Stochastic Calculus, Courant Institute, Fall 2018
http://www.math.nyu.edu/faculty/goodman/teaching/StochCalc2018/index.html
Jonathan Goodman, October, 2018

Lesson 6, Simulating, change of measure

November 5, 2018

1 Introduction

Suppose you have a stochastic model of something and you want numbers – specific facts about what your stochastic process does. You can get numbers by solving a backward or forward equation (depending on what you want to know), or by simulation. This lesson describes the basic tools for direct simulation of stochastic models.

For example, suppose X_t is a diffusion process that satisfies the SDE

$$dX_t = a(X_t)dt + b(X_t)dW_t \; . \tag{1}$$

The goal is to evaluate the expectation

$$f = \mathrm{E}_{x,0}\big[V(X_{[0,T]})\big] \; . \tag{2}$$

The "observable" V could be a *final time* function like $V(X_{[0,T]}) = X_T^2$, or it could be *path dependent* as in $V(X_{[0,T]}) = \int r(X_t)dt$.

Monte Carlo[1] analysis means finding numbers that themselves are not random but are related to a random process. and X_t is a diffusion process with a given SDE. We use $\widehat{f}$ for a Monte Carlo simulation. We have the computer create create a large number of *sample paths* for the SDE, called $X_t^{(n)}$ for t running from 0 to T and n running from 1 to N. The direct Monte Carlo estimate is

$$\widehat{f} = \frac{1}{N}\sum_{n=1}^{N} V(X_{[0,T]}^{(n)}) \; . \tag{3}$$

This is like random sampling in statists. As in statistics, it is important to have *error bars*, which give an idea how accurate $\widehat{f}$ is likely to be.

Most diffusion processes cannot be simulated exactly. Instead, they are simulated approximately using a time step Δt. The *Euler Maruyama* formula (often just called the *Euler* formula) is a *time stepping* method to create approximate sample paths. Define $t_k = k\Delta t$ and $X_k^{(n,\Delta t)}$ to be an approximation to $X_{t_k}^{(n)}$. We want the approximate process to have increments with approximately the right mean and variance over a *time step* of size Δt. This can be done using

$$X_{k+1}^{(n,\Delta t)} = X_k^{(n,\Delta t)} + a(X_k^{(n,\Delta t)})\Delta t + b(X_k^{(n,\Delta t)})\sqrt{\Delta t}Z_k^{(n)} \; . \tag{4}$$

[1] Monte Carlo is the capital city of the country Monaco. The country is so small and the city so big that most of Monaco is inside Monte Carlo. Monte Carlo is famous for gambling and car racing. Using random numbers in computation is like using random numbers in gabmling, which is how computing with random numbers came to named after a center of gambling.

The numbers $Z_k^{(n)}$ are independent standard normals

$$Z_k^{(n)} \sim \mathcal{N}(0,1) \ , \ \text{ i.i.d. .} \tag{5}$$

This equation has the property that

$$\mathrm{E}\Big[X_{k+1}^{(n,\Delta t)} - X_k^{(n,\Delta t)} \mid \mathcal{F}_k \Big] = a(X_k^{(n,\Delta t)})\Delta t \tag{6}$$

An exact path satisfies this only approximately

$$\mathrm{E}\Big[X_{t_{k+1}}^{(n)} - X_{t_k}^{(n)} \mid \mathcal{F}_{t_k} \Big] = a(X_{t_k}^{(n)})\Delta t + O(\Delta t^2) \ . \tag{7}$$

Direct simulation often is inaccurate because most paths make a small contribution to the sum (3). For example, suppose $dS_t = \sigma S_t dW_t$, $S_0 = 1$, and we want $1 = \mathrm{E}_{1,0}[S_T]$. We saw in an earlier lesson that $S_t \to 0$ almost surely as $t \to \infty$, so most paths have $S_T \ll 1$, There are rare *outliers* with large $S_T \gg 1$ that make $1 = \mathrm{E}_{1,0}[S_T]$ possible.

You can get more accurate Monte Carlo estimates by cheating. The technical term is *importance sampling*. Instead of generating $X_t^{(n)}$ from your diffusion process and finding the sample mean, you simulate a different process $Y_t^{(n)}$. You find a quantity called the *likelihood ratio*, $L(Y_{[0,T]})$. This has the property that

$$\mathrm{E}_X[V(X_T)] = \mathrm{E}_Y\left[V(Y_T)L(Y_{[0,T]})\right] \ . \tag{8}$$

The importance sampling procedure is to generate many Y sample paths and use the importance sampling estimate

$$\widehat{f}_{\text{is}} = \frac{1}{N}\sum_{n=1}^{N} V(Y_T^{(n)})\, L(Y_{[0,T]}^{(n)}) \ . \tag{9}$$

The trick is to find a process Y so that makes the important events more likely. The method is more complicated, but the answer can be much more accurate.

The formula (8) is a relationship between two random processes called a *change of measure*. For diffusions, the change of measure formula is described by *Girsanov's theorem*. The theorem tells us that one diffusion can be related to another in the sense of (8) if and only if they have the same noise term. For diffusions it is possible to change the infinitesimal mean but not the infinitesimal variance. When two processes have the same infinitesimal variance, the formula for L is *Girsanov's formula*.

