Stochastic Calculus, Courant Institute, Fall 2019
http://www.math.nyu.edu/faculty/goodman/teaching/StochCalc2019/StochCalc.html
Jonathan Goodman, September, 2019

Class 13, change of measure

1 Introduction

Change of measure is a deep subject with many practical applications. Two probability distributions P and Q may be related by a *likelihood ratio* L. A random variable X can have either the P or the Q distributions. We write expectations as

$$\mathrm{E}_P[\, V(X)] \ , \ \ \mathrm{E}_Q[\, V(X)] \ .$$

For example, P could represent ordinary Brownian motion and Q could represent Brownian motion with drift. Then $\mathrm{E}_P[\cdot]$ would mean expectation, assuming X is a standard Brownian motion path. The P and Q expectations generally are different, but they are related by the likelihood ratio L if, for "any" function V,

$$\mathrm{E}_P[\, V(X)] = \mathrm{E}_Q[\, V(X)L(X)] \ . \tag{1}$$

The likelihood ratio compensates for the difference between the P and Q distributions.

Importance sampling is one practical application of change of measure. It is common that "most" of the samples with large V in the P distribution are rare in the P distribution. Suppose you estimate $A = \mathrm{E}_P[\, V(X)]$ by simulation. This would mean generating many samples (sample paths if it's Brownian motion) and averaging the results

$$\widehat{A}_N = \frac{1}{N} \sum_{k=1}^{N} V(X^{(k)}) \ , \ \ X^{(k)} \sim P \ \text{ i.i.d.} \tag{2}$$

If "important" paths (paths with large V) are rare, then the samples you get might have few of none of them, $\widehat{A}_N$ not close to A. Importance sampling means designing a different probability distribution Q that makes "important" samples more likely. The more complicated, but hopefully more accurate importance sampling estimator is

$$\widehat{A}_N = \frac{1}{N} \sum_{k=1}^{N} V(X^{(k)})L(X^{(k)}) \ , \ \ X^{(k)} \sim Q \ \text{ i.i.d.} \tag{3}$$

The fake distribution creates more "hits" (samples $X^{(k)}$ with large $V(X^{(k)})$. You compensate for this by giving the hits less weight – $L(X)$ will be small. This situation is surprisingly common and importance sampling is used a lot.

Sensitivity analysis is a related practical application. You may have similar probability distributions P and Q. A finance example might be P and Q geometric Brownian motions with the same volatility and slightly different growth rates. You may be interested in the difference

$$D_{PQ} = A_P - A_Q = \mathrm{E}_P[\, V(X)] - \mathrm{E}_Q[\, V(X)] \ .$$

One approach to this would be to estimate A_P with N samples $X^{(k)} \sim P$ and N different samples $\widetilde{X}^{(k)} \sim Q$. This approach may fail if the statistical estimation error is larger than D_{PQ}. A different approach is to use one set of samples $X^{(k)} \sim P$ and re-weight them to estimate A_Q.

$$\widehat{D}_{PQ} = \frac{1}{N} \sum_{k=1}^{N} \left[V(X^{(k)}) - V(X^{(k)}) L(X^{(k)}) \right] \ . \tag{4}$$

If P is close to Q, then L will be close to 1 and the terms on the right of (4) will be smaller than the terms on the right of (2) or (3). In this context, the likelihood ratio function $L(x)$ may be called the *score function*. Sensitivity analysis in finance is often done this way, or should be.

A more theoretical use is understanding when things happen almost surely. The change of measure formula (1) implies (we will see) that something happens almost surely in the P distribution if and only if it happens almost surely in the Q distribution. For example, we showed that the $\Delta t \to 0$ (actually $m \to \infty$) limit that defines the Ito integral converges almost surely for Brownian motion. Now we learn that it converges almost surely for any process related to Brownian motion by a change of measure. This may seem surprising in view of the proof from Class 12. There, it was important that $\mathrm{E}[\Delta W] = 0$. But Brownian motion with drift has $\mathrm{E}[\Delta W] \neq 0$. The change of measure theorem implies that the Ito integral is defined for Brownian motion with drift.

