Stochastic Calculus Notes, Lecture 4

Last modified October 4, 2004

1 Continuous probability

1.1. Introduction: Recall that a set Ω is *discrete* if it is finite or countable. We will call a set *continuous* if it is not discrete. Many of the probability spaces used in stochastic calculus are continuous in this sense (examples below). Kolmogorov[1] suggested a general framework for continuous probability based on abstract integration with respect to abstract probability measures. The theory makes it possible to discuss general constructions such as conditional expectation in a way that applies to a remarkably diverse set of examples.

The difference between continuous and discrete probability is the difference between integration and summation. Continuous probability cannot be based on the formula

$$P(A) = \sum_{\omega \in A} P(\omega) \; . \qquad (1)$$

Indeed, the typical situation in continuous probability is that any event consisting of a single outcome has probability zero: $P(\{\omega\}) = 0$ for all $\omega \in \Omega$.

As we explain below, the classical formalism of probability densities also does not apply in many of the situations we are interested in. Abstract probability measures give a framework for working with probability in path space, as well as more traditional discrete probability and probabilities given by densities on R^n.

These notes outline the Kolmogorov's formalism of probability measures for continuous probability. We leave out a great number of details and mathematical proofs. Attention to all these details would be impossible within our time constraints. In some cases we indicate where a precise definition or a complete proof is missing, but sometimes we just leave it out. If it seems like something is missing, it probably is.

1.2. Examples of continuous probability spaces: Be definition, a *probability space* is a set, Ω, of possible outcomes, together with a $\sigma-$algebra, $\mathcal{F}$, of measurable events. This section discusses only the sets Ω. The corresponding algebras are discussed below.

R, the real numbers. If x_0 is a real number and $u(x)$ is a probability density, then the probability of the event $B_r(x_0) = \{x_0 - r \leq X \leq x_0 + r\}$ is

$$P([x_0 - r, x_0 + r]) = \int_{x_0 - r}^{x_0 + r} u(x)dx \to 0 \;\text{ as } r \to 0.$$

[1] The Russian mathematician Kolmogorov was active in the middle of the 20^{th} century. Among his many lasting contributions to mathematics are the modern axioms of probability and some of its most important theorems. His theories of turbulent fluid flow anticipated modern fractals be several decades.

Thus the probability of any individual outcome is zero. An event with positive probability ($P(A) > 0$) is made up entirely of outcomes $x_0 \in A$, with $P(x_0) = 0$. Because of countable additivity (see below), this is only possible when Ω is uncountable.

R^n, sequences of n numbers (possibly viewed as a row or column vector depending on the context): $X = (X_1 \ldots, X_n)$. Here too if there is a probability density then the probability of any given outcome is zero.

$\mathcal{S}^{\mathcal{N}}$. Let $\mathcal{S}$ be the discrete state space of a Markov chain. The space $\mathcal{S}^T$ is the set of sequences of length T of elements of $\mathcal{S}$. An element of $\mathcal{S}^T$ may be written $x = (x(0), x(1), \cdots, x(T-1))$, with each of the $x(t)$ in $\mathcal{S}$. It is common to write x_t for $x(t)$. An element of $\mathcal{S}^{\mathcal{N}}$ is an infinite sequence of elements of $\mathcal{S}$. The "exponent" $\mathcal{N}$ stands for "natural numbers". We misuse this notation because ours start with $t = 0$ while the actual natural numbers start with $t = 1$. We use $\mathcal{S}^{\mathcal{N}}$ when we ask questions about an entire infinite trajectory. For example the hitting probability is $P(X(t) \neq 1 \text{ for all } t \geq 0)$. Cantor proved that $\mathcal{S}^{\mathcal{N}}$ is not countable whenever the state space has more than one element. Generally, the probability of any particular infinite sequence is zero. For example, suppose the transition matrix has $P_{11} = .6$ and $u_0(1) = 1$. Let x be the infinite sequence that never leaves state 1: $x = (1, 1, 1, \cdots)$. Then $P(x) = u_0(1) \cdot .6 \cdot .6 \cdots$. Multiplying together an infinite number of .6 factors should give the answer $P(x) = 0$. More generally, if the transition matrix has $P_{jk} \leq r < 1$ for all (j, k), then $P(x) = 0$ for any single infinite path.

$C([0,T] \to R)$, the path space for Brownian motion. The C stands for "continuous". The $[0,T]$ is the time interval $0 \leq t \leq T$; the square brackets tell us to include the endpoints (0 and T in this case). Round parentheses $(0,T)$ would mean to leave out 0 and T. The final R is the "target" space, the real numbers in this case. An element of Ω is a continuous function from the interval $[0,T]$ to R. This function could be called $X(t)$ or X_t (for $0 \leq t \leq T$). In this space we can ask questions such as $P(\int_0^T X(t)dt > 4)$.

1.3. Probability measures: Let $\mathcal{F}$ be a $\sigma-$algebra of subsets of Ω. A *probability measure* is a way to assign a probability to each event $A \in \mathcal{F}$. In discrete probability, this is done using (1). In R^n a probability density leads to a probability measure by integration

$$P(A) = \int_A u(x)dx \; . \tag{2}$$

There are still other ways to specify probabilities of events in path space. All of these probability measures satisfy the same basic axioms.

Suppose that for each $A \in \mathcal{F}$ we have a number $P(A)$. The numbers $P(A)$ are a *probability measure* if

i. If $A \in \mathcal{F}$ and $B \in \mathcal{F}$ are disjoint events, then $P(A \cup B) = P(A) + P(B)$.

ii. $P(A) \geq 0$ for any event $A \in \mathcal{F}$.

iii. $P(\Omega) = 1$.

iv. If $A_n \in \mathcal{F}$ is a sequence of events each disjoint from all the others and $\cup_{n=1}^{\infty} A_n = A$, then $\sum_{n=1}^{\infty} P(A_n) = P(A)$.

The last property is called *countable additivity.* It is possible to consider probability measures that are not countably additive, but is not bery useful.

1.4. Example 1, discrete probability: If Ω is discrete, we may take $\mathcal{F}$ to be the set of all events (i.e. all subsets of Ω). If we know the probabilities of each individual outcome, then the formula (1) defines a probability measure. The axioms (i), (ii), and (iii) are clear. The last, countable additivity, can be verified given a solid undergraduate analysis course.

