beta(t)$, and $f(t)$ are such that α is nondecreasing and $f(t) \le \alpha(t) + \int_0^t \beta(s) f(s)\, ds$ for all $t \in [0, T]$. Then for all $t \in [0, T]$, we have

$$f(t) \le \alpha(t) \exp\left(\int_0^t \beta(s)\, ds \right).$$

As a consistency check, when $t = 0$ you're saying $f(0) \le \alpha(0)$, which is consistent with plugging $t = 0$ into this. Basically this is saying if you have some self-bounding inequality like this, you automatically get a bound on f in terms of α and β (you want to turn this thing into a bound on f that doesn't depend on f, only the inputs, and that's what Gronwall does).

We're going to apply this to get global existence.

Theorem 26.6

Let V be such that ∇V is locally Lipschitz and such that one of the following conditions is satisfied:

(a) $V(x) \to \infty$ as $x \to \infty$, and $|\nabla V|^2 - \Delta V$ is bounded below.

(b) There exist constants $a, b \in \mathbb{R}$ such that $x \cdot \nabla V(x) \ge -a|x|^2 - b$.

Then for any $x_0 \in \mathbb{R}^d$, almost surely the Langevin dynamics has a unique solution on $\mathbb{R}_+$ with initial data $X_0 = x_0$.

The condition (b) is kind of saying on some level that you approximate a quadratic function — if you took $V = x^2/2$ (the thing that gives you the OU process), then ∇V is x itself, and $x \cdot x = x^2$, so you satisfy this with $b = 0$ and $a = -1$. So this is saying you approximate a quadratic in some sense. And then ∇V is linear, and linear things are manageable — when you have something linear you should be able to prove whatever you want.

Basically if you assume the local existence theorem, all you need to show is that the stopping time T is ∞ almost surely — because if you prove $\mathbb{P}[T < \infty] = 0$, that basically means you have a global solution. So that's what you want to show.

Proof of (a). Let's assume (a) first. If you assume (a), it turns out the right thing to do is you apply Ito's formula to $V(X)$. You want to understand the evolution of $V(X)$, and you want to say why $V(X)$ will not

get large; because the assumption $V(X) \to \infty$ as $X \to \infty$ means that if V stays controlled, then X will also stay controlled (that's the contrapositive — if $X \to \infty$ then $V \to \infty$, so if V stays bounded then X stays bounded). So we want to study the evolution of V and show it doesn't blow up. Applying Ito's formula, we get

$$dV(X_t) = \nabla V(X_t) \cdot dX_t + \frac{1}{2}\partial_{ij} V(X_t)\, d\langle X^i, X^j\rangle_t.$$

Then you substitute in your equation for X. We specified what dX is, and when you plug that in you get

$$dV(X_t) = -\left|\nabla V(X_t)\right|^2 dt + \sqrt{2}\nabla V(X_t) \cdot dB_t + \Delta V(X_t)\, dt$$

(we're not going to care about the $\nabla V(X_t) \cdot dB_t$ term soon — it's just a (local) martingale term — and when you compute the QV term, the $\frac{1}{2}$ gets cancelled by the two factors of $\sqrt{2}$).

You can kind of see where the assumption is coming from — you want to assume something about your drift, because you're saying if we track the evolution of V, we have some drift term which is hopefully a restoring term so you never get too large, and maybe some Brownian fluctuations on top of it. But you want the drift term to generally take you back down to 0 or 1 or something, rather than down to ∞.

The right way to group these terms is

$$dV(X_t) = \sqrt{2}\nabla V(X_t)\, dB_t - (|\nabla V|^2 - \Delta V)(X_t)\, dt.$$

Now let's add in a localization — let $T_n = \inf\{t \geq 0 \mid |X_t| > n\}$. Now if you localize everything (so you stop at time T_n), the local martingale term is an actual martingale term (X becomes bounded and V is regular enough). And once you have a martingale you can take expectations; this guy is mean-0, so we're not going to worry about it. So we have that

$$\mathbb{E}[V(X_{t\wedge T_n})] = V(X_0) + \mathbb{E}\left[\int_0^{t\wedge T_n} (\Delta V(X_s) - |\nabla V(X_s)|^2)\, ds\right]$$

(here we're writing the integral version of this evolution equation and taking expectations).