The quantity L in the change of measure formula (8) is the called the *Radon Nikodym derivative*. The L is for *likelihood ratio*. If the processes X_t and Y_t had probability densities, L would be the ratio. But probabilities in path space do not have probability densities, though many quantities related to paths do have densities (such as the density of X_t at a specific time t and the hitting time density). Instead, probabilities for diffusion processes are given by *probability measures*.

A probability measure assigns probabilities directly to events, rather than using a probability density. Suppose $X \in \mathbb{R}$ is a random variable with $u(x)$ for its probability density. Suppose $A \subseteq \mathbb{R}$ is some *event*. In probability, an event is just a set of outcomes. For example, the event that $0 \leq X \leq 1$ is represented by the set $A = [0, 1]$. If there is a probability density, then integration gives the probability of an event:

$$\Pr(A) = \int_{x \in A} u(x)\, dx \ .$$

In abstract probability, the probabilities of events are given without using a probability density. A system of probabilities $P(A)$ for all "reasonable" events A is called a *probability measure* if it has some natural properties of probabilities and is continuous (technically, *countably additive*) in a certain sense. It is possible that two probability measures are related by a likelihood ratio even if they are not given by probability densities. The Girsanov theorem for diffusions is one of those cases.

2 Direct simulation and Monte Carlo

Suppose X is some kind of random object, like a random path for instance, and $V(X)$ is a function of the path, and that we want to know

$$f = \mathrm{E}[\, V(X)\,] \ .$$

Suppose that we are able to create *samples*. For now, this means independent copies $X^{(n)}$ with the same distribution as X. In more sophisticated Monte Carlo, it may be impossible to make independent samples – take the Courant Institute class on Monte Carlo Methods and pay attention to Markov chain Monte Carlo (MCMC) if you're interested. For diffusion processes it is usually impossible to create paths with the X distribution exactly. Instead we make approximate paths using Euler's method (4). But, for now, forget these pieces of reality and suppose the $X^{(n)}$ have exactly the desired distribution and that they are exactly independent.

The direct estimator (3) is a Monte Carlo method for estimating $f = \mathrm{E}[V]$. The next step is the direct Monte Carlo *error bar*, which estimates the accuracy. For now, we use the simplified notation

$$V_n = V(X^{(n)}) \ .$$

The direct error bar comes from the central limit theorem applied to the direct estimate (3). If N is large, then the sample mean $\widehat{f}$ is approximately normal with mean f and variance

$$\mathrm{var}\left(\widehat{f}\right) = \sigma_{\widehat{f}}^2 = \frac{1}{N}\mathrm{var}(V) = \frac{1}{N}\sigma_V^2 \ . \tag{10}$$

Let $\xi \sim \mathcal{N}(0,1)$ be a standard normal random variable. Then $\widehat{f}$ approximately (for large N) has a representation

$$\widehat{f} \;\approx\; f + \frac{\sqrt{\sigma_V^2}}{\sqrt{N}}\,\xi \;.$$

We turn this around for the error bar as

$$f \;\approx\; \widehat{f} + \frac{\sqrt{\sigma_V^2}}{\sqrt{N}}\,\xi \;.$$

Don't worry that ξ seems to have the wrong sign. If ξ is standard normal, then $-\xi$ also is standard normal. This doesn't say what the error is, but it does say that the error size is on the order of $\frac{\sqrt{\sigma_V^2}}{\sqrt{N}}$. This is the *one standard deviation* error bar. A Monte Carlo result would be expressed as

$$f = \widehat{f} \pm \frac{\sqrt{\sigma_V^2}}{\sqrt{N}} \;. \tag{11}$$

For example, an estimate might be $f = 2.48 \pm .08$. This indicates that your best guess is 2.48 and that it's probably off by something like .8.

Usually, you have to estimate the standard deviation σ_V from the data. The number you need to estimate is

$$\sigma_V^2 = \mathrm{E}\Big[\,(V-f)^2\Big] \;.$$

A natural Monte Carlo estimate is

$$\widehat{\sigma_V^2} = \frac{1}{N}\sum_{n=1}^{N}\left(V_n - \widehat{f}\right)^2 \;. \tag{12}$$

Some people suggest $\frac{1}{N-1}$ instead of $\frac{1}{N}$, because it gives an *unbiased* estimate, which means the expected value of the estimate is the actual value:

$$\sigma_V^2 = \mathrm{E}\left[\frac{1}{N-1}\sum_{n=1}^{N}\left(V_n - \widehat{f}\right)^2\right] \;.$$

This is true, but the standard deviation is what goes in the error bar, not the variance. The square root is a nonlinear function and our estimate of the standard deviation is

$$\widehat{\sigma_V} = \sqrt{\widehat{\sigma_V^2}} \;.$$

If U is a positive random variable that is truly random, then

$$\mathrm{E}\Big[\sqrt{U}\Big] \neq \sqrt{\mathrm{E}[U]} \;.$$

Therefore, if $\mathrm{E}\Big[\widehat{\sigma_V^2}\Big] = \sigma_V^2$, then $\mathrm{E}[\widehat{\sigma_V}] \neq \sigma_V$.

