Stochastic Analysis

Tyler Chang
May 7, 2019

The Standard Model of Probability

To model a stochastic phenomenon, it is typical to choose or fit a generic probability model, then evaluate its correctness. The *standard model of probability* consists of a *probability space* $(S, \Sigma, \mathbb{P})$, which is defined for a stochastic phenomenon.

- S is called the *state space* and is a set consisting of all possible outcomes of the phenomenon;
- Σ is a σ-algebra on S, and consists of all possible events, where an *event* $E \subset 2^S$ is a set of outcomes for which we can know the probability (i.e., a measurable set in Σ);
- $\mathbb{P}$ is a *probability measure*, i.e., a map $\mathbb{P} : \Sigma \to [0, 1]$ such that $\mathbb{P}(S) = 1$, $\mathbb{P}(\emptyset) = 0$, and if $E \cap F = \emptyset$ then $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F)$.

This model is convenient since the properties of a σ-algebra and probability measure agree with the intuitive properties of probability.

If the state space is $\mathbb{R}$, then the standard choice for Σ is the *Borel algebra* and the standard choice for $\mathbb{P}$ is $\mathbb{P}(E) = \int_E f(x)dx$, where f is Lebesgue measurable and dx is taken w.r.t. the Lebesgue measure. This function f essentially determines the measure $\mathbb{P}$ and is called the *probability density function* (pdf). In the special case where $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$, f is called a *Gaussian distribution* with mean μ and standard deviation σ. Another common case is that $S = \{0, 1\}$, $\Sigma = 2^S$, and $\mathbb{P}(0) = p \in [0, 1]$ and $\mathbb{P}(1) = 1 - p$ (Bernoulli).

To join two independent probability models $(S_1, \Sigma_1, \mathbb{P}_1)$ and $(S_2, \Sigma_2, \mathbb{P}_2)$, it suffices to take: $S^{(2)} = S_1 \times S_2$, $\Sigma^{(2)}$ is the smallest (i.e., intersection over all) σ-algebras containing $\Sigma_1 \times \Sigma_2$, and $\mathbb{P}^{(2)}(E_1 \times E_2) = \mathbb{P}(E_1)\mathbb{P}(E_2)$. Let $\{(S_n, \Sigma_n, \mathbb{P}_n)\}_{n=1}^{\infty}$ be an infinite sequence of independent probability spaces. Then we can use this process to iteratively construct a sequence of probability spaces $(S^{(n)}, \Sigma^{(n)}, \mathbb{P}^{(n)})$ that combines the first n spaces in the sequence. The elements of $S^{(n)}$ are called the *canonical coordinate projections* of the infinite sequence, and the elments of $\Sigma^{(n)}$ are called *cylinder sets*.

The **Kolmogorov Extension Theorem** guarantees that there exists a space $(S, \Sigma, \mathbb{P})$ such that $\Sigma^{(n)} \subseteq \Sigma$ for all n, and $\mathbb{P}(E) = \mathbb{P}_n(E)$ for $E \in \Sigma^{(n)}$. However, there is no valid $\mathbb{P}$ for the measurable space $(S, 2^S)$. Generally, we choose Σ to contain $E \times S_{n+1} \times \dots$ for every $E \in \Sigma^{(n)}$ for all n. I.e., every finite sequence of information is measurable, but infinite sequences of information are not measurable.

Random variables

A *random variable* y is a function $y : S \to \mathbb{R}$ that is measurable with respect to the probability space $(S, \Sigma, \mathbb{P})$. A random vairable is said to be *discrete* if Ran $y = \{y_1, \ldots, y_n\}$, and *continuous* if $\mathbb{P}(\{x \in S : y(x) \in E\}) = \int_E f(x)dx$. Here, y is said to be *distributed by* f (written $y \sim f$).

- In the discrete case, if Ran $y = \{0, 1\}$ and $\mathbb{P}(y = 0) = p$ then $\mathbb{P}(y = 1) = 1 - p$, and y is said to be a *Bernoulli random variable*.
- In the continuous case, if $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$, then y is said to be a *Gaussian random variable* with mean μ and standard deviation σ.