Now by our assumption (a), we're saying that thing up there is bounded below, which is the same as saying its negative is bounded above; and I exactly have the negative of that in my drift term. So I get that there exists some constant C (which is the lower bound of the thing from (a)) such that

$$\mathbb{E}[V(X_{t\wedge T_n})] \leq V(X_0) + Ct.$$

From here, now the intuition is that you can remove the localization and say $\mathbb{E}[V(X_t)]$ is bounded by something, and this is basically saying $V(X)$ is not going to blow up at any finite time (which also means X should not). But the way to make this a bit formal is that by (a), we can say V is bounded below by some $\alpha \in \mathbb{R}$; so if your truncation level n is larger than $|x_0|$ (which will be true because we'll be sending $n \to \infty$), then we get

$$\mathbb{E}[V(X_{t\wedge T_n})] \geq \alpha + V(n)\mathbb{P}[T_n \leq t]$$

(the first term corresponds to the case $T_n \geq t$; the second case corresponds to the case $T_n \leq t$, in which case at T_n you have to be exactly n).

And now we're basically done — we combine this lower bound of the left-hand side with whatever's on the right-hand side, and we get that

$$\alpha + n \cdot \mathbb{P}[T_n \leq t] \leq V(X_0) + Ct.$$

Thus we have

$$\mathbb{P}[T \leq t] \leq \limsup_{n\to\infty} \mathbb{P}[T_n \leq t] = 0$$

(where T is the maximal existence time appearing in the local theory). (The first inequality is because if you blow up, then you definitely have to have reached n for arbitrary n.) The RHS goes to 0 because we showed

$$\mathbb{P}[T_n \le t] \le \frac{\text{stuff that don't depend on } n}{n}.$$

So we've shown that $\mathbb{P}[T < \infty] = 0$, which means your maximal existence time is infinite, and so you have a global solution. □

That's the proof under (a). Under (b) you want to do something slightly different — basically you want to track a different statistic instead of V.

Proof for (b). Now let's assume (b). In this case, it turns out the right thing to look at is the evolution of $|X|^2$ (if this doesn't blow up, neither does X). By Ito's formula, we have

$$\frac{1}{2}d(|X_t|^2) = X_t \cdot dX_t + \partial_{ij} d\langle X^i, X^j\rangle_t.$$

When you input what you have fr dX, what you get in the end is

$$\sqrt{2}X_t \cdot dB_t - X_t \cdot \nabla V(X_t) + (2d)\, dt$$

(the first two terms come out of the first term, and the $2d$ from the QV term; the d is because you have a contribution of 2 from each index $i = j$).

Now let's write an inequality for this. Using our assumption (b), which is saying that $X \cdot \nabla V(X)$ is lower-bounded, now we have a minus so we get a lower bound; so we can say

$$\frac{1}{2}d(|X_t|^2) \le \sqrt{2}X_t \cdot dB_t + (a\,|X_t|^2 + b + 2d)\, dt.$$

Now we're almost in position to apply Gronwall — you're saying the evolution of X^2 is bounded in terms of itself plus some additional input, and you see when you take expectations, the martingale drops out. So we get that for all t and truncation levels n, we get

$$\frac{1}{2}\mathbb{E}\left[\left|X^2_{t\wedge T_n}\right|\right] \le \frac{1}{2}\,|X_0|^2 + \int_0^t a\mathbb{E}[|X_{s\wedge T_n}|^2]\, ds + (b+2d)^+ t$$

(the reason you localize is so that the local martingale term is an actual martingale and its expectation drops out). (We're interpreting the above equation in integral form, and then integrating it.)

Now we apply Gronwall to the fucntion $f(t) = \mathbb{E}[|X_{t\wedge T_n}|^2]$, and we obtain that

$$\mathbb{E}[\left|X^2_{t\wedge T_n}\right|] \le (|X_0|^2 + 2(b+2d)^+ t)e^{2at}$$

(where $(b+2d)^+$ is the positive part of the thing).

And once you get this, you're basically done — now you can send $n \to \infty$ and you get a bound on $\mathbb{E}[|X_t|^2]$. Sending $n \to \infty$, we obtain that $\mathbb{E}[|X_t|^2] < \infty$ (in fact you have some explicit bound in terms of your initial data and two input parameters, but at least for global existence you just need this to be finite), which means

$$\mathbb{P}[T > t] = 1,$$

which implies that $\mathbb{P}[T = \infty] = 1$. And that basically finishes the proof in this case. □

Student Question. *How do we define T?*

Answer. Using the local existence thing — the time at which your solution blows up.

Student Question. *For (a), why is V bounded below?*

Answer. We assumed V goes to ∞ when $x \to \infty$, and we're also assuming V is continuous (if ∇V is locally Lipschitz, that should imply V is at the least continuous). Then there exists R such that for all $|x| \geq R$ you have $V(x) \geq 1$ (this is a consequence of $V \to \infty$ as $x \to \infty$). (By $x \to \infty$ we mean $|x| \to \infty$.) And you combine this with the fact that V is continuous, so it's bounded on the ball of radius R.