We say that P and Q are *equivalent* probability distributions if they are related by a likelihood ratio as in (1). Be careful here, as "equivalent" distributions are not identical. If P and Q have a different idea of what happens almost surely, then P and Q are not equivalent. For example, the quadratic variation of Brownian motion is

$$[W]_t = t \ .$$

Another process with a different quadratic variation is not equivalent to Brownian motion.

2 Importance sampling in finite dimensions

Before we turn to probability distributions on paths (diffusions), explain how re-weighting works with probability densities in finite dimensional problems. Take $X \in \mathbb{R}^n$ to be a random variable with n components. Suppose there are two probability densities $p(x)$ and $q(x)$. Let $p(x)$ represent the P distribution

and $q(x)$ the Q distribution. Then

$$\mathrm{E}_P[\, V(X)] = \int_{\mathbb{R}^n} V(x)p(x)\, dx \ .$$

The likelihood ratio representation (1) asks to express the simple P representation as a Q representation with the L factor. We do this by dividing then multiplying by $q(x)$. The likelihood ratio turns out to be the ratio of probability densities:

$$L(x) = \frac{p(x)}{q(x)} \ . \tag{5}$$

The calculation is

$$\begin{aligned} \mathrm{E}_P[\, V(X)] &= \int_{\mathbb{R}^n} V(x)p(x)\, dx \\ &= \int_{\mathbb{R}^n} V(x)\frac{p(x)}{q(x)}q(x)\, dx \\ &= \int_{\mathbb{R}^n} V(x)L(x)q(x)\, dx \\ &= \mathrm{E}_Q[\, V(X)L(X)] \ . \end{aligned}$$

This shows that you can "change measure" from P to Q if $q(x) \neq 0$ for any x with $p(x) \neq 0$. For example, you can change $P = \mathcal{N}(0,1)$ to $Q = \mathcal{N}(\mu, \sigma)$ for any desired μ and σ. You cannot change from $P = \mathcal{N}(0,1)$ to Q being an exponential random variable because the exponential density is different from zero only for $x > 0$.

As a toy example, consider the probability that a Gaussian

3 Probability measure and expectation

The probability distributions of diffusion processes are not given by probability densities. Instead they are defined by approximations somewhat like the approximations used to define the Ito integral in Class 12. The random outcome (the random object?) for a diffusion process is a path $X_{[0,T]}$. An event is a set of paths. If $\mathcal{F}$ is a $\sigma-$algebra and $A \in \mathcal{F}$ is an event measurable with respect to $\mathcal{F}$, then we write $P(A)$ for the probability that $X_{[0,T]}$ is in A.

A *probability measure* is a function $P(A)$. For each $A \in \mathcal{F}$, there is a number $P(A)$. This represents the probability of the event A. It is a probability measure if it satisfies four basic axioms. First, $P(\Omega) = 1$. Second, $P(A) \geq 0$ for any event A (probabilities cannot be negative). The probability space Ω must include every (or "almost every") possible outcome. Third, if $A \cap B = \emptyset$ [The empty set is $\emptyset$, which is the set with no elements. This is a way to say that A and B are *disjoint*, that there are no elements in both A and B.] then $P(A \cup B) = P(A) + P(B)$. For example, A and the complement A^c are disjoint, and $A \cup A^c = \Omega$. Therefore $P(A) + P(A^c) = 1$, which may use in the form $P(A) = 1 - P(A^c)$.

Fourth, a measure must be *countably additive*. This means that if A_k is a family of events that is "pairwise disjoint" ($A_j \cap A_k = \emptyset$ for $j \neq k$), then

$$\sum_{k=1}^{\infty} P(A_k) = P(\cup_{k=1}^{\infty} A_k) \ . \tag{6}$$

Here is another form of countable additivity you may see. A family of events B_k is "increasing" if $B_k \subseteq B_{k+1}$ for all k. You can think of the "limit" of the B_k is the union. Countable additivity for an increasing family is

$$\lim_{k\to\infty} P(B_k) = P(\cup_{k=1}^{\infty} B_k) \ . \tag{7}$$

The two forms of countable additivity are equivalent to each other – each implies the other. Suppose P is finitely additive (the third axiom) and