The *expected value* of y is a linear functional given by

$$\mathbb{E}[y] = \sum_{i=1}^{n} y_i p_i \qquad \text{or} \qquad \mathbb{E}[y] = \int_S x f(x)dx$$

depending on whether y is discrete or continuous. The expected value of a function $g(y)$ is similarly given by

$$\mathbb{E}[g(y)] = \sum_{i=1}^{n} g(y_i) p_i \qquad \text{or} \qquad \mathbb{E}[y] = \int_S g(x) f(x)dx$$

Two special cases are

- The *characteristic function* of a continuous random variable y is given by $\phi(\lambda) = \mathbb{E}\left[e^{i\lambda y}\right]$. I.e., this is the *Fourier transform* of the distribution function f.
- The *variance* of a random variable is given by $\sigma^2 = \mathbb{E}\left[(y - \mathbb{E}[y])^2\right]$.

For two random variable $y_1 \sim f_1$ and $y_2 \sim f_2$, their *joint probability distribution function* $f(y_1, y_2)$ is such that $\mathbb{P}(y_1 \in E_1, y_2 \in E_2) = \int_{E_1} \int_{E_2} f(y_1, y_2)dy_1 dy_2$. Of course, the above formulation can be unwieldy to work with. However, we say that y_1 and y_2 are *independent* if $\mathbb{P}(y_1 \in E_1, y_2 \in E_2) = \mathbb{P}(y_1 \in E_1)\mathbb{P}(y_2 \in E_2)$.

Gaussian random variables have nice properties that make them easy to work with. For n random variables $y_1, \ldots, y_n$, the *covariance matrix* is the $n \times n$ matrix K with $K_{ij} = \mathbb{E}\left[(y_i - \mu_i)(y_j - \mu_j)\right]$. Two Gaussian random variables y_i and y_j are independent if and only their covariance $K_{ij} = 0$. Furthermore, if A is an $m \times n$ matrix, and u is a vector of Gaussian random variables, then $v = Au$ is also a vector of random variables.

Recall, the *conditional probability* of an event E given F is a number $\mathbb{P}(E|F) = \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(F)}$, or by **Baye's Theorem**, $\mathbb{P}(E|F) = \mathbb{P}(F|E)\frac{\mathbb{P}(E)}{\mathbb{P}(F)}$. However, the *conditional expected value* of y given $\Phi \subset \Sigma$ is another random variable (i.e., a function) $\mathbb{E}[y|\Phi] = w$ where w is a random variable on Φ such that for all measurable z on Φ, $\mathbb{E}[zw] = \mathbb{E}[zy]$. I.e., $\mathbb{E}\left[z\mathbb{E}\left[y|\Phi\right]\right] = \mathbb{E}[zy]$. To evaluate a conditional expectation, an event must be applied. If $y, w \sim f(y, w)$, then $\mathbb{E}[y|w = \omega] = \int_S y f(y, \omega)dy$ (where $\{w = \omega\} = \{x : w(x) = \omega\} \in \Phi$).

Stochastic Processes

A *stochastic process* is defined by any sequence of random variables y_t $(t = 1, \ldots, \infty)$, which act on the same probability space $(S, \Sigma, \mathbb{P})$. For a given stochastic process, the *sample path* is a particular sequence $y_t(E_t)$, where $E_t \in S$.

- A process y_t is said to be *independent identically distributed* (i.i.d.) if $y_i = y_j$ for all i, j. I.e., each y_t models an independent draw from the same distribution;
- A process y_t is said to be a *Martingale process* if $\mathbb{E}(y_{n+1}|y_n, \ldots, y_1) = y_n$;
- A process B_t is a *Wiener process* or *Brownian motion* if

  (i) $B_0 = 0$ almost surely,

  (ii) $B_{t+k} - B_t$ is independent of B_s for all $s \leq t$,

  (iii) $B_{t+k} - B_t$ is a Gaussian with mean 0 and variance k, and

  (iv) B_t is almost surely continuous in t;

  (Note that (iii) implies that Brownian motion is a Martingale.)
- A *Markov process* if $\mathbb{P}(y_{n+1} = \ell_j | y_n = \ell_{i_n}, \ldots, y_1 = \ell_{i_1}) = \mathbb{P}(y_{n+1} = \ell_j | y_n = \ell_{i_n})$ for all ℓ_j. I.e., given just the current state $y_n = \ell_{i_n}$, a Markov process can be perfectly propogated either forward or backward in time.