1.5. Borel sets: It is rare that one can define $P(A)$ for all $A \subseteq \Omega$. Usually, there are *non measurable* events whose probability one does not try to define (see below). This is not related to partial information, but is an intrinsic aspect of continuous probability. Events that are not measurable are quite artificial, but they are impossible to get rid of. In most applications in stochastic calculus, it is convenient to take the largest $\sigma-$algebra to be the *Borel sets*[2]

In a previous lecture we discussed how to generate a $\sigma-$algebra from a collection of sets. The Borel algebra is the $\sigma-$algebra that is generated by all *balls.* The *open ball* with center x_0 and radius $r > 0$ in n dimensional space is $B_r(x_0) = \{x \mid |x - x_0| < r$. A "ball" in one dimension is an interval. In two dimensions it is a disk. Note that the ball is solid, as opposed to the hollow *sphere*, $S_r(x_0) = \{x \mid |x - x_0| = r\}$. The condition $|x - x_0| \leq r$ instead of $|x - x_0| < r$, defines a *closed* ball. The $\sigma-$algebra generated by open balls is the same as the $\sigma-$algebra generated by closed balls (check this if you wish).

1.6. Borel sets in path space: The definition of Borel sets works the same way in the path space of Brownian motion, $C([0,T],R)$. Let $x_0(t)$ amd $x(t)$ be two continuous functions of t. The distance between them in the "sup norm" is

$$\|x - x_0\| = \sup_{0 \leq t \leq T} |x(t) - x_0(t)| \ .$$

We often use double bars to represent the distance between functions and single bar absolute value signs to represent the distance between numbers or vectors in R^n. As before, the open ball of radius r about a path x_0 is the set of all paths with $\|x - x_0\| < r$.

1.7. The $\sigma-$algebra for Markov chain path space: There is a convenient limit process that defines a useful $\sigma-$algebra on $\mathcal{S}^{\mathcal{N}}$, the infinite time horizon path space for a Markov chain. We have the algebras $\mathcal{F}_t$ generated by the first

[2] The larger $\sigma-$algebra of *Lebesgue sets* seems to more of a nuisance than a help, particularly in discussing convergence of probability measures in path space.

$t+1$ states $x(0)$, $x(1)$, $\ldots$, $x(t)$. We take $\mathcal{F}$ to be the $\sigma-$algebra generated by all these. Note that the event $A = \{X(t) \neq 1 \text{ for } t \geq 0\}$ is not in any of the $\mathcal{F}_t$. However, the event $A_t = \{X(t) \neq 1 \text{ for } 0 \leq t \leq T\}$ is in $\mathcal{F}_t$. Therefore $A = \cup_{t \geq 0} A_t$ must be in any $\sigma-$algebra that contains all the $\mathcal{F}_t$. Also note that the union of all the $\mathcal{F}_t$ is an algebra of sets, though it is not a $\sigma-$algebra.

1.8. Generating a probability measure: Let $\mathcal{M}$ be a collection of events that generates the $\sigma-$algebra $\mathcal{F}$. Let $\mathcal{A}$ be the algebra of sets that are finite intersections, unions, and complements of events in $\mathcal{M}$. Clearly the $\sigma-$algebra generated by $\mathcal{M}$ is the same as the one generated by $\mathcal{A}$. The process of going from the algebra $\mathcal{A}$ to the $\sigma-$algebra $\mathcal{F}$ is one of *completion*, adding all limits of countable intersections or unions of events in $\mathcal{A}$.

In order to specify $P(A)$ for all $A \in \mathcal{F}$, it suffices to give $P(A)$ for all events $A \in \mathcal{A}$. That is, *if there is* a countably additive probability measure $P(A)$ for all $A \in \mathcal{F}$, then it is completely determined by the numbers $P(A)$ for those $A \in \mathcal{A}$. Hopefully is is plausable that if the events in $\mathcal{A}$ generate those in $\mathcal{F}$, then the probabilities of events in $\mathcal{M}$ determine the probabilities of events in $\mathcal{F}$ (proof ommitted).

For example, in R^n if we specify $P(A)$ for event described by finitely many balls, then we have determined $P(A)$ for any Borel set. It might be that the numbers $P(A)$ for $A \in \mathcal{A}$ are inconsistent with the axioms of probability (which is easy to check) or can't be extended in a way that is countably additive to all of $\mathcal{F}$ (doesn't happen in our examples), but otherwise the measure is determined.

1.9. Non measurable sets (technical aside): A construction demonstrates that non measurable sets are unavoidable. Let Ω be the unit circle. The simplest probability measure on Ω would seem to be uniform measure (divided by 2π so that $P(\Omega) = 1$). This measure is *rotation invariant*: if A is a measurable event having probability $P(A)$ then the event $A + \theta = \{x + \theta \mid x \in A\}$ is measurable and has $P(A+\theta) = P(A)$. It is possible to construct a set B and a (countable) sequence of rotations, θ_n, so that the events $B + \theta_k$ and $B + \theta_n$ are disjoint if $k \neq n$ and $\bigcup_n B + \theta_n = \Omega$. This set cannot be measurable. If it were and $\mu = P(B)$ then there would be two choices: $\mu = 0$ or $\mu > 0$. In the former case we would have $P(\Omega) = \sum_n P(B + \theta_n) = \sum_n 0 = 0$, which is not what we want. In the latter case, again using countable additivity, we would get $P(\Omega) = \infty$.

The construction of the set B starts with a description of the θ_n. Write n in base ten, flip over the decimal point to get a number between 0 and 1, then multiply by 2π. For example for $n = 130$, we get $\theta_n = \theta_{130} = 2\pi \cdot .031$. Now use the θ_n to create an equivalence relation and partition of Ω by setting $x \sim y$ if $x = y + \theta_n \pmod{2\pi}$ for some n. The reader should check that this is an equivalence relation ($x \sim y \to y \sim x$, and $x \sim y$ and $y \sim z \to x \sim z$). Now, let B be a set that has exactly one representative from each of the equivalence classes in the partition. Any $x \in \Omega$ is in one of the equivalence classes, which means that there is a $y \in B$ (the representative of the x equivalence class) and an n so that $y + \theta_n = x$. That means that any $x \in \Omega$ has $x \in B + \theta_n$ for some n, which is to say that $\bigcup_n B + \theta_n = \Omega$. To see that $B + \theta_k$ is disjoint from

$B + \theta_n$ when $k \neq n$, suppose that $x \in B + \theta_k$ and $x \in \theta_n$. Then $x = y + \theta_k$ and $x = z + \theta_n$ for $y \in B$ and $z \in B$. But (and this is the punch line) this would mean $y \sim z$, which is impossible because B has only one representative from each equivalence class. The possibility of selecting a single element from each partition element without having to say how it is to be done is the *axiom of choice*.