Moreover, the difference between $\frac{1}{N}$ and $\frac{1}{N-1}$ is unimportant unless N is smaller than it should be for Monte Carlo. The *bias* of an estimator of a quantity A is $\mathrm{E}\Big[\widehat{A} - A\Big]$. The bias of $\widehat{\sigma_V^2}$ or $\widehat{\sigma_V}$ is order $\frac{1}{N}$, while the difference $|\widehat{\sigma_V} - \sigma_V|$ is on the order of $\frac{1}{\sqrt{N}}$. Correcting for the bias won't make the estimate significantly more accurate.

The big picture is the philosophy against spending lots of time making error bars precise. It is unprofessional to give Monte Carlo results without error bars. And it is a waste of time to make error bars very precise. They are a rough estimate of the error. "Don't put error bars on error bars."[2]

Summary of direct simulation Monte Carlo The problem is to estimate $\mathrm{E}\big[V(X_{[0,T]})\big]$. Here, X is the solution to an SDE (1). You choose computational parameters Δt, the time step for Euler's method (4), and N, the number of paths. You generate N paths. The total work is the number of paths times the number of time steps per path, which is $W = NT/\Delta t$. You compute $V_n = V(X^{(n)}$ and average (3). You compute the sample variance (12) and take the square root for the sample standard deviation. You report the estimate $\widehat{f}$ and the error bar $\widehat{\sigma_V}/\sqrt{N}$.

2.1 Histograms

A *histogram* is a graph of *bin counts*. Let $X^{(n)}$ be samples of a random variable. This could be from computer simulation or actual samples of something. A *bin* is an interval on the x axis whose *bin size* is the length Δx. We write B_k for bin k. It is convenient to take x_k as the center of B_k, so

$$B_k = \left[x_k - \tfrac{1}{2}\Delta x, x_k + \tfrac{1}{2}\Delta x\right] \ .$$

It is convenient in the mathematical discussion (but not in the code) not to specify the range of k or the location of x_0. In the code, there must be a largest and smallest k. The *bin count* N_k is the number of sample points in B_k:

$$N_k = \#\left\{n \mid X^{(n)} \in B_k\right\} \ .$$

A histogram is a plot of the bin counts.

Sometimes you plot the raw bin counts, but often you don't. You may be making the histogram to estimate the probability density $X^{(n)} \sim u(x)$. In this case the probability of a sample landing in bin k is (exactly or approximately)

$$\Pr\Big(X^{(n)} \in B_k\Big) = \int_{B_k} u(x)\,dx \approx \Delta x\, u(x_k) \ .$$

If you have N samples altogether, the expected count for bin k is N times the probability for one sample:

$$\mathrm{E}[N_k] = N\Pr\Big(X^{(n)} \in B_k\Big) \approx N\Delta x\, u(x_k) \ .$$

[2] A piece of advice from Malvin Kalos, one of the masters of Monte Carlo from his generation.

Some algebra turns this into an estimator of the probability:

$$\widehat{u(x_k)} = \frac{N_k}{\Delta x N} \ . \tag{13}$$

It may be more informative to plot $\widehat{u(x_k)}$ instead of the raw counts N_k. The difference is only a scaling of the y axis.

3 Importance sampling

Direct simulation simulation Monte Carlo is an impractical way to estimate $\mathrm{E}[V(X)]$ in many real applications. This is because values of X that contribute most to the expectation are very unlikely. Typical X values have much smaller $V(X)$ than the mean.

Geometric Brownian motion illustrates this. Consider the simple case of $\mu = 0$, $\sigma = 1$:

$$dS_t = S_t dW_t \ , \ \ S_0 = 1 \ .$$

This is a martingale so for all $t > 0$,

$$\mathrm{E}[S_t] = 1 \ .$$

But the solution formula is $S_t = e^{W_t - \frac{1}{2}t}$. For simulation, we can take $Z \sim \mathcal{N}(0,1)$ and take $W_t = \sqrt{t}Z$ (this has the same distribution as W_t, which is normal mean zero, variance t). This puts the formula is a more explicit form

$$S_t \sim e^{\sqrt{t}\,Z - \frac{1}{2}t} \ .$$

In order to have $S_t \geq 1$ (the mean value), we have to have

$$\sqrt{t}\,Z - \frac{1}{2}t \geq 0 \ .$$

This is

$$Z \geq \frac{1}{2}\sqrt{t} \ .$$

For example, with $t = 36$ it's $\Pr(Z > 3) \approx .0013$. If you simulated 1000 independent samples Z_k, the expected number of *hits* (samples with $Z_k > 3$, $S_{k,36} > 1$), is 1.3. The other 996.7 samples would be "wasted", contributing little to the expected value.

Importance sampling means changing the probability rules to make the important X values more likely – putting more X values in the region that is important for the expectation. This has to be done in a way that doesn't change the expectation value. If X has probability density $u(x)$, the trick is to find a different density v that puts samples where you want them, and then to

take into account the fact that you used the wrong density. Here is the algebra:

$$\begin{aligned}
\mathrm{E}_u[\,V(X)] &= \int_{-\infty}^{\infty} V(x)u(x)\,dx \\
&= \int_{-\infty}^{\infty} V(x)\frac{u(x)}{v(x)}\,v(x)\,dx \\
&= \int_{-\infty}^{\infty} V(x)L(x)\,v(x)\,dx\ , \quad L(x) = \frac{u(x)}{v(x)} \\
\mathrm{E}_u[\,V(X)] &= \mathrm{E}_v[\,V(X)L(X)]\ , \quad L(x) = \frac{u(x)}{v(x)}\ . \qquad (14)
\end{aligned}$$

In finance people imagine that there is a "$u-$world" where $X \sim u$ and a "$v-$world" where $X \sim v$. In the $u-$world, the number you want is $\mathrm{E}[\,V(X)]$. In the $v-$world, it's $\mathrm{E}[\,V(X)L(X)]$. A typical value of V or X in the $u-$world may be very different from typical values in the $v-$world. The *likelihood ratio* $L(x)$ makes the expectations equal.