§26.2 The Markov semigroup of Langevin dynamics

That's the proof of global existence under these assumptions. Now let's discuss some general facts about these Langevin dynamics.

Remark 26.7. Last time we talked about how Langevin dynamics when you do things in infinite dimensions is an active research field. In fact, later today Sky is giving a talk in the probability seminar at 3pm (in the room next to ours) about a case where you have some function V on an infinite dimensional space and try talking about global existence of the stochastic PDE. So this is still a topic of modern interest. If you are interested you can come; but unlike in classes, in seminars you don't go in expecting to understand everything.

Let $(\mathcal{P}_t)$ be the Markov semigroup corresponding to the Langevin dynamic. In the same way we saw that the OU process is always a Markov process, so are these more general solutions to SDEs, so you have its semigroup.

We're going to assume $(\mathcal{P}_t)$ is Feller. In general you have to prove this; you can look at the definition of Feller and prove each point, and somehow you have to prove these dynamics behave as nicely as you'd expect. There's some discussion in the notes, but we'll skip it.

Question 26.8. What's its generator?

You compute this the same way we did for the OU process, and it turns out

$$L = \Delta f - \nabla V \cdot \nabla f.$$

The reason is when you compute the generator, you want to compute for some smooth function f, you want to use Ito to compute $df(X_t)$. W basically did this computation when f was V, but basically you'll get a martingale part and some drift term, and you want to figure out what it is; adn it's

$$df(X_t) = \text{MG} + (\Delta f - \nabla V \cdot \nabla f)(X_t)\, dt.$$

And there's this general result that the generator is the right thing to subtract from $f(X_t)$ to get a martingale term; so that's the generator.

Remark 26.9. To see why this generalizes OU, if $V = |x|^2/2$ then $\nabla V(x) = x$, and this is precisely the generator we found for OU. So that's at least consistent.

Interestingly, there's still the exact same integration by parts relation between the generator and the L^2 space.

Proposition 26.10 (Integration by parts)

Letting μ_V be the probability measure defined before, we have

$$(Lf, g)_{L^2(\mu_V)} = -(\nabla f, \nabla g)_{L^2(\mu_V)} = (f, Lg)_{L^2(\mu_V)}.$$

Proof. This is by calculation, basically the same as in the OU process. You want to do integration by parts; you can go back to the one for the OU process, replace every instance of x by ∇V, and you prove this more general formula. □

A consequence of this is that:

Corollary 26.11

The measure μ_V is an invariant measure for $(\mathcal{P}_t)$.

The way you prove this is again the same as how you prove the corresponding statement for OU. We'll write this one out: To see this, recall that you have this integral formula

$$\mathcal{P}_t f = f + \int_0^t L\mathcal{P}_s f \, ds.$$

Now if I want to compute the expectation under my measure of $\mathcal{P}_t f$, I can write it as

$$\int \mathcal{P}_t f \, d\mu_V = \int f \, d\mu_V + \int \int_0^t L\mathcal{P}_s f \, ds \, d\mu_V.$$

So I want to say that the last thing is 0. Why? I can switch the integral and it becomes

$$\int_0^t ds \int L\mathcal{P}_s f \, d\mu_V.$$

And the thing on the inside, you want to think of it as a L^2 inner product $(L\mathcal{P}_s f, 1)_{L^2(\mu_V)}$. Then you use integration by parts to say this is $-(\nabla \mathcal{P}_s f, \nabla 1)_{L^2(\mu_V)}$. And 1 is a constant function, so this is 0.

§26.3 Variance decay and entropy decay

We'll also say that — because you have this integration by parts identity, the formulas for variance decay and entropy decay also look the same as for the Langevin dynamics.

Definition 26.12. We define $\mathsf{Ent}_{\mu_V}(f) = \mathbb{E}_{\mu_V}[f \log f] - \mathbb{E}_{\mu_V}[f] \log \mathbb{E}_{\mu_V}[f]$.