1.10. Probability densities in R^n: Suppose $u(x)$ is a probability density in R^n. Ifg A is an event made from finitely many balls (or rectangles) by set operations, we can define $P(A)$ by integrating, as in (2). This leads to a probability measure on Borel sets corresponding to the density u. Deriving the probability measure from a probability density does not seem to work in path space because there is nothing like the Riemann integral to use in[3] (2) Therefore, we describe path space probability measures directly rather than through probability densities.

1.11. Measurable functions: Let Ω be a probability space with a $\sigma-$algebra $\mathcal{F}$. Let $f(\omega)$ be a function defined on Ω. In discrete probability, f was measurable with respect to $\mathcal{F}$ if the sets $B_a = \{\omega \mid f(\omega = a)\}$ all were measurable. In continuous probability, this definition is replaced by the condition that the sets $A_{ab} = \{\omega \mid a \leq f(\omega) \leq b\}$ are measurable. Because $\mathcal{F}$ is countably additive, and because the event $a < f$ is the (countable) union of the events $a + \frac{1}{n} \leq f$, this is the same as requiring all the sets $\widetilde{A}_{ab} = \{\omega \mid a < f(\omega) < b\}$ to be measurable. If Ω is discrete (finite or countable), then the two definitions of measurable function agree.

In continuous probability, the notion of measurability of a function with respect to a $\sigma-$algebra plays two roles. The first, which is purely technical, is that f is sufficiently "regular" (meaning not crazy) that abstract integrals (defined below) make sense for it. The second, particularly for smaller algebras $\mathcal{G} \subset \mathcal{F}$, again involves incomplete information. A function that is measurable with respect to $\mathcal{G}$ not only needs to be regular, but also must depend on fewer variables (possibly in some abstract sense).

1.12. Integration with respect to a measure: The definition of integration with respect to a general probability measure is easier than the definition of the Riemann integral. The integral is written

$$E[f] = \int_{\omega \in \Omega} f(\omega) dP(\omega) \ .$$

We will see that in R^n with a density u, this agrees with the classical definition

$$E[f] = \int_{R^n} f(x) u(x) dx \ ,$$

[3]The *Feynman integral* in path space has some properties of true integrals but lacks others. The probabilist Mark Kac (pronounced "cats") discovered that Feynman's ideas applied to the heat equation rather than the Schrödinger equation can be interpreted as integration with respect to Wiener measure. This is now called the *Feynman Kac formula*.

if we write $dP(x) = u(x)dx$. Note that the abstract variable ω is replaced by the concrete variable, x, in this more concrete situation. The general definition is forced on us once we make the natural requirements

i. If $A \in \mathcal{F}$ is any event, then $E[1_A] = P(A)$. The integral of the indicator function if an event is the probability of that event.

ii. If f_1 and f_2 have $f_1(\omega) \leq f_2(\omega)$ for all $\omega \in \Omega$, then $E[f_1] \leq E[f_2]$. "Integration is monotone".

iii. For any reasonable functions f_1 and f_2 (e.g. bounded), we have $E[af_1 + bf_2] = aE[f_1] + bE[f_2]$. (*Linearity* of integration).

iv. If $f_n(\omega)$ is an increasing family of positive functions converging *pointwise* to f ($f_n(\omega) \geq 0$ and $f_{n+1}(\omega) \geq f_n(\omega)$ for all n, and $f_n(\omega \to f(\omega)$ as $n \to \infty$ for all ω), then $E[f_n] \to E[f]$ as $n \to \infty$. (This form of countable additivity for abstract probability integrals is called the *monotone convergence theorem.*)

A function is a *simple function* if there are finitely many events A_k, and weights w_k, so that $f = \sum_k w_k 1_{A_k}$. Properties (i) and (iii) imply that the expectation of a simple function is

$$E[f] = \sum_k w_k P(A_k) \; .$$

We can approximate general functions by simple functions to determine their expectations.

Suppose f is a nonnegative bounded function: $0 \leq f(\omega) \leq M$ for all $\omega \in \Omega$. Choose a small number $\epsilon = 2^{-n}$ and define the[4] "ring sets" $A_k = \{(k-1)\epsilon \leq f < k\epsilon$. The A_k depend on ϵ but we do not indicate that. Although the events A_k might be complicated, fractal, or whatever, each of them is measurable. A simple function that approximates f is $f_n(\omega) = \sum_k (k-1)\epsilon 1_{A_k}$. This f_n takes the value $(k-1)\epsilon$ on the sets A_k. The sum defining f_n is finite because f is bounded, though the number of terms is M/ϵ. Also, $f_n(\omega) \leq f(\omega)$ for each $\omega \in \Omega$ (though by at most ϵ). Property (ii) implies that

$$E[f] \geq E[f_n] = \sum_k (k-1)\epsilon P(A_k) \; .$$

In the same way, we can consider the upper function $g_n = \sum_k k\epsilon 1_{A_k}$ and have

$$E[f] \leq E[g_n] = \sum_k k\epsilon P(A_k) \; .$$

The reader can check that $f_n \leq f_{n+1} \leq f \leq g_{n+1} \leq g_n$ and that $g_n - f_n \leq \epsilon$. Therefore, the numbers $E[f_n]$ form an increasing sequence while the $E[g_n]$ are a

[4]Take $f = f(x, y) = x^2 + y^2$ in the plane to see why we call them ring sets.

decreasing sequence converging to the same number, which is the only possible value of $E[f]$ consistent with (i), (ii), and (iii).

It is sometimes said that the difference between classical (Riemann) integration and abstract integration (here) is that the Riemann integral cuts the x axis into little pieces, while the abstarct integral cuts the y axis (which is what the simple function approximations amount to).