The measure of success in importance sampling is *variance reduction*. You hope that the $v-$world variance is less than the original $u-$world variance. These variances are

$$\sigma_u^2(\,V(X))\ , \quad \text{and} \quad \sigma_v^2(\,V(X)L(X))\ .$$

In practical estimation, you can estimate the variances and see whether you decreased the variance. If you choose a bad strategy, then fancy $v-$world importance sampling can have a higher variance than direct $u-$world simulation.

Take the geometric Brownian motion example. If we want to make large Z more likely, we can sample from a Gaussian with a positive mean $Z \sim \mathcal{N}(\mu, 1)$. The likelihood ratio for this is (the random variable is z in this example)

$$\begin{aligned}
L(z) &= \frac{u(z)}{v(z)} \\
&= \frac{\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}}{\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(z-\mu)^2}} \\
&= \exp\left[\frac{z^2 - 2x\mu + \mu^2 - z^2}{2}\right] \\
&= e^{-z\mu}e^{\frac{1}{2}\mu^2}\ .
\end{aligned}$$

The importance sampling formula is

$$\mathrm{E}_{\mathcal{N}(0,1)}[\,V(Z)] = e^{\frac{1}{2}\mu^2}\mathrm{E}_{\mathcal{N}(\mu,1)}\left[\,V(Z)\,e^{-\mu Z}\right]\ . \qquad (15)$$

On the right side, we "pull" Z to the right by giving a mean $\mu > 0$. We "discount" the larger Z values with the discount factor $e^{-\mu Z}$. If you did this with $V = 1$, the expected value would go down because most of the samples would be discounted. The outside factor $e^{\frac{1}{2}\mu^2}$ fixes this effect, giving the exact answer even if $V = 1$.

4 Probability measure

The probabilities of paths that we use in stochastic calculus cannot be defined directly using probability densities. The expected values of random variables cannot be found directly by integration with respect to a probability density. The issue is that there is nothing in path space that is like dx in $\mathbb{R}^n$. Instead of integration on $\mathbb{R}^n$ with respect to dx, we integrate in probability space with a probability measure dP.

The first step is to define abstract probability measure and integration (expected value) with respect to a general (abstract) probability measure. The second step is to define the specific probability spaces and probability measures that are relevant for stochastic calculus. These are path space and versions of *Wiener measure*. This Stochastic Calculus class is not a course in abstract measure theory and integration any more than an ordinary Calculus class is a course on mathematical analysis. Many mathematical details are missing, as they are in an ordinary calculus class. Still, abstract probability measures seem to be the simplest way to understand some important topics such as importance sampling and change of measure for diffusion processes.

Probability measure is an abstract concept that forms the basis for most modern probability theory. Here is a superficial description of abstract measure based probability theory. A good graduate probability theory book has a more complete discussion. In the abstract approach, a "probability" consists of three things:

- A *probability space*, Ω. We think of this as the set of all possible "outcomes". A specific outcome is $\omega \in \Omega$.
- A $\sigma-$*algebra*, $\mathcal{F}$, of subsets of Ω. We think of $A \in \Omega$ as an *event*, which is a set of outcomes whose probability we know. We say A is *measurable* if $A \in \mathcal{F}$.
- A *probability measure*, P, which is a number $P(A) \in [0, 1]$. We think of $P(A)$ as the probability that the event A happens. In terms of random outcomes, $P(A) = \Pr(\omega \in A)$.

Probability theory requires $\mathcal{F}$ to be "complete" in a sense similar to the completeness of the real numbers. This makes the algebra a $\sigma-$algebra. The measure P must be "continuous" in the sense that the probability of a limit event is the limit of the probabilities. This is called *countable additivity*. The term *complete* in probability does not refer to the $\sigma-$algebra property, but to something more technical that is irrelevant in this course.[3]

[3] A probability is complete if any set of outcomes that "should" have probability zero does have probability zero. More technically, if $A \in \mathcal{F}$ and $P(A) = 0$, and if $B \subseteq A$, then $B \in \mathcal{F}$ and $P(B) = 0$. This may seem natural, but it is inconvenient in the common setting where we have two probability measures P_1 and P_2 with different events of probability zero. It is hard to have $\mathcal{F}_1 = \mathcal{F}_2$ (the same measurable events), when this happens. (Comment for experts: For $\Omega = [0, 1]$, this corresponds to using Borel measure, which is not complete, rather than the complete Lebesgue measure.)