  A Markov process is best modeled by a *stochastic matrix* A, where $A_{ij} = p_{ij} = \mathbb{P}(y_{n+1} = \ell_j | y_n = \ell_i)$, i.e., the probability of y_t transitioning from $y_t = \ell_i$ to $y_{t+1} = \ell_j$ in a time step $t \to t + 1$. Each value ℓ_i in the range of y_t is called a *site*, and is considered as a node in the transition graph for y_t. Then if the row vector $\pi^{(t)}$ denotes the distribution of probabilities for y_t (i.e., $\pi_i^{(t)} = \mathbb{P}(y_t = \ell_i)$), then $\pi^{(t+1)} = \pi^{(t)} A$ and $\mathbb{E}[f(y)] = \pi^{(t)} A f$, where f is a column vector with $f_i = f(\ell_i)$.

  A Markov process is called *irreducible* if there is a nonzero probability of eventually transitioning from a site $y_t = \ell_i$ to $y_n = \ell_j$ for all i and j, and $n > t$. I.e., there exists $n > t$ such that $\mathbb{P}(y_n = \ell_i | y_t = \ell_j) > 0$ for all i, j. The *period* of a site ℓ_i is gcd $(n - t)$ (where $n > t$) such that $y_n \ell_i$, given $y_t = \ell_i$. If $\mathbb{P}(y = \ell_i | y = \ell_i) \neq 0$, then the period of ℓ_i is trivially one (and we say that ℓ_i is *aperiodic*). If a process is irreducible then all its sites have the same period, which is referred to as the period of the process.

There are several types of convergence for a stochastic process y_n.

- $y_n \to y$ *in distribution* if for all continuous functions f, $\mathbb{E}(f(y_n)) \to \mathbb{E}(f(y))$.
- $y_n \to y$ *in probability* if for all $\varepsilon > 0$, $\mathbb{P}(\{x \in S : |y_n(x) - y(x)| > \varepsilon\}) \to 0$;
- $y_n \to y$ *almost surely* if $\mathbb{P}(\{x \in S : y_n(x) \to y(x)\}) = 1$;

Note that convergence almost surely implies convergence in probability implies convergence in distribution. Convergence in distribution is most common since it does not require the definition of a probability measure $\mathbb{P}$.

The **Weak Law of Large Numbers** states that for an i.i.d. sequence y_n, if $\mathbb{E}[y_n] < \infty$ then $y_1 + \ldots + y_n \to \mathbb{E}[y_i]$ (for any i) in probability. The **Strong Law of Large Numbers** repeats the same statement, but the convergence is almost surely. Note that the strong law implies the weak law, making it redundant. More generally, a stochastic process S_n is said to have the *strong law property* if the strong law holds, and the *weak law property* if only the weak law holds for that process. The **Central Limit Theorem** states that if an i.i.d. sequence y_n satisfies $\mathbb{E}[y_n] = 0$ and $\mathbb{E}[y_n^2] = \sigma^2 < \infty$, then $\frac{y_1+\ldots+y_n}{\sqrt{n}} \to y$ in distribution, where y is a Gaussian random variable with mean 0 and variance σ^2.

Now, consider a Markov process, y_t defined by a stochastic matrix A, with distribution $\pi^{(t)}$. It is guaranteed that y_t has at least one stationary distribution π^* such that $\pi^* A = \pi^*$. A cannot have eigenvalues with modulus greater than one, so if π^* is the only left eigenvector for $\lambda = 1$, and no other eigenvalues have modulus one, then $\pi^{(t)} \to \pi^*$ almost surely, since $\pi^{(t+1)} = \pi^{(t)} A$. The **Perron-Frobenius Theorem** states that if y_t is irreducible, then it has a unique stationary distribution satisfying $\pi_i^* \geq 0$ and $\sum_i \pi_i^* = 1$ for all i. Furthermore, if y_t is aperiodic, then all other eigenvalues have modulus strictly less than one. Therefore, for all functions f, $\frac{1}{n}\sum_{j=1}^{\infty} f(y_t) \to \sum_{j=1}^{N} f(j)\pi_j$ almost surely. Otherwise, if y_t has period n, the modulo one eigenvalues of A must come from the n roots of unity.

