Stochastic Calculus Notes, Lecture 4

Last modified October 5, 2002

1 Continuous probability

This section is a quick and sketchy introduction to the modern terminology of probability following Kolmogorov in what we call continuous spaces. Although the modern approach has lots of baggage, it ultimately makes things easier, as we will begin to see here.

1.1. Continuous spaces I use this to mean probability spaces that are not countable (discrete). In discrete probability, we first defined $P(\omega)$, the probability of any particular outcome. Then the probability of an event, A was the sum of the probabilities of the outcomes that make up that event:

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

In continuous probability, the rule (though there are exceptions), is that the probability of any particular outcome is zero. Also, there are uncountably many outcomes in a typical event. Both of these make (1) inapplicable. We do not know how to sum uncountable many numbers, and, we might expect such a sum rule to give the answer zero if all the terms in the sum were zero.

Examples of continuous probability spaces:

R, the real numbers. If ω is a real number and $u(x)$ is a probability density, then the probability of a small interval $(\omega - \epsilon, \omega + \epsilon)$ containing ω is (with an abuse of notation)

$$P(\omega - \epsilon, \omega + \epsilon) = \int_{\omega - \epsilon}^{\omega + \epsilon} u(x) dx \to 0 \;\text{ as } \epsilon \to 0.$$

Thus the probability of ω itself should naturally be zero.

R^n, sequences of n numbers (possibly viewed as a row or column vector depending on the context): $X = (x_1 \ldots, X_n)$.

$\mathcal{S}^{\mathcal{N}}$. Here $\mathcal{S}$ is the state space for a Markov chain (might be finite or countable) and $\mathcal{N}$ is the "natural" numbers, $1, 2, 3, \ldots$. An element is an infinite sequence of elements of $\mathcal{S}$: $X = (X_1, X_2, \ldots)$. Generally, the probability of any particular infinite sequence is zero. For example, if we have a two state Markov chain with transition matrix $\begin{pmatrix} .6 & .4 \\ .3 & .7 \end{pmatrix}$. If we call the states U and D, then the probability of the infinite string $UUU \cdots$ should be $u(U) \cdot .6 \cdot .6 \cdot \cdots = 0$: multiplying together infinitely many .6 numbers converges to zero.

$C([0,T] \to R)$, the path space for Brownian motion. The C stands for "continuous". The $[0,T]$ is the time interval $0 \le t \le T$; 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. If we call this function X_t for $0 \le t \le T$, X_t is a real number for each $t \in [0,T]$ and X is a continuous function of t.

1.2. Probability measures: We want to define the probabilities of events $A \subset \Omega$. Since we cannot base these on the probabilities of the individual outcomes in A, we just assume the probabilities are defined for events. For this we first define $\sigma-$algebra. An algebra of events is a $\sigma-$algebra if, for any sequence of events $A_n \in \Omega$, the union union $\cup_{n=1}^{\infty} A_n$ is also an event in $\mathcal{F}$. Suppose $\mathcal{F}$ is a $\sigma-$algebra of events in Ω. The numbers $P(A)$ for $A \in \mathcal{F}$ 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. $P(A) \ge 0$ 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". All the probability measures we deal with in this course are countably additive.

1.3. R^n: A "ball" in n dimensional space is any of the sets $B_r(x) = \{y \mid |x-y| < r$. This might be called an interval in one dimension and a disk in two, but the term ball applies to any dimension, including 1 and 2. With $|x-y| \le r$, we would have a "closed" ball, as opposed to the "open" ball above. This makes no difference here. In fact, a $\sigma-$algebra that contains all open balls also contains all closed balls, and any set in R^n you can describe without advanced mathematical analysis. The $\sigma-$algebra generated by open balls is called the Borel algebra, and events measurable in this algebra are called Borel sets. A function $u(x)$ is a probability density if it is never negative and $\int_{R^n} u(x)dx = 1$. Such a probability density defines a probability measure on the Borel algebra by

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

It is can be shown that if u is measurable with respect to the Borel sets then this probabiity measure is countable additive.