$$B_k = A_1 \cup \cdots \cup A_k \ .$$

Finite additivity implies that

$$\begin{aligned} P(B_k) &= P(A_1 \cup \cdots \cup A_k) \\ &= P([A_1 \cup \cdots \cup A_{k-1}] \cup A_k) \\ &= P(B_{k-1}) + P(A_k) \\ &= [(B_{k-2}) + P(A_{k-1})] + P(A_k) \\ &\ \ \vdots \\ P(B_k) &= P(A_1) + \cdots + P(A_k) \ . \end{aligned}$$

Since all of these numbers are non-negative,

$$\lim_{k\to\infty} P(B_k) = \sum_{k=1}^{\infty} P(A_k) \ .$$

Also (you might have to think about this for a minute)

$$B = \cup_{k=1}^{\infty} B_k = \cup_{k=1}^{\infty} A_k \ .$$

Therefore, each of the formulas (6) and (7) implies the other.

The word "measure" comes from "measure theory", which began as a study of which sets $A \in \mathbb{R}$ or $A \in \mathbb{R}^d$ could be assigned a length or area or volume (a *measure*). The point there (as it turned out) was to find a family of measurable sets that formed a $\sigma-$algebra and included basic sets like intervals or balls. The measure of an interval or ball is the length or volume. The measures of the rest of the measurable sets (it turned out) is determined by countable additivity. A class on "real variables" usually starts with a proof of these facts – countable additive measure on $\mathbb{R}^d$ with a $\sigma-$algebra of measurable sets, so that the measure of a "simple" set is what it's supposed to be. Kolmogorov realized

that abstract countably additive probability measures allow for a useful abstract theory of probability without probability density functions.

Measure theory allows for a simple definition of the integral. In the Riemann integral, you define Δx and $x_k = k\Delta x$ and then

$$\int_a^b f(x)dx \approx \sum_{a<x_k<b} f(x_k)\Delta x \ .$$

The left side has a limit as $\Delta t \to 0$, which is the definition of the right side. This can be done "sideways" by defining a $\Delta y > 0$. Then divide the $y-$axis into intervals of length Δy with endpoints $y_k = k\Delta y$. The sets A_k are the sets on the $x-$axis that go to these intervals on the $y-$axis:

$$A_k = \{x \mid y_k \le f(x) < y_{k+1}\} \ .$$

If $f(x) \ge 0$ for all x, you can approximate the integral as

$$\int f(x)dx \approx \sum_{k=0}^{\infty} y_k \, (\text{ measure of } A_k) \ .$$

The right side has a limit as $\Delta y \to 0$, which is the definition of the left side.

In abstract probability, we use $\omega \in \Omega$ to represent the basic outcome. If $\Omega = \mathbb{R}^d$ (probability densities), then ω is a point in $\mathbb{R}^d$. We usually denote this by x. If we're talking about Brownian motions or diffusions, then Ω is the set of paths (functions of t), and ω represents a path. In this setting, expected values are defined as abstract integrals. If $V(\omega)$ is a function of the random outcome $\omega \in \Omega$ (often called a "random variable"), then the expected value is written using one of two equivalent notations,

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

This is defined if V is non-negative and measurable.

The definition is as before. Define events

$$A_k = \{\omega \mid y_k \le V(\omega) < y_{k+1}\} \ .$$

Every $\omega \in \Omega$ is in one of the A_k and the A_k are pairwise disjoint (because of $\le$ on one side and $<$ on the other). Choose a positive integer m and define $\Delta y_m = 2^{-m}$. The approximate expectation is

$$\mathrm{E}_P^{(m)}[\, V(\omega)] = \sum_{k=0}^{\infty} y_k P(A_k) \ . \tag{8}$$

From the definition it is "clear" (think about it for a few minutes, draw a picture) that

$$\mathrm{E}_P^{(m)}[\, V(\omega)] \le \mathrm{E}_P^{(m+1)}[\, V(\omega)] \le \mathrm{E}_P^{(m)}[\, V(\omega)] + \tfrac{1}{2}2^{-m} \ .$$