Here are more details of $\sigma-$algebra and probability measure. It may help to look ahead to the examples if this seems too vague. Suppose $\mathcal{F}$ is a collection of subsets of Ω. We want $A \in \mathcal{F}$ to mean "we know whether $\omega \in A$". We say $\mathcal{F}$ is an *algebra* if it is *closed* under the operations of intersection (*and*), union (*or*), and complement (*not*). *Closed* means that doing one of the operations does not take you out of $\mathcal{F}$. For example, suppose $A_1 \in \mathcal{F}$ and $A_2 \in \mathcal{F}$. If we know whether $\omega \in A_1$ and whether $\omega \in A_2$, then it is reasonable to assume that we know whether $\omega \in A_1 \cap A_2$. The intersection $A_1 \cap A_2$ is the set of outcomes in A_1 and in A_2. The union $A_1 \cup A_2$ is the set of outcomes in A_1 or in A_2, or both. The complement A_1^c is the set of outcomes not in A_1. If we know whether $\omega \in A_1$, then we know whether $\omega \in A_1^c$, which is the same as $\omega \notin A_1$. Ordinary algebra (ordinary arithmetic, actually) has *binary* operations $+$, $*$ (operations on two numbers), and a *uniary* operation $-$ (taking the negative of a number). The algebra of sets has binary operations union ($\cup$) and intersection ($\cap$) and the uniary operation of complement ($A \to A^c$).

There may be subsets $A \subseteq \Omega$ so that we don't know whether $\omega \in A$ or not. Not every set is measurable. Two particular sets must be measurable, $A = \Omega$ and $A = \emptyset$ (the *empty set* is the set with no elements). This is natural from the "what we know" interpretation. We know whether $\omega \in \Omega$ (it is). We require that an algebra of sets satisfy the axiom $\Omega \in \mathcal{F}$. We know whether $\omega \in \emptyset$ (is isn't). We require that $\emptyset \in \mathcal{F}$. Note that if $\Omega \in \mathcal{F}$, then the complement axiom ($A \in \mathcal{F} \Longrightarrow A^c \in \mathcal{F}$) implies that $\emptyset = \Omega^c \in \mathcal{F}$.

A set algebra is a $\sigma-$algebra if it is closed under infinite sequences of union or intersection operations. If $A_n \in \mathcal{F}$ is an infinite sequence of measurable events, then the infinite union

$$A = \bigcup_{n=1}^{\infty} A_n$$

also has $A \in \mathcal{F}$. This is something like *completeness* of the real number system. Suppose $a_k > 0$ is a sequence of real numbers whose sum is bounded in the sense that

$$\sum_{k=1}^{n} a_k < C$$

for all n (there is a $C > 0$ so that ...). In the real number system, there is an S so that

$$S = \sum_{k=1}^{\infty} a_k \ .$$

This is not true in the rational numbers (fractions with integers on the top and bottom). For example, the Taylor series for e^x gives

$$e = \sum_{k=0}^{\infty} \frac{1}{k!} \ .$$

All the terms on the left are rational numbers, but the infinite sum, $e = 2.718..$ is not a rational number. The rational numbers are not complete because a

limit or an infinite sum of rational numbers may not be a rational number. A $\sigma-$algebra is a family of sets that includes set limits (unions and intersections of infinite sequences of sets).

A *measure* is a number $P(A)$ associated to each measurable $A \in \mathcal{F}$. For a *probability measure*, $0 \leq P(A) \leq 1$ for every $A \in \mathcal{F}$. This represents the probability that the event A happens, which is the probability that $\omega \in A$. A measure must be *additive*, which means that if $A_1 \in \mathcal{F}$ and $A_2 \in \mathcal{F}$, and if $A_1 \cap A_2 = \emptyset$ (A_1 and A_2 are *disjoint*), then $P(A_1 \cup A_2) = P(A_1) + P(A_2)$. This implies that if $A \subset B$ then $P(A) \leq P(B)$. This is because $B = A \cup (A^c \cap B$, and A is disjoint from $(A^c \cap B)$, and therefore $P(B) = P(A) + P(A^c \cap B) \geq P(A)$.

A measure is *countably additive* if it respects limits in the following way. Suppose $A_1 \subset A_2 \cdots$ is an "increasing" sequence of events. The "limit" event is the event is

$$A = \bigcup_{n=1}^{\infty} A_n \ .$$

The definition of countable additivity is that the probability of the limit event is the limit of the probabilities:

$$P(A) = \lim_{n \to \infty} P(A_n) \ . \tag{16}$$

From this abstract point of view, to make a probability model of random something you have to say what sets are measurable events and say how the probability of these events is defined. The usual way to define $\mathcal{F}$ is to define some sets that you want to be measurable (whose probability you want to define) and then say that $\mathcal{F}$ is the smallest $\sigma-$algebra that contains these events. This $\mathcal{F}$ will contain all the sets you said, and all limits of those, and limits of those, and so on. This $\sigma-$algebra is *generated* by the sets you give. Any collection of sets *generates* a $\sigma-$algebra.

Measure from a probability density. A probability density on $\mathbb{R}$ defines a probability measure with $\Omega = \mathbb{R}$ as its measure space. The $\sigma-$algebra is generated by all intervals $[a, b]$. The $\sigma-$algebra that contains these also contains infinite intervals. There are many ways to see this including

$$[0, \infty) = \bigcup_{n=0}^{\infty} [n, n+1] = \bigcup_{n=0}^{\infty} [0, n] \ .$$

It contains "open" intervals (intervals that do not contain the endpoints) such as (note: $(-\infty, a]^c = (a, \infty)$, etc.)

$$(a, b) = (-\infty, a]^c \cap [b, \infty)^c \ .$$

This $\sigma-$algebra is called the *Borel sets*.