1.4. 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. Let Ω be a probability space, $\mathcal{F}$ a $\sigma-$algebra of events,

and P a probability measure. A function $f(\omega)$ is measureable with respect to $\mathcal{F}$ if all of the events $A_{ab} = \{a \leq f \leq b\} = \{\omega \mid a \leq f(\omega) \leq b\}$ are in $\mathcal{F}$. Because $\mathcal{F}$ is an algebra, the condition $a \leq f$ can be replaced by $a < f$, etc. Any function on R^n (i.e. any function of n real variables), no matter how many wierd discontinuities you try to throw in, will be measurable with respect to the Borel algebra, unless you know serious advanced analysis. It happens in general that a function may fail to be measurable with respect to some $\mathcal{F}$, but this will always (in this course) be due to a lack of information (small $\mathcal{F}$) rather than discontinuities in u.

The integral is written

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

In R^n with a density u, this agrees with teh classical definition

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

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 $f_1(\omega) \leq f_1(\omega)$ 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 $E[af_1 + bf_2] = aE[f_1] + bE[f_2]$. "Integration is linear".

Now suppose f is a nonnegative bounded function: $0 \leq f(\omega) \leq M$ for all $\omega \in \Omega$. The integral of f is determined by the three properties above. Choose a small number ϵ and define the "ring sets" $A_n = \{(n-1)\epsilon \leq f < n\epsilon$. The A_n depend on ϵ but we do not indicate that. Although the events A_n might be complicated, fractal, or whatever, Each of them is measurable. The "step function" $g(\omega) = \sum_n (n-1)\epsilon 1_{A_n}$ takes the value $(n-1)\epsilon$ on each of the sets A_n (each ω is in only one A_n. For any ω, only one of the terms in the sum is different from zero.). The sum defining g is finite because f is bounded, though the number of terms is M/ϵ. Also, $g(\omega) \leq f(\omega)$ for each $\omega \in \Omega$ (though by at most ϵ). Therefore, the three properties of integration imply that

$$E[f] \geq E[g] = \sum_n (n-1)\epsilon E[A_n] = \sum_n (n-1)\epsilon P((n-1)\epsilon \leq f < n\epsilon) \ .$$

In the same way, we can consider the upper function $h = \sum_n n\epsilon 1_{A-n}$ and have

$$E[f] \leq E[h] = \sum_n n\epsilon E[A_n] = \sum_n n\epsilon P((n-1)\epsilon \leq f < n\epsilon) \ .$$

If you draw a picture of this situation for $\Omega = R$, you will see the lower (g) and upper (h) step functions bracketing f. When you replace ϵ by $\epsilon/2$, the lower step goes up and the upper step goes down. This gives a sequence of approximations $G(\epsilon) \leq E[f] \leq H(\epsilon)$ with $G(\epsilon)$ increasing and $H(\epsilon)$ decreasing as $\epsilon \to 0$. Finally, note that $H(\epsilon) - G(\epsilon) \leq \epsilon$, because that is how close the upper and lower step approximations h and g are. Thus, as $\epsilon \to 0$, the upper and lower approximations converge to the same number, which must be $E[f]$. It is sometimes said that the difference between classical (Riemann) integration and modern integration (here) is that we used to cut the x axis into little pieces, but it is simpler to cut the y axis instead.

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 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 $E[f_-] = -\infty$.

1.5. Markov chains with $T = \infty$: The probability space, Ω, is the set of all infinite sequences $X = (X_1, X_2, \ldots)$, where each X_t is one of the states in the state space $\mathcal{S}$. Just as the Borel algebra of sets can be generated by balls, the algebra of sets here can be generated by "cylinder" sets (don't ask me how they got that name). For each sequence of length L, $x = (x_1, \ldots, x_l)$, there is a cylinder set $B_x = \{X \mid X_1 = x_1, \ldots, X_L = x_L\}$. Other sets can be made from countable set operations starting with these. For example, the event containing the single sequence $UUU\cdots$ is the intersection of the events having the first L entries U. In a slightly more complicated way, it is possible to express the event "the first $UUDDU$ occurs before the first $DDUD$" in terms of cylinder sets. The probabilities $P(B_x) = u_1(X_1) \prod_{t=1}^{L-1} P_{x_t, x_{t+1}}$ give rise to a probability measure that is countably additive on this $\sigma-$algebra, another theorem of Kolmogorov.