If the function f is positive but not bounded, it might happen that $E[f] = \infty$. The "cut off" functions, $f_M(\omega) = \min(f(\omega), M)$, might have $E[f_M] \to \infty$ as $M \to \infty$. If so, we say $E[f] = \infty$. Otherwise, property (iv) implies that $E[f] = \lim_{M\to\infty} E[f_M]$. If f is both positive and negative (for different ω), we integrate the positive part, $f_+(\omega) = \max(f(\omega), 0)$, and the negative part $f_-(\omega) = \min(f(\omega), 0$ separately and subtract the results. We do not attempt a definition if $E[f_+] = \infty$ and $E[f_-] = -\infty$. We omit the long process of showing that these definitions lead to an integral that actually has the properties (i) - (iv).

1.13. Markov chain probability measures on $\mathcal{S}^{\mathcal{N}}$: Let $\mathcal{A} = \cup_{\sqcup \geq \prime} \mathcal{F}_{\sqcup}$ as before. The probability of any $A \in \mathcal{A}$ is given by the probability of that event in $\mathcal{F}_t$ if $A \in \mathcal{F}_t$. Therefore $P(A)$ is given by a formula like (1) for any $A \in \mathcal{A}$. A theorem of Kolmogorov states that the *completion* of this measure to all of $\mathcal{F}$ makes sense and is countably additive.

1.14. Conditional expectation: We have a random variable $X(\omega)$ that is measurable with respect to the $\sigma-$algebra, $\mathcal{F}$. We have $\sigma-$algebra that is a sub algebra: $\mathcal{G} \subset \mathcal{F}$. We want to define the conditional expectation $Y = E[X \mid \mathcal{G}]$. In discrete probability this is done using the partition defined by $\mathcal{G}$. The partition is less useful because it probably is uncountable, and because each partition element, $B(\omega) = \cap A$ (the intersection being over all $A \in \mathcal{G}$ with $\omega \in A$), may have $P(B(\omega)) = 0$ (examples below). This means that we cannot apply Bayes' rule directly.

The definition is that $Y(\omega)$ is the random variable measurable with respect to $\mathcal{G}$ that best approximates X in the least squares sense

$$E[(Y - X)^2] = \min_{Z \in \mathcal{G}} E[(Z - X)^2] \ .$$

This is one of the definitions we gave before, the one that works for continuous and discrete probability. In the theory, it is possible to show that there is a minimizer and that it is unique.

1.15. Generating a $\sigma-$algebra: When the probability space, Ω, is finite, we can understand an algebra of sets by using the partition of Ω that generates the algebra. This is not possible for continuous probability spaces. Another way to specify an algebra for finite Ω was to give a function $X(\omega$, or a collection of functions $X_k(\omega)$ that are supposed to be measurable with respect to $\mathcal{F}$. We noted that any function measurable with respect to the algebra generated by functions X_k is actually a function of the X_k. That is, if $F \in \mathcal{F}$ (abuse of notation), then there is some function $u(x_1, \ldots, x_n)$ so that

$$F(\omega) = u(X_1(\omega), \ldots, X_n(\omega)) \ . \tag{3}$$

The intuition was that $\mathcal{F}$ contains the information you get by knowing the values of the functions X_k. Any function measurable with respect to this algebra is determined by knowing the values of these functions, which is precisely what (3) says. This approach using functions is often convenient in continuous probability.

If Ω is a continuous probability space, we may again specify functions X_k that we want to be measurable. Again, these functions generate an algebra, a $\sigma-$algebra, $\mathcal{F}$. If F is measurable with respect to this algebra then there is a (Borel measurable) function $u(x_1, \ldots)$ so that $F(\omega) = u(X_1, \ldots)$, as before. In fact, it is possible to define $\mathcal{F}$ in this way. Saying that $A \in \mathcal{F}$ is the same as saying that $\mathbf{1}_A$ is measurable with respect to $\mathcal{F}$. If $u(x_1, \ldots)$ is a Borel measurable function that takes values only 0 or 1, then the function F defined by (3) defines a function that also takes only 0 or 1. The event $A = \{\omega \mid F(\omega) = 1$ has (obviously) $F = \mathbf{1}_A$. The $\sigma-$algebra generated by the X_k is the set of events that may be defined in this way. A complete proof of this would take a few pages.

1.16. Example in two dimensions: Suppose Ω is the unit square in two dimensions: $(x, y) \in \Omega$ if $0 \leq x \leq 1$ and $0 \leq y \leq 1$. The "x coordinate function" is $X(x, y) = x$. The information in this is the value of the x coordinate, but not the y coordinate. An event measurable with respect to this $\mathcal{F}$ will be any event determined by the x coordinate alone. I call such sets "bar code" sets. You can see why by drawing some.

1.17. Marginal density and total probability: The abstract situation is that we have a probability space, Ω with generic outcome $\omega \in \Omega$. We have some functions $(X_1(\omega), \ldots, X_n(\omega)) = X(\omega)$. With Ω in the background, we can ask for the joint PDF of $(X_1, \ldots, X_n)$, written $u(x_1, \ldots, x_n)$. A formal definition of u would be that if $A \subseteq R^n$, then

$$P(X(\omega) \in A) = \int_{x \in A} u(x) dx \ . \tag{4}$$

Suppose we neglect the last variable, X_n, and consider the reduced vector $\tilde{X}(\omega) = (X_1, \ldots, X_{n-1})$ with probability density $\tilde{u}(x_1, \ldots, x_{n-1})$. This $\tilde{u}$ is the "marginal density" and is given by integrating u over the forgotten variable:

$$\tilde{u}(x_1, \ldots, x_{n_1}) = \int_{-\infty}^{\infty} u(x_1, \ldots, x_n) dx_n \ . \tag{5}$$

This is a continuous probability analogue of the law of total probability: integrate (or sum) over a complete set of possibilities, all values of x_n in this case.