If you go through the proof of the entropy evolution proposition for OU, you do the exact same calculations and see that

$$\partial_t \mathsf{Ent}_{\mu_V}(\mathcal{P}_t f) = -\mathbb{E}_{\mu_V}\left[\frac{|\nabla \mathcal{P}_t f|^2}{\mathcal{P}_t f}\right].$$

When $\mu_V = \gamma_d$ this is what we computed before; if you look at the proof, all we needed was the integration by parts identity; so you get the same entropy evolution here. You also get the same variance evolution — if we look at the second moment,

$$\partial_t (\mathcal{P}_t f, \mathcal{P}_t f)_{L^2(\mu_V)} = -2(\nabla \mathcal{P}_t f, \mathcal{P}_t f)_{L^2(\mu_V)}.$$

The calculation is again the same.

§26.4 Poincare and log-Sobolev

All this is to say is, now you can ask:

Question 26.13. When does the Poincare or log-Sobolev inequality hold for μ_V?

(You basically want to understand, when is the dynamics rapidly mixing?)

What does it mean for a measure to satisfy a Poincare or log-Sobolev inequality?

Definition 26.14. We say that μ_V satisfies the Poincare inequality with constant α if for all $f \imath C_b^1(\mathbb{R}^d)$, we have

$$\mathrm{Var}_{\mu_V}(f) \leq \alpha(\nabla f, \nabla f)_{L^2(\mu_V)}.$$

We say μ_V satisfies the log-Sobolev inequality with constant α if

$$\mathsf{Ent}_{\mu_V}(f) \leq \alpha \mathbb{E}_{\mu_V}\left[\frac{|\nabla f|^2}{f}\right].$$

(This means bounded and bounded first derivative.)

The motivation is we want to understand when variance or entropy decays exponentially over time. And with these conditions, you'd be saying the evolution of the entropy is $-\alpha$ times the entropy itself, and then by Gronwall you get it's decaying exponentially.

If Poincare or log-Sobolev holds, then you have rapid convergence to equilibrium — all the discussion in the Gaussian case exactly carries over. (You use Gronwall on this thing.)

So basically, we want to understand, when does a measure satisfy these nice functional inequalities? One sufficient condition is if it has enough convexity or concavity — if $\nabla^2 V(x) \geq \gamma \,\mathrm{id}$ for all x (the Hessian of V, which is a matrix, is uniformly lower-bounded by some constant times the identity), then it turns out you have a sub-commutation relation

$$|\nabla \mathcal{P}_t f| \leq e^{-\gamma t} |\mathcal{P}_t \nabla f|,$$

and this will imply Poincare and log-Sobolev with $\alpha = 1/\gamma$ and $1/2\gamma$, respectively (but the exact values don't matter). The way you use this to prove Poincare or log-Sobolev is the same as in the Gaussian case — the key part is to use this commutation identity (in the Gaussian case it was equality, but we remarked in class that all you need was an inequality, and then you still get these functional inequalities; the reason is basically that this is saying you converge rapidly to equilibrium, because your gradient after you apply $\mathcal{P}_t$ is tiny so you're quickly approaching a constant function). The proof is applying Ito's formula to the right thing.

The point we're trying to make is all the things in the Gaussian case generalize, though you have to make assumptions, such as convexity. Nowadays the problems of modern interest are when you don't have these assumptions — V might have many local minima and not be convex, and you still want to know if you have rapid mixing and so on.

§26.5 The final

Sky made the announcement that you can make 1 cheat sheet (front and back). He doesn't intend it to be a cleverness test; he treats this class more like skill-building, so he wants to see we know the basic concepts (can you apply Ito's formula? do you know what a stopping time is?). So hopefully the problems should just be applications of concepts; they shouldn't require a bunch of clever tricks.

Student Question. *Can we type the cheat sheet?*

Answer. Sure, but let's limit the font; let's say 8pt font. Which is small enough, he thinks. The cheat sheet will mostly help with review — his expectation is when you understand a concept deeply, you can kind of just immediately recall it. It definitely helps you to study to put in some time recalling everything; but math is not about memorization, but also often when you've really worked on something you can kind of do things very quickly. So that's also kind of why he wants to limit how much of a cheat sheet you have; even though math isn't about memorization, it still helps to at least know some concepts.

Student Question. *What do you mean by clever tricks; does the homework use clever tricks?*

Answer. His intention is with homework you have way more time so the problems are allowed to be more involved. He doesn't think the average problem should be as hard as the average homework problem. Usually with tests there's some meta-thinking where you see a problem and want to figure out which concept they're trying to test you on; sometimes that helps you because there's only finitely many things you shoudl be using. With research it's different; but on a test if you don't know how to do a problem, you can ask yourself this.

Student Question. *Is the exam going to be 3 hours?*

Answer. Yes. It might be a longer exam but maybe each problem shouldn't be too involved. But he hasn't actually written it; he'll write it this week and then they'll figure out how long it should be.