Stochastic Calculus, Courant Institute, Fall 2022
http://www.math.nyu.edu/faculty/goodman/teaching/StochCalc2022/index.html

Assignment 4

Correction. Exercise 2 of part 2 has been corrected to say q is $\mathcal{N}(s,1)$. Amazingly, this is a small change from the garble in the original version.

Part 1, Maximum likelihood.

This exercise illustrates a use of the fact that AB = BA if the matrices are compatible for multiplication both ways. This is the calculation showing that the empirical covariance matrix is the maximum likelihood estimate of the true covariance matrix. Suppose $p(x,\theta)$ is a probability density function as a function of x with parameters θ (both x and θ may have more than one component). Suppose X_k for $k = 1, \cdots, N$ are independent samples $X_k \sim p(\cdot, \theta_*)$. The *maximum likelihood estimator* is a way to use the samples to estimate the unknown parameters, which are the components of θ. The method is to find $\widehat{\theta}$ that maximizes the probability in θ to get the samples X_k. Since the X_k are independent, the joint PDF for the whole dataset is

$$P(x_1, \cdots, x_N, \theta) = \prod_{k=1}^{N} p(x_k, \theta) \; .$$

This function is called *likelihood* when thinking of it as a function of the parameters θ. The maximum likelihood *estimator* is

$$\widehat{\theta} = \arg \max_{\theta} P(X_1, \cdots, X_N, \theta) \; .$$

The "arg max" refers to the value of θ that gives the largest value of P. It can be helpful to maximize the likelihood by using directional derivatives instead of gradients. For any function $F(\theta)$ this means that you look for $\widehat{\theta}$ using the condition that if $\dot{\theta}$ is any "perturbation direction", then

$$\frac{d}{d\epsilon} F(\widehat{\theta} + \epsilon\dot{\theta})\Big|_{\epsilon=0} = 0 \; .$$

This condition is the same as $\nabla_\theta(F(\widehat{\theta}) = 0$, because of the general directional derivative relation:

$$F(\widehat{\theta} + \epsilon\dot{\theta})\Big|_{\epsilon=0} = \dot{\theta}^T \nabla_\theta F(\widehat{\theta}) \; .$$

If θ has several "pieces", you can optimize over all of θ by optimizing over the pieces separately.

The multivariate normal density in d dimensions with mean μ and covariance matrix C is

$$p(x, \mu, C) = (2\pi)^{-\frac{d}{2}} \det(C)^{-\frac{1}{2}} \, e^{-\frac{1}{2}(x-\mu)^T C^{-1}(x-\mu)} \; .$$

It may be convenient to express this in terms of the *precision matrix*[1] $H = C^{-1}$.

$$p(x, \mu, H) = (2\pi)^{-\frac{d}{2}} \det(H)^{\frac{1}{2}} \, e^{-\frac{1}{2}(x-\mu)^T H (x-\mu)} \; .$$

If you maximize over H and take $\widehat{C} = \widehat{H}^{-1}$, you get the same answer when you maximize over C directly. This exercise is about finding $\widehat{\mu}$ and $\widehat{C}$ from N samples. In the abstract theory, $\theta = (\mu, C)$ or, equivalently, $\theta = (\mu, H)$. It is necessary that H and C are positive definite.

1. Show that $\nabla_\theta P(x_1, \cdots, x_N, \theta) = 0$ is equivalent to

$$\sum_{k=1}^{N} \nabla_\theta \log(p(x_k, \theta)) \; .$$

   Hint. $\nabla(uv) = \left(\frac{\nabla u}{u} + \frac{\nabla v}{v}\right) uv$ (why?).

2. Show that the maximum likelihood estimate of the mean is the empirical mean of the data:

$$\widehat{\mu} = \frac{1}{N} \sum_{k=1}^{N} X_k \; .$$

   Explain why we may take $\mu = \widehat{\mu}$ when maximizing over C or H to find $\widehat{C}$. *Hint.* To show this minimizes the likelihood, you need to use the fact that H is positive definite.

3. Show that

$$\det(H + \epsilon \dot{H}) = \det(H) \Big(1 + \epsilon \mathrm{Tr} \Big(H^{-1} \dot{H} \Big) + O(\epsilon^2) \Big) \; .$$

   Hint. You can see that $\det(\, I + \epsilon K) = 1 + \epsilon \mathrm{Tr}(K) + O(\epsilon^2)$ directly from the definition of the determinant as a sum over permutations of products. If K is diagonalizable, you can use the formula for determinant in terms of eigenvalues.

4. The empirical covariance matrix is

$$C_e = \frac{1}{N} \sum_{k=1}^{N} (X_k - \widehat{\mu}) \, (X_k - \widehat{\mu})^T \; .$$

   Show that

$$\sum_{k=1}^{N} (X_k - \widehat{\mu})^T H (X_k - \widehat{\mu}) = N \mathrm{Tr}(\, H C_e) \; .$$

5. Show that $\widehat{C} = C_e$. It may be helpful to first show $\widehat{H} = C_e^{-1}$ considering directional derivatives in all possible "directions" $\dot{H}$.

[1] *Precision* can be thought of as the inverse of *variance* in that a random variable has high precision if it has low variance.

Part 2, Importance sampling.