We can prove (5) from (4) by considering a set $B \subseteq R^{n-1}$ and the corresponding set $A \subseteq R^n$ given by $A = B \times R$ (i.e. A is the set of all pairs $\tilde{x}, x_n$) with $\tilde{x} = (x_1, \ldots, x_{n-1}) \in B$). The definition of A from B is designed so that $P(X \in A) = P(\tilde{X} \in B)$. With this notation,

$$\begin{aligned}
P(\tilde{X} \in B) &= P(X \in A) \\
&= \int_A u(x)dx \\
&= \int_{\tilde{x} \in B} \int_{x_n=-\infty}^{\infty} u(\tilde{x}, x_n) dx_n d\tilde{x} \\
P(\tilde{X} \in B) &= \int_B \tilde{u}(\tilde{x}) d\tilde{x} \; .
\end{aligned}$$

This is exactly what it means for $\tilde{u}$ to be the PDF for $\tilde{X}$.

1.18. Classical conditional expectation: Again in the abstract setting $\omega \in \Omega$, suppose we have random variables $(X_1(\omega), \ldots, X_n(\omega))$. Now consider a function $f(x_1, \ldots, x_n)$, its expectated value $E[f(X)]$, and the conditional expectations

$$v(x_n) = E[f(X) \mid X_n = x_n] \; .$$

The Bayes' rule definition of $v(x_n)$ has some trouble because both the denominator, $P(X_n = x_n)$, and the numerator,

$$E[f(X) \cdot \mathbf{1}_{X_n = x_n}] \; ,$$

are zero.

The classical solution to this problem is to replace the exact condition $X_n = x_n$ with an approximate condition having positive (though small) probability: $x_n \leq X_n \leq x_n + \epsilon$. We use the approximaion

$$\int_{x_n}^{x_n+\epsilon} g(\tilde{x}, \xi_n) d\xi_n \approx \epsilon g(\tilde{x}, x_n) \; .$$

The error is roughly proportional to ϵ^2 and much smaller than either the terms above. With this approximation the numerator in Bayes' rule is

$$\begin{aligned}
E[f(X) \cdot \mathbf{1}_{x_n \leq X_n \leq x_n+\epsilon}] &= \int_{\tilde{x} \in R^{n-1}} \int_{\xi_n = x_n}^{\xi_n = x_n+\epsilon} f(\tilde{x}, \xi_n) u(\tilde{x}, x_n) d\xi_n d\tilde{x} \\
&\approx \epsilon \int_{\tilde{x}} f(\tilde{x}, x_n) u(\tilde{x}, x_n) d\tilde{x} \; .
\end{aligned}$$

Similarly, the denominator is

$$P(x_n \leq X_n \leq x_n + \epsilon) \approx \epsilon \int_{\tilde{x}} u(\tilde{x}, x_n) d\tilde{x} \; .$$

If we take the Bayes' rule quotient and let $\epsilon \to 0$, we get the classical formula

$$E[f(X) \mid X_n = x_n] = \frac{\int_{\tilde{x}} f(\tilde{x}, x_n) u(\tilde{x}, x_n) d\tilde{x}}{\int_{\tilde{x}} u(\tilde{x}, x_n) d\tilde{x}} \quad . \tag{6}$$

By taking f to be the characteristic function of an event (all possible events) we get a formula for the probability density of $\tilde{X}$ given that $X_n = x_n$, namely

$$\tilde{u}(\tilde{x} \mid X_n = x_n) = \frac{u(\tilde{x}, x_n)}{\int_{\tilde{x}} u(\tilde{x}, x_n) d\tilde{x}} \quad . \tag{7}$$

This is the classical formula for conditional probability density. The integral in the denominator insures that, for each x_n, $\tilde{u}$ is a probability density as a function of $\tilde{x}$, that is

$$\int \tilde{u}(\tilde{x} \mid X_n = x_n) d\tilde{x} = 1 \ ,$$

for any value of x_n. It is very useful to notice that as a function of $\tilde{x}$, u and $\tilde{u}$ almost the same. They differ only by a constant normalization. For example, this is why conditioning Gaussian's gives Gaussians.

1.19. Modern conditional expectation: The classical conditional expectation (6) and conditional probability (7) formulas are the same as what comes from the "modern" definition from paragraph 1.6. Suppose $X = (X_1, \ldots, X_n)$ has density $u(x)$, $\mathcal{F}$ is the $\sigma-$algebra of Borel sets, and $\mathcal{G}$ is the $\sigma-$algebra generated by X_n (which might be written $X_n(X)$, thinking of X as ω in the abstract notation). For any $f(x)$, we have $\tilde{f}(x_n) = E[f \mid \mathcal{G}]$. Since $\mathcal{G}$ is generated by X_n, the function $\tilde{f}$ being measurable with respect to $\mathcal{G}$ is the same as it's being a function of x_n. The modern definition of $\tilde{f}(x_n)$ is that it minimizes

$$\int_{R^n} \left(f(x) - \tilde{f}(x_n) \right)^2 u(x) dx \ , \tag{8}$$

over all functions that depend only on x_n (measurable in $\mathcal{G}$).

To see the formula (6) emerge, again write $x = (\tilde{x}, x_n)$, so that $f(x) = f(\tilde{x}, x_n)$, and $u(x) = u(\tilde{x}, x_n)$. The integral (8) is then

$$\int_{x_n=-\infty}^{\infty} \int_{\tilde{x} \in R^{n-1}} \left(f(\tilde{x}, x_n) - \tilde{f}(x_n) \right)^2 u(\tilde{x}, x_n) d\tilde{x} dx_n \ .$$

In the inner integral:

$$R(x_n) = \int_{\tilde{x} \in R^{n-1}} \left(f(\tilde{x}, x_n) - \tilde{f}(x_n) \right)^2 u(\tilde{x}, x_n) d\tilde{x} \ ,$$

$\tilde{f}(x_n)$ is just a constant. We find the value of $\tilde{f}(x_n)$ that minimizes $R(x_n)$ by minimizing the quantity

$$\int_{\tilde{x} \in R^{n-1}} \left(f(\tilde{x}, x_n) - g \right)^2 u(\tilde{x}, x_n) d\tilde{x} =$$

$$\int f(\tilde{x})^2 u(\tilde{x}, x_n) d\tilde{x} + 2g \int f(\tilde{x}) u(\tilde{x}, x_n) d\tilde{x} + g^2 \int u(\tilde{x}, x_n) d\tilde{x} \ .$$

The optimal g is given by the classical formula (6).

