Assignment #6 Solutions

1. Stochastic root-finding for drug design. (Drug-design problems motivated a lot of early work on root-finding algorithms. This homework problem is a highly simplified example.)
   (a) See the class website for Python code that will solve the problem. Of course, we are actually looking for the root of the shifted function $E[g(X,\theta)-20]$. Our code exploits numpy arrays to speed things up, but old-fashioned looping through arrays can be used instead. Our pilot run of 100 samples indicated that we needed roughly $n = 2400$ samples for our final estimation. We then obtained a final point estimate of $\theta_n \approx 0.8502$.
   (b) Again, see the website for Python code. As per the hint, minimizing the function $E[g(X,\theta)] = E[e^{-\theta X} + (\theta X / 2)]$ corresponds to finding a root of the expected derivative. I.e., by the hint, we have $\frac{\partial}{\partial\theta}E[g(X,\theta)] = E\left[g'(X,\theta)\right]$, where $g'(x,\theta) = \frac{\partial}{\partial\theta}g(x,\theta) = -xe^{-\theta x} + \frac{x}{2}$. Thus, solving $\frac{\partial}{\partial\theta}E[g(X,\theta)] = 0$ is equivalent to solving $E\left[g'(X,\theta)\right] = 0$. This latter problem is a root-finding problem of the same type as in Part (a), except using a different function. Our pilot run of 100 samples indicated that we needed roughly $n = 170$ samples for our final estimation. We then obtained a final point estimate of $\theta_n \approx 0.2085$. The estimated minimal discomfort level corresponding to θ_n is $g_n^{\text{opt}} \approx 0.8623$, which is indeed is slightly less than the true answer $g^{\text{opt}} \approx 0.8634$. So we think that we can do better than we actually can! Even worse, in *constrained* optimization problems, where we only consider values of θ that lie in a specified *feasible set* Θ, it is often the case that, if we use too few samples *n*, not only does g_n^{opt} look better that the unknown true solution g^{opt}, but the estimated solution θ_n, which is feasible with respect to the approximate sample-based optimization problem (that we solve to get our point estimate), is **not** feasible with respect to the actual problem. This erroneous sense of optimism is sometimes called the "optimizer's curse".
   (c) Following the hint, we have that $g_n^{\text{opt}} = \min_\theta \frac{1}{n}\sum_{i=1}^n g(X_i,\theta) \le \frac{1}{n}\sum_{i=1}^n g(X_i,\theta^*)$ for any θ^*. Taking expectations on both sides, we have $E[g_n^{\text{opt}}] \le E\left[\frac{1}{n}\sum_{i=1}^n g(X_i,\theta^*)\right] = \frac{1}{n}\sum_{i=1}^n E[g(X_i,\theta^*)] = E[g(X,\theta^*)]$. Since θ^* is arbitrary, it follows that $E[g_n^{\text{opt}}] \le \min_{\theta^*} E[g(X,\theta^*)] = g^{\text{opt}}$.

2. As suggested, write $\text{Corr}[U,V]^2 = g(\mu_1,\mu_2,\mu_3,\mu_4,\mu_5)$, where

$$g(x_1,x_2,x_3,x_4,x_5) = \frac{(x_5 - x_1x_2)^2}{(x_3 - x_1^2)(x_4 - x_2^2)}$$

   (a) Taylor-series method. The point estimate is

$$\begin{aligned}\alpha_n &= g\left(\bar{U}_n, \bar{V}_n, \overline{U_n^2}, \overline{V_n^2}, \overline{UV_n}\right) \\ &= g(14.4, 29.7, 282.2, 1115.9, 540) = 0.72097\end{aligned}$$

   To compute a confidence interval, follow the hint and observe that

$$\frac{\partial g}{\partial x_1}(x_1,\ldots,x_5) = 2a(acx_1q^{-2} - x_2q^{-1})$$
$$\frac{\partial g}{\partial x_2}(x_1,\ldots,x_5) = 2a(abx_2q^{-2} - x_1q^{-1})$$
$$\frac{\partial g}{\partial x_3}(x_1,\ldots,x_5) = \frac{-a^2c}{q^2}$$
$$\frac{\partial g}{\partial x_4}(x_1,\ldots,x_5) = \frac{-a^2b}{q^2}$$
$$\frac{\partial g}{\partial x_5}(x_1,\ldots,x_5) = \frac{2a}{q},$$

Where $a = (x_5 - x_1x_2)$, $b = (x_3 - x_1^2)$, $c = (x_4 - x_2^2)$, and $q = bc$. So

$$d_i = \frac{\partial g}{\partial x_i}\left(\bar{U}_n, \bar{V}_n, \overline{U_n^2}, \overline{V_n^2}, \overline{UV_n}\right) \text{ for } i = 1,2,3,4,5.$$
$$d_1 = -0.10384$$
$$d_2 = -0.00170$$
$$d_3 = -0.00963$$
$$d_4 = -0.00308$$
$$d_5 = 0.01284$$

And

$$s_n^2 = \frac{1}{9}\sum_{i=1}^{10}\left[d_1\left(U_i - \bar{U}_n\right) + d_2\left(V_i - \bar{V}_n\right) + d_3\left(U_i^2 - \overline{U_n^2}\right) + d_4\left(V_i^2 - \overline{V_n^2}\right) + d_5\left(U_iV_i - \overline{UV_n}\right)\right]^2$$
$$= 0.3308$$

95% confidence interval is

$$\left[0.721 - \frac{(1.96)(0.3308)^{1/2}}{\sqrt{10}}, 0.721 + \frac{(1.96)(0.3308)^{1/2}}{\sqrt{10}}\right] = [0.364, 1.077]$$

Since the correlation coefficient is always ≤ 1, we can take the CI to be [0.364,1.000].