If u is a probability density and A is a Borel set, then the probability measure is

$$P(A) = \int_A u(x) \, dx \ .$$

This definition works (we are not going to show) because it makes sense if A is an interval and because the integral respects limits.

Combining measures. Suppose P_1 and P_2 are two probability measures with the same probability space Ω and the same $\sigma-$algebra $\mathcal{F}$. Suppose that $q_1 \geq 0$ and $q_2 \geq 0$ and $q_1 + q_2 = 1$. Then there is a combined probability measure $P(A) = q_1 P_1(A) + q_2 P_2(A)$. You can check that P is countably additive and has $P(\Omega) = 1$ as a probability measure should. You can think of $\omega \sim P$ as first tossing a coin – with probability q_1 you take $\omega \sim P_1$ and otherwise you take $\omega \sim P_2$. You can take combinations of n measures, if you take measure P_k with probability q_k. This would be $P(A) = \sum_k q_k P_k(A)$. You can even take integral combinations, integrating P_t with respect to a probability density $u(t)$. That would be $P(A) = \int P_t(A)\, u(t)\, dt$.

Singular measures. Some measures on $\Omega = \mathbb{R}$ do not have a probability density. A measure like that is called *singular*, or, more properly, *singular with respect to Borel measure*. The most singular measure on $\mathbb{R}$ is a *point mass*, or *delta measure*. This is the measure that has all it's probability at the point $x = a$ and no probability at any other point. This is called δ_a. It is defined by $\delta_a(A) = 1$ if $a \in A$ and $\delta_a(A) = 0$ if $a \notin A$. This measure is countably additive, and checking this fact clarifies something about countable additivity. Suppose, for example A_n is the interval

$$A_n = \left[\frac{1}{n}, 1\right] .$$

This is an increasing family of events because $A_n \subset A_{n+1}$. All of them have $\delta_0(A_n) = 0$ because $0 \notin \left[\frac{1}{n}, 1\right]$. The limit (the union) of the A_n is

$$A = (0, 1] = \bigcup_{n=1}^{\infty} A_n .$$

Note that $0 \notin A$, because $(0, 1]$ does not include the left endpoint 0. The limit of the numbers $\frac{1}{n}$ is 0, but the union of the sets A_n does not include zero.

The *Dirac delta function*, which is written $\delta(x)$, is an informal way to express the singular measure δ_0. This function is infinite at $x = 0$ and zero elsewhere in a way that $\int_a^b \delta(x)dx = 1$ if $a < 0 < b$ and zero if $b < 0$ or $a > 0$. The point mass probability measure makes sense in any dimension and even in any probability space. Integral combinations of point mass measures give other singular measures.

For example, in 2d ($\Omega = \mathbb{R}^2$), define $a(t) = (\cos(t), \sin(t))$ and consider the probability measure

$$P = \frac{1}{2\pi} \int_0^{2\pi} \delta_{a(t)}\, dt .$$

This is a uniform density on the unit circle. If A is the event that (x, y) is in the "first quadrant" (i.e., $x > 0$ and $y > 0$), then $P(A) = \frac{1}{4}$, because one quarter of

the unit circle is in the first quadrant. IF $A \subseteq \mathbb{R}^2$ is any measurable event, then

$$P(A) = \Pr(\,(\cos(t), \sin(t)\,) \in A) \ .$$

This singular measure "lives" on the unit circle.

Continuous path space, diffusion measures. For this example, a *path* is a continuous function X_t defined for $0 \leq t \leq T$. The space of paths like this is written $C([0,T])$. The probability space is $\Omega = C([0,T])$. This is often called *path space*.

There can be several ways to generate a desired $\sigma-$algebra. The standard one for diffusions can be generated by events that depend on X at some time $t \in [0,T]$, such as

$$X_{[0,T]} \in A \ \text{ if } \ a \leq X_t \leq b \ .$$

By taking intersections we get events selected by criteria like

$$a_1 \leq X_{t_1} \leq b_1 \ \text{ and } \ a_2 \leq X_{t_2} \leq b_2 \ .$$

Taking intersections of an infinite sequence (because it's a $\sigma-$algebra) we can get the event[4]

$$X_t \geq 0 \ \text{ for all } \ t \in [0,T] \ .$$

5 Expectation and integration

In abstract probability, the expected value is the integral with respect to the probability measure. This section describes abstract measure-theoretic integration with respect to an abstract probability measure. Abstract probability measure is useful, in part, because this abstract integral is easy to define. Suppose Ω is a probability space with $\mathcal{F}$ and P, and that $V(\omega)$ is a function we want to integrate. The integral we need to define is

$$\mathrm{E}_P[V] = \int_\Omega V(\omega)\, dP(\omega) \ . \tag{17}$$

When $\Omega = \mathbb{R}$ (and ω is x), and $u(x)$ is a probability density, the abstract expectation (17) is the same as

$$\mathrm{E}_u[V] = \int_{-\infty}^{\infty} V(x)\, u(x)\, dx \ .$$

The abstract and concrete expectations should agree in the concrete setting.