1.20. Modern conditional probability: We already saw that the modern approach to conditional probability for $\mathcal{G} \subset \mathcal{F}$ is through conditional expectation. In its most general form, for every (or almost every) $\omega \in \Omega$, there should be a probability measure P_ω on Ω so that the mapping $\omega \to P_\omega$ is measureable with respect to $\mathcal{G}$. The measurability condition probably means that for every event $A \in \mathcal{F}$ the function $p_A(\omega) = P_\omega(A)$ is a $\mathcal{G}$ measurable function of ω. In terms of these measures, the conditional expectation $\tilde{f} = E[f \mid \mathcal{G}]$ would be $\tilde{f}(\omega) = E_\omega[f]$. Here E_ω means the expected value using the probability measure P_ω. There are many such subscripted expectations coming.

A subtle point here is that the conditional probability measures are defined on the original probability space, Ω. This forces the measures to "live" on tiny (generally measure zero) subsets of Ω. For example, if $\Omega = R^n$ and $\mathcal{G}$ is generated by x_n, then the conditional expectation value $\tilde{f}(x_n)$ is an average of f (using density u) only over the hyperplane $X_n = x_n$. Thus, the conditional probability measures P_X depend only on x_n, leading us to write P_{x_n}. Since $\tilde{f}(x_n) = \int f(x) dP_{x_n}(x)$, and $\tilde{f}(x_n)$ depends only on values of $f(\tilde{x}, x_n)$ with the last coordinate fixed, the measure dP_{x_n} is some kind of δ measure on that hyperplane. This point of view is useful in many advanced problems, but we will not need it in this course (I sincerely hope).

1.21. Semimodern conditional probability: Here is an intermediate "semimodern" version of conditional probability density. We have $\Omega = R^n$, and $\tilde{\Omega} = R^{n-1}$ with elements $\tilde{x} = (x_1, \ldots, x_{n-1})$. For each x_n, there will be a (conditional) probability density function $\tilde{u}_{x_n}$. Saying that $\tilde{u}$ depends only on x_n is the same as saying that the function $x \to \tilde{u}_{x_n}$ is measurable with respect to $\mathcal{G}$. The conditional expectation formula (6) may be written

$$E[f \mid \mathcal{G}](x_n) = \int_{R^{n-1}} f(\tilde{x}, x_n) \tilde{u}_{x_n}(\tilde{x}) d\tilde{x} \ .$$

In other words, the classical $u(\tilde{x} \mid X_n = x_n)$ of (7) is the same as the semimodern $\tilde{u}_{x_n}(\tilde{x})$.

2 Gaussian Random Variables

The central limit theorem (CLT) makes Gaussian random variables important. A generalization of the CLT is Donsker's "invariance principle" that gives Brownian motion as a limit of random walk. In many ways Brownian motion is a multivariate Gaussian random variable. We review multivariate normal random variables and the corresponding linear algebra as a prelude to Brownian motion.

2.1. Gaussian random variables, scalar: The one dimensional "standard normal", or Gaussian, random variable is a scalar with probability density

$$u(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \; .$$

The normalization factor $\frac{1}{\sqrt{2\pi}}$ makes $\int_{-\infty}^{\infty} u(x)dx = 1$ (a famous fact). The mean value is $E[X] = 0$ (the integrand $xe^{-x^2/2}$ is antisymmetric about $x = 0$). The variance is (using integration by parts)

$$\begin{aligned}
E[X^2] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2} dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x \left(xe^{-x^2/2} \right) dx \\
&= -\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x \left(\frac{d}{dx} e^{-x^2/2} \right) dx \\
&= -\frac{1}{\sqrt{2\pi}} \left(xe^{-x^2/2} \right) \Big|_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} dx \\
&= 0 + 1
\end{aligned}$$

Similar calculations give $E[X^4] = 3$, $E[X^6] = 15$, and so on. I will often write Z for a standard normal random variable. A one dimensional Gaussian random variable with mean $E[X] = \mu$ and variance $\text{var}(X) = E[(X-\mu)^2] = \sigma^2$ has density

$$u(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \; .$$

It is often more convenient to think of Z as the random variable (like ω) and write $X = \mu + \sigma Z$. We write $X \sim \mathcal{N}(\mu, \sigma^2)$ to express the fact that X is normal (Gaussian) with mean μ and variance σ^2. The standard normal random variable is $Z \sim \mathcal{N}(0, 1)$

2.2. Multivariate normal random variables: The $n \times n$ matrix, H, is positive definite if $x^* H x > 0$ for any n component column vector $x \neq 0$. It is symmetric if $H^* = H$. A symmetric matrix is positive definite if and only if all its eigenvales are positive. Since the inverse of a symmetric matrix is symmetric, the inverse of a symmetric positive definite (SPD) matrix is also SPD. An n component random variable is a mean zero multivariate normal if it has a probability density of the form

$$u(x) = \frac{1}{z} e^{-\frac{1}{2} x^* H x} \; ,$$

for some SPD matrix, H. We can get mean $\mu = (\mu_1, \ldots, \mu_n)^*$ either by taking $X + \mu$ where X has mean zero, or by using the density with $x^* H x$ replaced by $(x-\mu)^* H (x-\mu)$.

If $X \in R^n$ is multivariate normal and if A is an $m \times n$ matrix with rank m, then $Y \in R^m$ given by $Y = AX$ is also multivariate normal. Both the cases $m = n$ (same number of X and Y variables) and $m < n$ occur.

2.3. Diagonalizing H: Suppose the eigenvalues and eigenvectors of H are $Hv_j = \lambda_j v_j$. We can express $x \in R^n$ as a linear combination of the v_j either in vector form, $x = \sum_{j=1}^n y_j v_j$, or in matrix form, $x = Vy$, where V is the $n \times n$ matrix whose columns are the v_j and $y = (y_1, \ldots, y_n)^*$. Since the eigenvectors of a symmetric matrix are orthogonal to each other, we may normalize them so that $v_j^* v_k = \delta_{jk}$, which is the same as saying that V is an orthogonal matrix, $V^*V = I$. In the y variables, the "quadratic form" x^*Hx is diagonal, as we can see using the vector or the matrix notation. With vectors, the trick is to use the two expressions $x = \sum_{j=1}^n y_j v_j$ and $x = \sum_{k=1}^n y_k v_k$, which are the same since j and k are just summation variables. Then we can write

$$\begin{aligned}
x^*Hx &= \left(\sum_{j=1}^n y_j v_j\right)^* H \left(\sum_{k=1}^n y_k v_k\right) \\
&= \sum_{jk} \left(v_j^* H v_k\right) y_j y_k \\
&= \sum_{jk} \lambda_k v_j^* v_k y_j y_k \\
x^*Hx &= \sum_k \lambda_k y_k^2 \,. \qquad (9)
\end{aligned}$$