Therefore the limit in this definition exists:

$$\mathrm{E}_P[\,V(\omega)] = \lim_{m\to\infty} \mathrm{E}_P^{(m)}[\,V(\omega)] \ . \tag{9}$$

It is possible (it happens a lot in probability) that $\mathrm{E}_P[\,V(\omega)] = \infty$. If V is not positive, we can define

$$V(\omega) = V_+(\omega) - V_-(\omega) \ , \ \ V_\pm(\omega) \geq 0 \ \text{ for all } \omega \in \Omega \ .$$

These are the "positive" (non-negative) and "negative" parts of V. If

$$\mathrm{E}_P[\,V_-(\omega)] < \infty \ ,$$

we can define

$$\mathrm{E}_P[\,V(\omega)] = \mathrm{E}_P[\,V_+(\omega)] - \mathrm{E}_P[\,V_-(\omega)] \ .$$

If $\mathrm{E}_P[\,V_+(\omega)] = \infty$ then $\mathrm{E}_P[\,V(\omega)] = \infty$. It doesn't seem possible to define the overall expectation if both the positive and negative parts have infinite expectation. We can define $\infty - (\text{finite}) = \infty$, but $\infty - \infty = ?$.

This definition of integral implies the "obvious" formula, for a measurable event A,

$$\mathrm{E}_P[\,\mathbf{1}_A(\omega)] = P(A) \ .$$

Here $\mathbf{1}_A(\omega)$ is the *indicator function* which is 1 if $\omega \in A$ and 0 if $\omega \notin A$. In the definition (9), the sets A_k will be empty except for the k with $y_k = 1$. That one will have $y_k = 1$ (DUH) and $A_k = A$. The definition (8) and (9) makes expectation linear, in that

$$\mathrm{E}_P[\,V_1(\omega)] + c\mathrm{E}_P[\,V_2(\omega)] = \mathrm{E}_P[\,V_1(\omega) + cV_2(\omega)] \ .$$

Any "countably additive" version of this for infinite sums has an extra hypothesis. Examples are the *monotone convergence theorem* (all $V_k \geq 0$) and the *dominated convergence theorem*, which has the hypothesis that involves the *maximal function* $M(\omega)$ defined by

$$M(\omega) = \max_n \left| \sum_{k=1}^n V_k(\omega) \right| \ .$$

The hypothesis is

$$\mathrm{E}_P[\,M(\omega)] < \infty \ .$$

A function is a *simple function* if it may be represented as a finite linear combination of indicator functions

$$V(\omega) = \sum_{k=1}^n u_k \mathbf{1}_{A_k}(\omega) \ .$$

The expectation (using the one indicator function property and finite linearity) is

$$\mathrm{E}_P[\,V(\omega)] = \sum_{k=1}^n u_k P(A_k) \ . \tag{10}$$

Many modern textbooks base the definition of abstract integration/expectation on this. You approximate your function by a simple function, define the integral for simple functions, and take a limit. Our definition (8) and (9) is like this (you can check), where the approximating functions are

$$V(\omega) \approx V^{(m)}(\omega) = \sum_{k=0}^{m} y_k \mathbf{1}_{A_k}(\omega) \; .$$

These definitions have the following "obvious" property. [Here "obvious" often means a property the definition should have. If the definition does not have the property, then the definition is wrong, or our preconception was wrong.] As before, $V = 0$ almost surely if $P(V \neq 0) = 0$, or if $A = \{\omega \mid V(\omega) \neq 0\}$ implies $P(A) = 0$. If $V = 0$ almost surely, then $\mathrm{E}_P[\, V(\omega)] = 0$. This is a feature of our definition, because (8) has $P(A_k) = 0$ except A_0.