[4]It is possible to put the positive rational numbers in into a single list. For example, you can make a list $(q_1, q_2, q_3, \cdots) = (1/1, 2/1, 2/2, 3/1, 3/2, 3/3, \cdots)$. It's OK for the list to have duplicates (like $1/1 = 2/2$). Take the event A_n to be $X_{q_n} \geq 0$. The intersection of the sequence $A = \cap A_n$ has $X \in A$ if and only if $X_q \geq 0$ for every positive rational number q. But X_t is a continuous function of t, so this implies that $X_t \geq 0$ for every t.

The two expressions for expectation are related through the informal identity

$$dP(x) = u(x)dx \ . \tag{18}$$

This says that the probability of a little bit of x space around x is equal to $u(x)$ multiplied by the length of that little bit. Earlier, we expressed this as $\Pr(x \leq X \leq x + dx) = u(x)dx$. In view of this, many people feel it's more natural to write $P(dx)$ than $dP(x)$. Either way, (18) expresses the probability measure P in terms of the "natural" measure, usually called *Lebesgue*[5] measure. For this class, the point of abstract probability measures is that there is no natural measure like dx to help define the probability measure for diffusions. There is a natural dx in $\mathbb{R}$ or $\mathbb{R}^n$, but not on $C([0,T])$. Diffusion measure is a probability measure without a probability density.

Think of finding the expected value of a function $V(x)$ when the random variable $X \sim u(x)$ is one dimensional with probability density u. The expectation is the area "under" the graph of a function $V(x)u(x)$,

$$\mathrm{E}[V] = I = \int_{-\infty}^{\infty} V(x)u(x)\,dx \ .$$

The Riemann integral approach is to divide the $x-$axis into pieces of size Δx. An $x-$point $x_k = k\Delta x$ has an approximate piece of area $A_k = \Delta x V(x_k)u(x_k)$. The Δx approximation to the total area is

$$I_{\mathrm{R},\Delta x} = \sum_k \Delta x V(x_k)u(x_k) \ .$$

The Riemann integral is the limit

$$I_{\mathrm{R}} = \lim_{\Delta x \to 0} I_{\mathrm{R},\Delta x} \ .$$

Measure theoretic approach to integration (outlined below) was invented because the limit is problematic if V is a "general" function (not continuous, not monotone, not the sum continuous and/or monotone functions).

The Riemann approach to integration has another disadvantage for general measure spaces: there is no analogue of little x intervals of length Δx, if you are integrating over a general probability space Ω. The trick to avoid this is to consider little intervals of length Δv on the $y-$axis instead. The $v-$points, which are $v_k = k\Delta v$, divide the $v-$axis into small pieces of height Δv. Define events

$$A_k = \{k\Delta y \leq V < (k+1)\Delta y\} \ .$$

The approximation $V(\omega) \approx v(x)$ is accurate with an error less than Δv in A_k. Therefore, the part of the expectation/integral over A_k is approximately

$$v_k P(A_k) \ .$$

[5]After the French mathematician who participated in developing measure theory, pronounced "luh-**beg**".

The Δv approximation to (17) comes from adding up these approximate integrals:

$$I_{a,\Delta v} = \sum_k v_k P(A_k) \ . \tag{19}$$

The measure-theoretic expectation/integral is the limit as $\Delta v \to 0$.

It is "easy" to show that the limit (19) exists. The first step is to make sure the approximations are defined, which is the hard part if there is a hard part. A function $V(\omega)$ is called *measurable* if sets defined by inequalities are measurable

$$L_v = \{\omega \mid V(\omega) < v\} \ , \ \ M_v = \{\omega \mid V(\omega) \leq v\} \ . \tag{20}$$

Measurable means that $L_v \in \mathcal{F}$ and $M_v \in \mathcal{F}$ for all v. The hypothesis $L_v \in \mathcal{F}$, informally, is that $P(V < v)$ is well defined. If the probabilities of the events, $V < v$ and $V \leq v$ are not defined, then (in this theory), $\mathrm{E}[V]$ is not defined either. Strict inequality, $P(V < v)$ can be different from non-strict inequality, $P(V \leq v)$ if P is a delta measure, or if the random variable V is constant a lot of the time.

In earlier lessons we replaced the general limit $\Delta t \to 0$ with the specific limit $\Delta t = 2^{-n}$, with $n \to \infty$. We use that philosophy here and define $\Delta v = 2^{-n}$ and take $n \to \infty$. With this trick, the step from n to $n+1$ means dividing an event A_k into two disjoint pieces:

$$A_k = B_k \cup C_k \ , \ \ B_k \cap C_k = \emptyset \ ,$$

with

$$B_k = \left\{ v_k \leq V < v_k + \tfrac{1}{2}\Delta v \right\} \ , \ \ C_k = \left\{ v_k + \tfrac{1}{2}\Delta v \leq V < v_{k+1} \right\} \ .$$

Note that the event $V = v_k + \frac{1}{2}\Delta v$ (if V lands exactly on the boundary between B_k and C_k) is assigned to C_k and is not in B_k. When you go from n to $n+1$, the contribution from A_k becomes the sum of contributions from B_k and C_k. The result changes a little:

$$v_k P(A_k) = v_k \left[P(B_k) + P(C_k) \right] \ \stackrel{n \to n+1}{\longrightarrow} \ v_k P(B_k) + (v_k + \tfrac{1}{2}\Delta v) P(C_k) \ .$$