(b) Jackknife method. (Almost the same answer as Part(a); using a spreadsheet makes the calculations go a lot faster.) Set

$$\alpha_n = g\left(\bar{U}_n, \bar{V}_n, \overline{U_n^2}, \overline{V_n^2}, \overline{UV_n}\right) = 0.72097 \text{ from part (a)}$$
$$\bar{U}_n(i) = \frac{1}{n-1}\sum_{\substack{j=1\\j\neq i}}^{n} U_j,\ \overline{U_n^2}(i) = \frac{1}{n-1}\sum_{\substack{j=1\\j\neq i}}^{n} U_j^2,\ \bar{V}_n(i) = \frac{1}{n-1}\sum_{\substack{j=1\\j\neq i}}^{n} V_j,$$
$$\overline{V_n^2}(i) = \frac{1}{n-1}\sum_{\substack{j=1\\j\neq i}}^{n} V_j^2,\ \overline{UV_n}(i) = \frac{1}{n-1}\sum_{\substack{j=1\\j\neq i}}^{n} U_jV_j$$
$$\alpha_n^i = g\left(\bar{U}_n(i), \bar{V}_n(i), \overline{U_n^2}(i), \overline{V_n^2}(i), \overline{UV_n}(i)\right) \quad i = 1,2,\ldots,10$$
$$\alpha_n^1 = 0.726568, \ldots, \alpha_n^{10} = 0.69062$$

$$\alpha_n(i) = n\alpha_n - (n-1)\alpha_n^i = 10\cdot(0.72097) - 9\alpha_n^i$$
$$\alpha_n(1) = 0.670587,\ldots,\alpha_n(10) = 0.99412$$
$$\alpha_n^J = \frac{1}{10}\big(\alpha_n(1)+\cdots+\alpha_n(10)\big) = \boxed{0.70088} \quad \text{(point estimator)}$$
$$V_n^J = \frac{1}{9}\Big(\big(\alpha_n(1)-\alpha_n^J\big)^2+\cdots+\big(\alpha_n(10)-\alpha_n^J\big)^2\Big) = 0.434552$$

95% asymptotic confidence interval is

$$\left[\alpha_n^J - \frac{1.96(V_n^J)^{1/2}}{\sqrt{n}}, \alpha_n^J + \frac{1.96(V_n^J)^{1/2}}{\sqrt{n}}\right] = [0.292, 1.109] \text{ so the answer is } \boxed{[0.292, 1.000]}$$

3. Multiple performance measures.
   (a) Following the hint, let

   $$I_j = \begin{cases} 1 & \text{if the } j\text{th CI brackets the } j\text{th performance measure} \\ 0 & \text{otherwise.} \end{cases}$$

   Also let N be the number of confidence intervals that do not contain their corresponding performance measure. Then $N = \sum_{j=1}^k (1-I_j) = k - \sum_{j=1}^k I_j$ and

   $$E[N] = E\left[k - \sum_{j=1}^k I_j\right] = k - \sum_{j=1}^k E[I_j] = k - \sum_{j=1}^k P(I_j = 1) = k - \sum_{j=1}^k (1-\alpha) = k\alpha.$$

   (b) Let A denote the event that all of the CIs bracket their respective performance measures. Again following the hint, we have, by Bonferroni's inequality,

   $$P(A) = P(A_1 \cap A_2 \cap \cdots \cap A_k) \ge 1 - P(A_1^c) - P(A_2^c) - \cdots - P(A_k^c) = 1 - k\alpha^*$$

   So set $\alpha^* = \alpha / k$ to ensure that $P(A) \ge 1-\alpha$. This procedure works reasonably well as long as k is relatively small. Note that we do not need to assume either normality or independence of the point estimators for the k measures. On the other hand, the bound may be "crude" in the sense that the true value of $P(A)$ might be much larger than $1-\alpha$; this means that our confidence intervals are wider than necessary.

4. Discounted reward. This problem shows that a number of interesting performance measures can be handled within the regenerative estimation framework discussed in class

   (a) Following the hint, we have

   $$\begin{aligned} r &= E\left[\int_0^{T_1} e^{-\beta u} q\big(X(u)\big)\,du\right] + E\left[e^{-\beta T_1}\int_{T_1}^{\infty} e^{-\beta(u-T_1)} q\big(X(u)\big)\,du\right] \\ &= E\left[\int_0^{T_1} e^{-\beta u} q\big(X(u)\big)\,du\right] + E\left[e^{-\beta T_1}\right] E\left[\int_{T_1}^{\infty} e^{-\beta(u-T_1)} q\big(X(u)\big)\,du\right] \\ &= E\left[\int_0^{T_1} e^{-\beta u} q\big(X(u)\big)\,du\right] + E\left[e^{-\beta T_1}\right] r, \end{aligned}$$

   Where the 2nd equality follows from the independence-from-the-past property of a regeneration point and the 3rd equality follows from identical-distribution property. Solving for *r*, we get

$$r = \frac{E\left[\int_0^{T_1} e^{-\beta u} q\big(X(u)\big)\,du\right]}{1 - E\left[e^{-\beta T_1}\right]} = \frac{E\left[\int_0^{T_1} e^{-\beta u} q\big(X(u)\big)\,du\right]}{E\left[1 - e^{-\beta T_1}\right]}$$

Thus, we take

$$X = \int_0^{T_1} e^{-\beta u} q\big(X(u)\big)\,du \quad \text{and} \quad Y = 1 - e^{-\beta T_1}.$$

(b) For the *i*th cycle, take

$$X_i = \int_{T_{i-1}}^{T_i} e^{-\beta(u - T_{i-1})} q\big(X(u)\big)\,du \quad \text{and} \quad Y_i = 1 - e^{-\beta(T_i - T_{i-1})}.$$