4 Likelihood ratio change of measure

We say that measures P and Q are *equivalent* if there is a likelihood function $L(\omega)$ so that for "any" function V,

$$\mathrm{E}_P[\, V(X)] = \mathrm{E}_Q[\, V(X)L(X)] \;\; , \;\; \mathrm{E}_Q[\, V(X)] = \mathrm{E}_P\left[V(X) \frac{1}{L(X)} \right] \; . \tag{11}$$

The *Radon Nykodim* theorem says that two probability measures are equivalent in this sense if they agree on what "almost surely" means. That is,

$$P(A) \iff Q(A) = 0 \;\; \text{for any measurable event } A \; . \tag{12}$$

This theorem is easy to prove in the "easy direction" (which is the definition of "easy direction"). If P and Q are equivalent by the L definition (11), and if $Q(A) = 0$, then

$$P(A) = \mathrm{E}_P[\, \mathbf{1}_A(\omega)] = \mathrm{E}_Q[\, \mathbf{1}_A(\omega) L(\omega)] = 0 \; .$$

We know the last expectation is zero because $V(\omega) = \mathbf{1}_A(\omega)L(\omega) = 0$ almost surely with respect to Q ($V(\omega) \neq 0$ only if $\omega \in A$, which it almost surely is not in Q). The hard direction is the theorem that the of P and Q agree on "almost surely", then there is a likelihood ratio L with the property (11). That argument takes too long to include here.

In finite dimensions with probability densities $p(x)$ and $q(x)$, the probability distributions are equivalent if the places $p \leq 0$ and $q \neq 0$ allow it. For example, any two Gaussians are equivalent but a Gaussian is not equivalent to an exponential. If $q(x)$ is the PDF of an exponential random variable, then $Q(X < 0) = 0$. If P is Gaussian, then $P(X < 0) \leq 0$.

For measures in path space, equivalence is more subtle. The theorem for diffusions, *Girsanov's theorem*, is that they are equivalent if the have the same

infinitesimal variance. For example there is a probability distribution for ordinary Brownian motion

$$dX = dW \ ,$$

and a distribution for Brownian motion with a different infinitesimal drift and variance

$$dX = a dt + b \, dW \ .$$

These two processes are equivalent if and only if $b^2 = 1$ (the same infinitesimal variance).

An easy part of Girsanov's theorem concerns a measure P on path space determined by the diffusion

$$dX = a(X)dt + b(X)dW \ , \ \ X_0 = 0 \ .$$

This is not equivalent to Brownian motion measure $dX = dW$, $X_0 = 0$, unelss $b^2 = 1$. The proof involves quadratic variation. In the P measure, almost surely,

$$[X]_t = \lim_{\Delta t \to 0} \sum_{t_k < t} \left(X_{t_{k+1}} - X_{t_k} \right)^2 = \int_0^t b^2(X_s) ds \ .$$

In the Q (Brownian motion) measure, it is

$$[X]_t = \lim_{\Delta t \to 0} \sum_{t_k < t} \left(X_{t_{k+1}} - X_{t_k} \right)^2 = t \ .$$

Therefore, unless $b^2(X_t) = 1$ almost surely, we can tell the paths apart (see next paragraph).

You can understand equivalence of measures from the point of view of hypothesis testing in statistics. In hypothesis testing, you are given a random sample and asked to judge whether the null hypothesis or the alternative hypothesis is true. Usually you don't know with certainty, but you can say what is likely to be true. The null hypothesis (usually called H_0) is that data came from probability measure Q. The alternative hypothesis (called H_1) is that the data came from measure P. A hypothesis test procedure is equivalent to an event A, which is the set of outcomes that are classified as H_0. The event A may be defined implicitly by the hypothesis testing procedure you use to decide H_0 or H_1. The test is perfect if $Q(A) = 1$ and $P(A) = 0$. That means that the hypothesis test is correct (almost surely). In finite dimensions, two probability measures are likely to be equivalent so there can be no perfect test.

With paths, you have one path $X_{[0,t]}$ and you are asked whether is it Brownian motion or some other diffusion process. To decide, you compute the quadratic variation for your path and find b. If $b^2 \neq 1$ you know it is not Brownian motion. If H_0 is Brownian motion and H_1 is another diffusion with $b^2 \neq 1$, there is a perfect test.