The $n+1$ contribution is larger (technically, not smaller). Therefore the approximate integral (19) increases (doesn't decrease), but not by much (the last step uses $\sum P(C_k) \leq 1$):

$$\begin{aligned} I_{a,\frac{1}{2}\Delta v} &= \sum_k v_k P(B_k) + (v_k + \frac{1}{2}\Delta v) P(C_k) \\ &\leq I_{a,\Delta v} + \frac{1}{2}\Delta v \sum_k P(C_k) \\ &\leq I_{a,\Delta v} + \frac{1}{2}\Delta v \ . \end{aligned}$$

We have a sequence of approximations satisfies (writing n for $\Delta v_n = 2^{-n}$)

$$|I_{a,n+1} - I_{a,n}| \leq 2^{-n} \ .$$

We saw in an earlier lesson that $\sum 2^{-n} < \infty$ implies that

$$\mathrm{E}[V] = \int_\Omega V(\omega) dP(\omega) = I_a = \lim_{n \to \infty} I_{a,n} \tag{21}$$

exists.

This definition can also be given in terms of *simple functions*. The *indicator function* of an event $D \subseteq \Omega$ is

$$\mathbf{1}_D(\omega) = \begin{cases} 1 & \text{if } \omega \in D \\ 0 & \text{if } \omega \notin D \ . \end{cases}$$

The integral of an indicator function, hich is its expected value if $\omega \sim P$, should be

$$\int_\Omega \mathbf{1}_D(\omega) dP(\omega) = \mathrm{E}[\,\mathbf{1}_D] = P(D) \ .$$

A function $W(\omega)$ is a simple function if it takes only finitely many values. This is the same as saying there are events $D_j \subseteq \Omega$ and numbers w_j so that

$$W(\omega) = \sum_{j=1}^{M} w_j \mathbf{1}_{D_j}(\omega) \ .$$

The integral of a simple function should be

$$\int W(\omega) dP(\omega) = \sum_{j=1}^{M} w_j \int_\Omega \mathbf{1}_{D_j}(\omega) dP(\omega) = \sum_{j=1}^{M} w_j P(D_j) \ .$$

If $V \geq 0$ is any measurable function, and if W is a simple function with $W(\omega) \leq V(\omega)$ for all $\omega \in \Omega$, then we should have

$$\int_\Omega V(\omega) dP(\omega) \geq \int_\Omega W(\omega) dP(\omega) \ .$$

The definition (21) is equivalent to

$$\int_\Omega V(\omega) dP(\omega) = \sup \int_\Omega W(\omega) dP(\omega) \ ,$$

over all simple functions $W \leq V$. On the right, sup means *supremum*. This is like *maximum* except that the supremum may not be *attained*.[6] The approximations (19) are integrals of simple functions

$$W = \sum v_k \mathbf{1}_{A_k} \ .$$

[6]Suppose S is some collection of numbers. The supremum is the largest number you can get as a limit of numbers $s \in S$. It is a theorem in mathematical analysis that if S is *bounded* (there is some t with $s \leq t$ for all $s \in S$), then S has a supremum. For example, the supremum of the numbers $1 - \frac{1}{n}$ is 1, which is not attined because $1 - \frac{1}{n} < 1$ for all n. If S is not bounded we say the supremum is ∞.

It often happens that

$$\int_\Omega V(\omega) dP(\omega) = \infty \ .$$

In this case we say the integral *diverges*. If the integral is finite, we sometimes say V is *integrable*. This is not to be confused with the term *measurable*, which refers to the level sets of V.

If V has both negative and positive values, we write $V_+(\omega) = \max(V(\omega), 0)$ and $V_-(\omega) = |V(\omega) - V_+(\omega)|$ for the *positive part* and *negative part* of V (some people define the negative part without $|\cdot|$ to be negative). If V_+ and V_- are integrable (finite integrals), then we say that V is integrable and define the integral as

$$\int_\Omega V(\omega) dP(\omega) = \int_\Omega V_+(\omega) dP(\omega) - \int_\Omega V_-(\omega) dP(\omega) \ .$$

The condition that V_+ and $V + -$ are integrable is the same as the condition that $|V|$ is integrable.

In probability language, suppose X is a random variable. A mathematical probabilist would say that the expected value of X exists if

$$\mathrm{E}[\,|X|] < \infty \ .$$

The expected value is

$$\mu_X = \mathrm{E}[X] \ .$$

The *Kolmogorov strong law of large numbers* says that if $|X|$ is integrable, and if X_n are independent "copies" of X (independent with the same probability distribution), then the sample means converge to μ_X almost surely. The sample means are

$$S_n = \frac{1}{n} \sum_{k=1}^{n} X_k \ .$$

The strong law says

$$S_n \to \mu_X \ \text{ as } \ n \to \infty \ \text{ almost surely } \ .$$

The hypothesis $|X| < \infty$ is crucial. Consider the Cauchy random variable with $u(x) = \frac{1}{\pi} \frac{1}{1+x^2}$. It may seem that $\mathrm{E}[X] = 0$ by symmetry, but

$$\mathrm{E}[|X|] = \frac{2}{\pi} \int_0^\infty \frac{x}{1+x^2} \, dx = \infty \ .$$

The sample means of a Cauchy random variable do not converge at all (a homework exercise).

6 Change of measure in diffusions