Simulation is often used to estimate the expected value of some random quantity. Suppose X is a random "object" (maybe just a number or maybe a multi-component path). Suppose $V(X)$ is some function of this random object and we want

$$A = \mathrm{E}[\, V(X)] \, .$$

The direct simulation estimate of A involves n independent "samples" X_k that are simulated to be from the distribution of X. The estimate is the sample mean using the samples generated:

$$\widehat{A} = \frac{1}{n} \sum_{k=1}^{n} V(X_k) \, .$$

The empirical variance from the samples is

$$\widehat{\sigma^2}_V = \frac{1}{n} \sum_{k=1}^{n} \Big(V(X_k) - \widehat{A} \Big)^2 \, .$$

This would be the maximum likelihood estimate of $\mathrm{var}(V(X))$ if it were Gaussian, which it is unlikely to be. The variance of $\widehat{A}$ is

$$\mathrm{var}\Big(\widehat{A}\Big) = \frac{1}{n} \mathrm{var}(V(X)) = \frac{1}{n} \sigma_V^2 \, .$$

The estimate of σ_V^2 gives an estimate for the standard deviation of the simulation based estimator, which is

$$\sigma_{\widehat{A}} \approx \widehat{\sigma}_{\widehat{A}} = \left(\frac{1}{n} \, \widehat{\sigma^2}_V \right)^{\frac{1}{2}} \, .$$

The quantity on the left is the one standard deviation *error bar*[2] for the estimator $\widehat{A}$. The *relative accuracy* of an estimate (of anything) is the ratio of the error in the estimate to the actual value. It is the "percentage" of error. When you are doing simulation based estimation, you can estimate the relative accuracy by

$$\text{rel err size} \approx \frac{\widehat{\sigma}_{\widehat{A}}}{\widehat{A}} \, . \tag{1}$$

To be clear, the relative error size (1) is positive while the actual error can be positive or negative. The actual relative error can be bigger than this, but it is unlikely to be more than twice as large.

[2]The term "error bar" comes from the practice of drawing the estimated value $\widehat{A}$ as a point in a graph with a bar extending up and down from $\widehat{A}$ of length $\sigma_{\widehat{A}}$ to indicate the uncertainty in the estimate. In general, experimental measurements are often put in graphs with error bars.

Importance sampling is a way to estimate A by sampling from a different probability density and then compensating by putting a *likelihood ratio* factor. If $X \sim p(\cdot)$, then

$$A = \int V(X)p(x)\,dx \ .$$

If q is another PDF, the likelihood ratio is

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

The importance sampling estimator of A with sampling distribution q is

$$A = \mathrm{E}_p[V(X)] = \mathrm{E}_q[V(X)L(X)] \ . \tag{2}$$

The notation E_p refers to the expected value when $X \sim p$. Importance sampling means using samples $X_k \sim q$ instead of $X_k \sim p$ and using the second part of (2). A well chosen distribution q can reduce the simulation estimate substantially. For this reason, importance sampling is routine in many or most serious uses of simulation for estimation.

1. Derive the second equality of (2). It is necessary to assume that L is never infinite, which means that $p(x) = 0$ whenever $q(x) = 0$.

2. Suppose $Z \sim \mathcal{N}(0,1)$ and we want to estimate $A = \Pr(Z > s)$. Let q be the distribution $\mathcal{N}(s,1)$. Calculate $L(z)$ in this case. Let $V(z)$ be the indicator function $V(z) = 1$ if $z > s$ and $V(z) = 0$ otherwise. Show that if s is large, then the variance of the q importance sampling estimator of A is exponentially (in s more accurate than the direct p estimator. *Hint.* What is the variance of each estimator?

3. Suppose $X = X_{[0,T]}$ is an approximate Brownian motion path starting from $X_0 = x_0$, generated using the Euler Maruyama approximation with time step $\Delta t = T/M$ (taking M steps to go from $t = 0$ to $t = T$). Let $p(x)$ be the joint PDF of the numbers $X_k \approx X_{t_k}$, for $k = 1, \cdots, M$. Write a formula for $p(x)$.

4. Now, let $X = X_{[0,T]}$ satisfy the SDE $dX = -r dt + dW_t$ and let $q(x)$ be the joint density of X_k for the Euler Maruyama approximation of this process with the same x_0 and Δt. Find a formula for the likelihood ratio $L(x)$. *Note.* This exercise is just (complicated) algebra. The result is related to *Girsanov's formula*, but that formula involves the limit $\Delta t \to 0$ and things we haven't discussed in class yet. If you look up and interpret that formula (which I do not necessarily recommend), you may recognize that your answer is consistent with it.

5. Use Euler Maruyama simulation of Brownian motion starting from x_0 to estimate $A = \Pr[\tau \leq T]$ with $T = 1$. Choose x_0 so that the probability is around a half (between, say, 35% and 70%, and x_0 so that the probability

is much smaller. Most of the code can come from a similar simulation from an earlier assignment. Implement the formula (1) to estimate the number of sample paths needed to get the answer to within 5% relative error. Compare the work needed for this accuracy for likely and unlikely hitting events.

6. Implement the importance sampling strategy from Exercise 4 to estimate the same probability using Brownian motion with a constant negative drift. Experiment to see how much computer time you can save using a good drift rate. How does this depend on x_0 and the probability that $\tau < T$? You do not need to record precise computer times. The runs should take long enough that you will experience the time they take.