5 Girsanov – change of measure from Brownian motion

The hard part of Girsanov's theorem is that Brownian motion is equivalent to "any other" diffusion with $b = 1$. That is, you can change the drift with a likelihood ratio, but not the infinitesimal variance. This part of Girsanov's theorem comes with *Girsanov's formula* for the likelihood ration function L.

If were were a PDF for P and Q, then we could divide to get L. Without a PDF we choose a Δt and look at the "observations"

$$X_k^{(\Delta t)} = X_{t_k} \ , \ \ 1 \leq t_k \leq t \ .$$

These observations form a random variable with finitely many components

$$X^{(\Delta t)} = (X_1^{(\Delta t)}, X_2^{(\Delta t)}, \ldots,) \ .$$

Let $p^{(\Delta t)}$ and $q^{(\Delta t)}$ be the corresponding probability densities. (Actually, we will cheat with p and write an approximation to it.) The approximate likelihood ratio is

$$L^{(\Delta t)}(x) = \frac{p^{(\Delta t)}(x^{(\Delta t)})}{q^{(\Delta t)}(x^{(\Delta t)})} \ .$$

We will see that $L^{(\Delta t)}$ has a limit (almost surely) as $\Delta t \to 0$.

The PDF of the observations of Brownian motion is given by the independent increments property of Brownian motion

$$q(x_1, x_2, \ldots) = q(x_1) \cdot q(x_2|x_1) \cdot q(x_3|x_2, x_1) \cdot \cdots \ .$$

Each of the conditional expectations is that X_{k+1} is normal with mean X_k and variance Δt. Therefore

$$\begin{aligned} q(x_1) &= \frac{1}{2\pi\Delta t} e^{-\frac{x_1^2}{2\Delta t}} \\ q(x_2|x_1) &= \frac{1}{2\pi\Delta t} e^{-\frac{(x_2 - x_1)^2}{2\Delta t}} \\ q(x_3|x_2, x_1) = q(x_3|x_2) &= \frac{1}{2\pi\Delta t} e^{-\frac{(x_3 - x_2)^2}{2\Delta t}} \ . \end{aligned}$$

As a result (n is the largest k with $t_k < t$),

$$q(x_1, x_2, \ldots) = \left(\frac{1}{2\pi\Delta t} \right)^n e^{-\frac{1}{2\Delta t}\left[\sum_{t_k < t} (x_k - x_{k-1})^2\right]} \ . \tag{13}$$

This formula includes x_0, which is equal to 0.

With drift, $X_{k+1} \approx X_k + a(X_k)\Delta t + \Delta W_k$. We write the PDF for X_{k+1} assuming this is exact. This means that X_{k+1} is normal with mean $X_k +$

$\Delta t a(X_k)$ and variance Δt. The conditional probabilities are

$$\begin{aligned} p(x_1) &= \frac{1}{2\pi\Delta t} e^{-\frac{(x_1 - a(0)\Delta t)^2}{2\Delta t}} \\ p(x_2|x_1) &= \frac{1}{2\pi\Delta t} e^{-\frac{(x_2 - x_1 - a(x_1)\Delta t)^2}{2\Delta t}} \\ p(x_3|x_2, x_1) = q(x_3|x_2) &= \frac{1}{2\pi\Delta t} e^{-\frac{(x_3 - x_2 - a(x_2)\Delta t)^2}{2\Delta t}} \ . \end{aligned}$$

This leads to

$$p(x_1, x_2, \ldots) = \left(\frac{1}{2\pi\Delta t}\right)^n e^{-\frac{1}{2\Delta t}\left[\sum_{t_k < t}(x_k - x_{k-1} - a(x_{k-1})\Delta t)^2\right]} \ . \qquad (14)$$

We calculate the ratio.

First, note that the 2π factors are the same in p and q, so they cancel in the quotient. Then calculate a term in the exponent:

$$-\frac{1}{2\Delta t}(x_k - x_{k-1} - a(x_{k-1})\Delta t)^2 = -\frac{1}{2\Delta t}\left[(x_k - x_{k-1})^2 - 2(x_k - x_{k-1})a(x_{k-1})\Delta t + a(x_{k-1})^2\Delta t^2\right]$$