The matrix version of the eigenvector/eigenvalue relations is $V^*HV = \Lambda$ (Λ being the diagonal matrix of eigenvalues). With this we have $x^*Hx = (Vy)^*HVy = y^*(V^*HV)y = y^*\Lambda y$. A diagonal matrix in the quadratic form is equivalent to having a sum involving only squares $\lambda_k y_k^2$. All the λ_k will be positive if H is positive definite. For future reference, also remember that $\det(H) = \prod_{k=1}^n \lambda_k$.

2.4. Calculations using the multivariate normal density: We use the y variables as new integration variables. The point is that if the quadratic form is diagonal the muntiple integral becomes a product of one dimensional gaussian integrals that we can do. For example,

$$\begin{aligned}
\int_{R^2} e^{-\frac{1}{2}(\lambda_1 y_1^2 + \lambda_2 y_2^2)} dy_1 dy_2 &= \int_{y_1=-\infty}^{\infty} \int_{y_2=-\infty}^{\infty} e^{-\frac{1}{2}(\lambda_1 y_1^2 + \lambda_2 y_2^2)} dy_1 dy_2 \\
&= \int_{y_1=-\infty}^{\infty} e^{-\lambda_1 y_1^2/2} dy_1 \cdot \int_{y_2=-\infty}^{\infty} e^{-\lambda_2 y_2^2/2} dy_2 \\
&= \sqrt{2\pi/\lambda_1} \cdot \sqrt{2\pi/\lambda_2} \,.
\end{aligned}$$

Ordinarily we would need a Jacobian determinant representing $\left|\frac{dx}{dy}\right|$, but here the determinant is $\det(V) = 1$, for an orthogonal matrix. With this we can find the normalization constant, z, by

$$\begin{aligned}
1 &= \int u(x) dx \\
&= \frac{1}{z} \int e^{-\frac{1}{2}x^*Hx} dx
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{z}\int e^{-\frac{1}{2}y^*\Lambda y}dy \\
&= \frac{1}{z}\int \exp(-\frac{1}{2}\sum_{k=1}^{n}\lambda_k y_k^2))dy \\
&= \frac{1}{z}\int \left(\prod_{k=1}^{n} e^{-\lambda_k y_k^2}\right) dy \\
&= \frac{1}{z}\prod_{k=1}^{n}\left(\int_{y_k=-\infty}^{\infty} e^{-\lambda_k y_k^2} dy_k\right) \\
&= \frac{1}{z}\prod_{k=1}^{n}\sqrt{2\pi/\lambda_k} \\
1 &= \frac{1}{z}\cdot\frac{(2\pi)^{n/2}}{\sqrt{\det(H)}} .
\end{aligned}
$$

This gives a formula for z, and the final formula for the multivariate normal density

$$
u(x) = \frac{\sqrt{\det H}}{(2\pi)^{n/2}} e^{-\frac{1}{2}x^* Hx} . \tag{10}
$$

2.5. The covariance, by direct integration: We can calculate the covariance matrix of the X_j. The jk element of $E[XX^*]$ is $E[X_jX_k] = \mathrm{cov}(X_j, X_k)$. The covariance matrix consisting of all these elements is $C = E[XX^*]$. Note the conflict of notation with the constant C above. A direct way to evaluate C is to use the density (10):

$$
\begin{aligned}
C &= \int_{R^n} xx^* u(x)dx \\
&= \frac{\sqrt{\det H}}{(2\pi)^{n/2}}\int_{R^n} xx^* e^{-\frac{1}{2}x^* Hx}dx .
\end{aligned}
$$

Note that the integrand is an $n \times n$ matrix. Although each particular xx^* has rank one, the average of all of them will be a nonsingular positive definite matrix, as we will see. To work the integral, we use the $x = Vy$ change of variables above. This gives

$$
C = \frac{\sqrt{\det H}}{(2\pi)^{n/2}}\int_{R^n} (Vy)(Vy)^* e^{-\frac{1}{2}y^*\Lambda y}dy .
$$

We use $(Vy)(Vy)^* = V(yy^*)V^*$ and take the constant matrices V outside the integral. This gives C as the product of three matrices, first V, then an integral involving yy^*, then V^*. So, to calculate C, we can calculate all the matrix elements

$$
B_{jk} = \frac{\sqrt{\det H}}{(2\pi)^{n/2}}\int_{R^n} y_j y_k^* e^{-\frac{1}{2}y^*\Lambda y}dy .
$$

Clearly, if $j \neq k$, $B_{jk} = 0$, because the integrand is an odd (antisymmetric) function, say, of y_j. The diagonal elements B_{kk} may be found using the fact that the integrand is a product:

$$B_{kk} = \frac{\sqrt{\det H}}{(2\pi)^{n/2}} \prod_{j \neq k} \left(\int_{y_j} e^{-\lambda_j y_j^2/2} dy_j \right) \cdot \int_{y_k} y_k^2 e^{-\lambda_k y_k^2/2} dy_k \ .$$

As before, λ_j factors (for $j \neq k$) integrate to $\sqrt{2\pi/\lambda_j}$. The λ_k factor integrates to $\sqrt{2\pi/(\lambda_k)^{3/2}}$. The λ_k factor differs from the others only by a factor $1/\lambda_k$. Most of these factors combine to cancel the normalization. All that is left is

$$B_{kk} = \frac{1}{\lambda_k} \ .$$

This shows that $B = \Lambda^{-1}$, so

$$C = V\Lambda^{-1}V^* \ .$$

Finally, since $H = V\Lambda V^*$, we see that

$$C = H^{-1} \ . \tag{11}$$

The covariance matrix is the inverse of the matrix defining the multivariate normal.

