0.1 Law of Total Expectation

Suppose X and Y are random varialbes over the same probability space (i.e., the events or outcomes happens simultaneously, e.g., gender=girl, boy, and haircolor=black, brown, other happens at the same time), then the Law is

$$\mathrm{E}(X) = \mathrm{E}_Y(\mathrm{E}_{X|Y}(X|Y)) \tag{1}$$

0.1.1 Proof in the discrete case

$$\begin{aligned} \mathrm{E}(X) &= \mathrm{E}_Y(\sum_x xP(X=x|Y)) \qquad \text{weighted sum of the random outcomes of x} \\ &= \sum_y(\sum_x xP(X=x|Y=y))P(Y=y) \\ &= \sum_x x(\sum_y P(X=x|Y=y)P(Y=y)) \\ &= \sum_x xP(X=x) \end{aligned} \tag{2}$$

0.2 Law of Total Variance (conditional variance formula)

$$\mathrm{Var}(Y) = \mathrm{E}(\mathrm{Var}(Y|X)) + \mathrm{Var}(\mathrm{E}(Y|X)) \tag{3}$$

0.2.1 Proof

$$\mathrm{Var}(Y) = \mathrm{E}(Y^2) - \mathrm{E}^2(Y) \tag{4}$$

0.3 Kernel Density Estimation

Draw n samples randomly from some unknown distribution f, the samples $(x_1, \cdots, x_n)$ are thus iid. The shape of f is estimated by using the n samples

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{hn}\sum_{i=1}^{n} K(\frac{x - x_i}{h}), \tag{5}$$

where K is a nonnegative kernel that integrates to one and has mean zero, $K_h(x) = \frac{1}{h}K(\frac{x}{h})$, and h is a positive smoothing parameter called the bandwidth. Notice that $\hat{f}_h(x)$ becomes a deterministic function once the n samples are determined, but changes as new n samples are drawn. If K is a standard normal kernel with $z = (x - x_i)/h \sim \mathcal{N}(0, 1)$, x is treated as a random variable with mean x_i, and standard deviation h.

Finding the best bandwidth

The mean integrated squared error (mean of the integrated squared error over multiple batches of n samples, where each batch gives an estimation of $\hat{f}_h(x)$)

$$\text{MISE}(h) = \text{E} \int (\hat{f}_h(x) - f(x))^2 dx. \tag{6}$$

0.4 Subgrid-scale Parametrization with CMC (Conditional Markov Chain)

The Markov chain is a stochastic process/sequence indexed by t with the Markov property. The multi-dimensional state X has N_x outcomes indexed by i and j as $X^i(t)$ and $X^j(t + dt)$, with a corresponding N_b outcomes subgrid-scale parameter $B^n(t)$ and $B^m(t + dt)$. The i, j, n and m are the indices for the uniformly separated intervals of the domain of X and B. Suppose $i = \{1, \cdots, N_x\}$, and $n = \{1, \cdots, N_B\}$ At fixed i and j, the stochastic transition matrix of size $N_B \times N_B$ is

$$\mathbf{P}^{ij}_{nm} = P(B^m_j \mid B^n_i, X^i, X^j), \tag{7}$$

therefore there are N_x^2 numbers of matrices. Given the initial stochastic row vector of size N_x with elements summed to one, we can multiply the transition matrix to get the next stochastic vector. The ith index of the vector indicates the probability of ending up at the ith state.

Example:

$$[0.3 0.7] \begin{pmatrix} 0.2 & 0.8 \\ 0.3 & 0.7 \end{pmatrix} \tag{8}$$

The probability of ending up at the 1st state equals $0.3 \times 0.2 + 0.7 \times 0.3$, *which is the transition probability of 1 to 1, and 2 to 1 added.*

Instead of using the transition matrix to calculate the final probability vector, we can use it to sample the states to obtain realizations of stochastic processes.

Example:

Suppose one given the triplet (i, n, j), the mth state of B is sampled according to the nth row stochastic vector of the transition matrix $\mathbf{P}^{ij}$. If m is sampled at $m = 2$, then the next triplet $(j, 2, k)$, where $X^k(t+dt)$ is determined from the dynamic system, will sample at the 2nd row of the transition matrix $\mathbf{P}^{jk}$.

0.5 Hidden Markov Model