Estimation Theory

Given a model of a stochastic process $M(X, \rho)$ with hyperparameters ρ, the goal of estimation theory is to estimate $\hat{\rho} \approx \rho$ based on labeled data X^*. This estimation is said to be *consistent* if $\hat{\rho} \to \rho$ as $|X^*|$ grows, and is said to be *unbiased* if $\mathbb{E}[\hat{\rho}] = \mathbb{E}[\rho]$ for a fixed size of $|X^*|$. The *Fisher information* $F(\rho)$ tells how well the data X^* can be used to predict $\hat{\rho}$. For an unbiased estimator, $\mathbb{E}[(\rho - \hat{\rho})^2] \geq \frac{1}{F(\rho)}$.

- The *maximum likelihood estimate* (MLE) is given by the parameter choices that make the observed data X^* most probable.
- The *Bayesian estimate* is given by the most likely parameter choices, given the observed data X^* and some apriori assumptions about the values of ρ.

Itô's Calculus

Itô's calculus allows us to integrate and differentiate functions of Brownian motion (random variables) B_t. This allows us to solve stochastic ODEs. *Itô's integral* is defined similarly as the forward Riemann integral:

$$\int_0^t g(s, B_s) dB_s = \lim_{N \to \infty} \sum_{j=0}^{N-1} g(t_j, B_{t_j})(B_{t_{j+1}} - B_{t_j})$$

where $t_j = \frac{t}{N}j$. Note that ΔB_t is approximated by $B_{t_{j+1}} - B_{t_j}$, which points to the future, and estimates the integral based on the left endpoints. Using a different point to estimate g on $[t_j, t_{j+1}]$ will yield a different integration rule, and unlike Riemann, the different rules do NOT necessarilly converge to the same value. Itô's integral is a linear operator. Also, for $x_t = \int_0^t g(s, B_s)dB_s$, $\mathbb{E}[x_t] = 0$ and $\mathbb{E}[x_t^2] = \int_0^t \mathbb{E}[g^2]ds$. Some basic rules for Itô's integral include:

(i) $\int_0^t dB_s = B_t$

(ii) $\int_0^t B_s dB_s = \frac{1}{2}(B_t - t)^2$

(iii) $(dB_t)^2 = dt$

Itô's lemma defines the stochastic differentials, and allows for us to solve stochastic ODEs: Let u be a continuously differentiable function, and let $x_t - x_0 = \int_0^t g(s, B_s)dB_s + \int_0^t f(s, B_s)ds$. Then

$$u(x_t) - u(x_0) = \int_0^t u'(x_s)g(s, B_s)dB_s + \int_0^t \frac{1}{2}u''(x_s)g(s, B_s)^2(dB_s)^2 + \int_0^t u'(x_s)f(s, B_s)ds.$$

The second term above is often referred to as *Itô's drift*. Written in differential form, we have

$$du(x_t) = u'(x_t)g(t, B_t)dB_t + \frac{1}{2}u''(x_t)g(t, B_t)^2(dB_t)^2 + u'(x_t)f(t, B_t)dt.$$

Using (iii) above, this can be simplified to

$$du(x_t) = u'(x_t)g(t, B_t)dB_t + \left(\frac{1}{2}u''(x_t)g(t, B_t)^2 + u'(x_t)f(t, B_t)\right)dt.$$

One common application of Itô's lemma is the *Black-Scholes* model for pricing European options. Assume an *arbitrage free* market, and a *self-financing* portfolio strategy. Then if the stock price at time t is given by S_t, the risk-free interest rate is r, the volatility is σ, the maturity time is T, and the strike price is K, we have the following stochastic PDE

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} = rV - rS\frac{\partial V}{\partial S},$$

where $V(S_t, t)$ is the fair price for the option at time t. Solving with Itô's calculus and assuming the boundary condition $V(S, T) = (S - K)_+$, we get the Black-Scholes formula

$$V(S_t, t) = N(d_1)S_t - N(d_2)Ke^{-r(T-t)}$$

where $N(x)$ denotes the normal distribution with mean 0 and standard deviation x, $d_1 = \frac{1}{\sigma\sqrt{T-t}}\left(\log\frac{S_t}{K} + (r + \frac{\sigma^2}{2})(T - t)\right)$, and $d_2 = d_1 - \sigma\sqrt{T-t}$.