2.6. Linear functions of multivariate normals: A fundamental fact about multivariate normals is that a linear transformation of a multivariate normal is also multivariate normal, provided that the transformation is onto. Let A be an $m \times n$ matrix with $m \leq n$. This A defines a linear transformation $y = Ax$. The transformation is "onto" if, for every $y \in R^m$, there is at least ibe $x \in R^n$ with $Ax = y$. If $n = m$, the transformation is onto if and only if A is invertable ($\det(A) \neq 0$), and the only x is $A^{-1}y$. If $m < n$, A is onto if its m rows are linearly independent. In this case, the set of solutions is a "hyperplane" of dimension $n - m$. Either way, the fact is that if X is an n dimensional multivariate normal and $Y = AX$, then Y is an m dimensional multivariate normal. Given this, we can completely determine the probability density of Y by calculating its mean and covariance matrix. Writing μ_X and μ_Y for the means of X and Y respectively, we have

$$\mu_Y = E[Y] = E[AX] = AE[X] = A\mu_X \ .$$

Similarly, if $E[Y] = 0$, we have

$$C_Y = E[YY^*] = E[(AX)(AX)^*] = E[AXX^*A^*] = AE[XX^*]A^* = AC_XA^* \ .$$

The reader should verify that if C_X is $n \times n$, then this formula gives a C_Y that is $m \times m$. The reader should also be able to derive the formula for C_Y in terms of C_X without assuming that $\mu_Y = 0$. We will soon give the proof that linear functions of Gaussians are Gaussian.

2.7. Uncorrelation and independence: The inverse of a symmetric matrix is another symmertic matrix. Therefore, C_X is diagonal if and only if H is diagonal. If H is diagonal, the probability density function given by (10) is a product of densities for the components. We have already used that fact and will use it more below. For now, just note that C_X is diagonal if and only if the components of X are uncorrelated. Then C_X being diagonal implies that H is diagonal and the components of X are independent. The fact that uncorrelated components of a multivariate normal are actually independent firstly is a property only of Gaussians, and secondly has curious consequences. For example, suppose Z_1 and Z_2 are independent standard normals and $X_1 = Z_1 + Z_2$ and $X_2 = Z_1 - Z_2$, then X_1 and X_2, being uncorrelated, are independent of each other. This may seem surprising in view of that fact that increasing Z_1 by $1/2$ increases both X_1 and X_2 by the same $1/2$. If Z_1 and Z_2 were independent uniform random variables (PDF $= u(z) = 1$ if $0 \leq z \leq 1$, $u(z) = 0$ otherwise), then again X_1 and X_2 would again be uncorrelated, but this time not independent (for example, the only way to get $X_1 = 2$ is to have both $Z_1 = 1$ and $Z_2 = 1$, which implies that $X_2 = 0$.).

2.8. Application, generating correlated normals: There are simple techniques for generating (more or less) independent standard normal random variables. The Box Muller method being the most famous. Suppose we have a positive definite symmetric matrix, C_X, and we want to generate a multivariate normal with this covariance. One way to do this is to use the Choleski factorization $C_X = LL^*$, where L is an $n \times n$ lower triangular matrix. Now define $Z = (Z_1, \ldots, Z_n)$ where the Z_k are independent standard normals. This Z has covariance $C_Z = I$. Now define $X = LZ$. This X has covariance $C_X = LIL^* = LL^*$, as desired. Actually, we do not necessarily need the Choleski factorization; L does not have to be lower triangular. Another possibility is to use the "symmetric square root" of C_X. Let $C_X = V\Sigma V^*$, where Σ is the diagonal symmetric matrix with eigenvalues of C_X ($\Sigma = \Lambda^{-1}$ where Λ is given above), and V is the orthogonal matrix if eigenvectors. We can take $A = V\sqrt{\Sigma}V^*$, where $\sqrt{\Sigma}$ is the diagonal matrix. Usually the Choleski factorization is easier to get than the symmetric square root.

2.9. Central Limit Theorem: Let X be an n dimensional random variable with probability density $u(x)$. Let $X^{(1)}$, $X^{(2)}$, $\ldots$, be a sequence of independent samples of X, that is, independent random variables with the same density u. Statisticians call this iid (independent, identically distributed). If we need to talk about the individual components of $X^{(k)}$, we write $X_j^{(k)}$ for component j of $X^{(k)}$. For example, suppose we have a population of people. If we choose a person "at random" and record his or her height (X_1) and weight (X_2), we get a two dimensional random variable. If we measure 100 people, we get 100 samples,

$X^{(1)}$, ..., $X^{(100)}$, each consisting of a height and weight pair. The weight of person 27 is $X_2^{(27)}$. Let $\mu = E[X]$ be the mean and $C = E[(X-\mu)(X-\mu)^*]$ the covariance matrix. The Central Limit Theorem (CLT) states that for large n, the random variable

$$R^{(n)} = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} (X^{(k)} - \mu)$$

has a probability distribution close to the multivariate normal with mean zero and covariance C. One interesting consequence is that if X_1 and X_2 are uncorrelated then an average of many independent samples will have $R_1^(n)$ and $R_2^{(n)}$ nearly independent.

2.10. What the CLT says about Gaussians: The Central Limit Theorem tells us that if we avarage a large number of independent samples from the same distribution, the distribution of the average depends only on the mean and covariance of the starting distribution. It may be surprising that many of the properties that we deduced from the formula (10) may be found with almost no algebra simply knowing that the multivariate normal is the limit of averages. For example, we showed (or didn't show) that if X is multivariate normal and $Y = AX$ where the rows of A are linearly independent, then Y is multivariate normal. This is a consequence of the averaging property. If X is (approximately) the average of iid random variables U_k, then Y is the average of random variables $V_k = AU_k$. Applying the CLT to the averaging of the V_k shows taht Y is also multivariate normal.

Now suppose U is a univariate random variable with iid samples U_k, and $E[U_k] = 0$, $E[U_k^2 = \sigma^2]$, and $E[U_k^4] = a_4 < \infty$ Define $X_n = \frac{1}{\sqrt{n}} \sum_{k=n}^{n} U_k$. A calculation shows that $E[X_n^4] = 3\sigma^4 + \frac{1}{n} a_4$. For large n, the fourth moment of the average depends only on the second moment of the underlying distribution. A multivariate and slightly more general version of this calculation gives "Wick's theorem", an expression for the expected value of a product of components of a multivariate normal in terms of covariances.