1.6. Conditional expectation: We have a random variable $X(\omega)$ that is measurable with respect to the $\sigma-$algebra, $\mathcal{F}$, and a subalgebra $\mathcal{G} \subset \mathcal{F}$. We want to define the conditional expectation $Y = E[X \mid \mathcal{G}]$. When Ω is finite we can define $Y(\omega)$ be knowing which partition block ω is in. In continuous probability, a subalgebra might or might not be generated by a partition (I don't know), but even if it were, the sets in the partition would usually have probability zero so Bayes' rule would not be applicable. For example, suppose we have a two dimensional random variable $X = (X_1, X_2)$ with a density $u(x_1, x_2)$ and we want $P(X_1 > 3 \mid X_2 = 0)$. The event $B = \{X_2 = 0\}$ has probability $P(B) = 0$. There is a "classical" definition of conditional expectation for this case (see homewrok 1), but the one "modern" definition works for all cases. 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.

1.7. 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

$$F(\omega) = u(X_1(\omega), \ldots, X_n(\omega)) \, . \tag{2}$$

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 (2) 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 (2) 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.

1.8. 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.

1.9. 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

$$P(X(\omega) \in A) = \int_{x \in A} u(x) dx \, . \tag{3}$$

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:

$$\tilde{u}(x_1, \ldots, x_{n_1}) = \int_{-\infty}^{\infty} u(x_1, \ldots, x_n) dx_n \ . \tag{4}$$

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 (4) from (3) 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 $\tilde{X}$.

1.10. 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

$$E[f(X) \mid X_n = x_n] = \frac{\int_{\tilde{x}} f(\tilde{x}, x_n) u(\tilde{x}, x_n) d\tilde{x}}{\int_{\tilde{x}} u(\tilde{x}, x_n) d\tilde{x}} \ . \tag{5}$$

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

$$\tilde{u}(\tilde{x} \mid X_n = x_n) = \frac{u(\tilde{x}, x_n)}{\int_{\tilde{x}} u(\tilde{x}, x_n) d\tilde{x}} \ . \tag{6}$$

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.

1.11. Modern conditional expectation: The classical conditional expectation (5) and conditional probability (6) 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

$$\int_{R^n} \Big(f(x) - \tilde{f}(x_n) \Big)^2 u(x) dx \ , \tag{7}$$

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

To see the formula (5) 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 (7) is then

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

In the inner integral:

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

$\tilde{f}(x_n)$ 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 (5).

1.12. 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).

1.13. 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 (5) 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 (6) 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 Brow-

nian 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 $\int_{-\infty}^{\infty} u(x)dx = 0$ (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)

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

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^*Hx 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. Diagonaiizing 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

$$\begin{aligned} x^*Hx &= \left(\sum_{j=1}^n y_j v_j\right)^* H \left(\sum_{k=1}^n y_k v_k\right) \\ &= \sum_{jk} \left(v_j^* H v_k\right) y_j y_k \\ &= \sum_{jk} \lambda_k v_j^* v_k y_j y_k \\ x^*Hx &= \sum_k \lambda_k y_k^2 \,. \end{aligned} \tag{8}$$

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

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

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

$$u(x) = \frac{\sqrt{\det H}}{(2\pi)^{n/2}}e^{-\frac{1}{2}x^*Hx}\ . \tag{9}$$

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 (9):

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

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

$$C = H^{-1} \ . \tag{10}$$

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 (9) 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 $X^{(1)}$, $X^{(2)}$, ..., 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 (9) 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.