

Problem 1: Consider Monte Carlo integration with $g \geq 0$ and assume that $g \leq ch$ for some $c > 1$ and some density h . The *hit-or-miss* estimator of $z \stackrel{\text{def}}{=} \int_0^1 g(u)du$ is $c\mathbb{1}_{\{Uch(Y) \leq g(Y)\}}$, where U, Y are independent with U as uniform(0, 1) and Y with density h (note that h could be defined over the whole real line, but we restrict this to h only defined on $[0, 1]$). Show that its expectation is $z = \int_0^1 g(u)du$ as desired, but that the variance is always at least the variance of the importance sampling estimator that uses sampling from h .

Solution: Using a simple condition argument, we compute the expectation as,

$$\begin{aligned} \mathbf{E}[c\mathbb{1}_{\{Uch(Y) \leq g(Y)\}}] &= c \int_0^1 \mathbf{E}[\mathbb{1}_{\{Uch(y) \leq g(y)\}} | Y = y] h(y) dy \\ &= c \int_0^1 \frac{g(y)}{ch(y)} h(y) dy \\ &= \int_0^1 g(y) dy = z. \end{aligned}$$

Taking $H = c\mathbb{1}_{\{Uch(Y) \leq g(Y)\}}$, we have,

$$\text{Var}(H) = \mathbf{E}[H^2] - z^2 = c\mathbf{E}[H] - z^2 = cz - z^2.$$

For the importance sampling estimator Z , we have,

$$\text{Var}(Z) = \mathbf{E}_h[g(Y)^2/h(Y)^2] - z^2 = \int_0^1 g(y)^2/h(y) dy - z^2 \leq c \int_0^1 g(y) dy - z^2 = \text{Var}(H),$$

where we used $g \leq ch$ to get the inequality.

Problem 2: With $\phi(y)$ as in Problem 7 from Homework 1, show the following.

1. Prove that, for all functions g satisfying $\mathbf{E}[g(Y)^2] < \infty$, $\mathbf{E}[(X - \phi(Y))g(Y)] = 0$.
2. Prove that $\phi(Y)$ is the best mean square predictor of X across all predictors $g(Y)$ such that $\mathbf{E}[g(Y)^2] < \infty$.

Solution:

1. Via the law of iterated expectation, we have that

$$\mathbf{E}[(X - \phi(Y))g(Y)] = \mathbf{E}[\mathbf{E}[(X - \phi(Y))g(Y)|Y]].$$

Now we compute $\mathbf{E}[(X - \phi(Y))g(Y)|Y]$. We have

$$\mathbf{E}[(X - \phi(Y))g(Y)|Y] = g(Y)\mathbf{E}[X - \phi(Y)|Y],$$

since given Y , we certainly know $g(Y)$ so we take it out of the expectation. Now, previously, we have $\phi(Y) = \mathbf{E}[X|Y]$ so

$$\mathbf{E}[X - \phi(Y)|Y] = \mathbf{E}[X - \mathbf{E}[X|Y]|Y] = \mathbf{E}[X|Y] - \mathbf{E}[X|Y] = 0$$

Thus, we have

$$\mathbf{E}[(X - \phi(Y))g(Y)] = \mathbf{E}[\mathbf{E}[(X - \phi(Y))g(Y)|Y]] = \mathbf{E}[0] = 0,$$

as desired.

2. Proving this claim amounts to showing that for any predictor $g(Y)$, $\mathbf{E}[(X - g(Y))^2] \geq \mathbf{E}[(X - \phi(Y))^2]$.

$$\begin{aligned} \mathbf{E}[(X - g(Y))^2] &= \mathbf{E}[(X - \phi(Y) + \phi(Y) - g(Y))^2] \\ &= \mathbf{E}[(X - \phi(Y))^2] + \mathbf{E}[(\phi(Y) - g(Y))^2] + 2\mathbf{E}[(X - \phi(Y))(\phi(Y) - g(Y))] \\ &\geq \mathbf{E}[(X - \phi(Y))^2] + 2\mathbf{E}[(X - \phi(Y))\phi(Y)] - 2\mathbf{E}[(X - \phi(Y))g(Y)] \\ &= \mathbf{E}[(X - \phi(Y))^2] \end{aligned}$$

since both $\mathbf{E}[(X - \phi(Y))\phi(Y)]$ and $\mathbf{E}[(X - \phi(Y))g(Y)]$ are zero by the previous part. It follows that $\phi(Y)$ is the best mean square predictor of X across all predictors $g(Y)$.

Problem 3: VARIANCE REDUCTION WITH CONTROL VARIATES

Suppose that we wish to compute $\alpha = \mathbf{E}X$ via Monte Carlo. Assume that there exist r.v.'s Y_1, \dots, Y_d such that $\mathbf{E}Y_i = \mu_i$ is known for $1 \leq i \leq d$. The r.v. $C_i = Y_i - \mu_i$ is called a *control variate*.

1. If $C = (C_1, \dots, C_d)^T$, prove that

$$\mathbf{E}X(\lambda) = \alpha,$$

where $X(\lambda) = X - \lambda^T C$, $\lambda \in \mathbb{R}^d$.

2. Find the vector λ^* which minimizes $\text{Var}(X(\lambda))$ over λ .
3. How would you estimate λ^* in an implementation of this approach?
4. How general is this method? (i.e., are there typically r.v.'s for which the means μ_1, \dots, μ_d can be computed?) Explain your answer.

Solution:

- 1.

$$\mathbf{E}X(\lambda) = \mathbf{E}X - \mathbf{E}[\lambda^T C] = \mathbf{E}X - \lambda^T \mathbf{E}C = \alpha$$

since $\mathbf{E}C = 0$.

- 2.

$$\text{Var}(X(\lambda)) = \mathbf{E}[(X - \lambda^T C - \alpha)^2] = \mathbf{E}[(X - \alpha)^2] - 2\lambda^T \mathbf{E}[(X - \alpha)C] + \lambda^T \mathbf{E}[CC^T]\lambda$$

Setting σ^2 as the variance of X , σ_{XC} as the cross-covariance vector, Σ_C as the covariance matrix of C , we differentiate with respect to λ and set equal to 0:

$$-2\sigma_{XC} + 2\Sigma_C \lambda = 0$$

Thus, we set $\lambda^* = \Sigma_C^{-1} \sigma_{XC}$ which gives

$$X(\lambda^*) = X - \sigma_{XC}^T \Sigma_C^{-1} C,$$

and minimum variance

$$\sigma^2 - \sigma_{XC}^T \Sigma_C^{-1} \sigma_{XC}.$$

3. For sample means \bar{x} and \bar{c} , take the sample cross-covariance and covariance matrix:

$$\begin{aligned} \hat{\sigma}_{XC} &= \frac{1}{n} \sum_{i=1}^n (X_i - \bar{x})(C_i - \bar{c}) \\ \hat{\Sigma}_C &= \frac{1}{n} \sum_{i=1}^n (C_i - \bar{c})(C_i - \bar{c})^T \end{aligned}$$

The estimated λ^* is then

$$\hat{\lambda} = \hat{\Sigma}_C^{-1} \hat{\sigma}_{XC}.$$

However, if you notice, we can't just use this value right away. We would have to run some simulations in order to compute the estimate $\hat{\lambda}$. From there, we could run more simulations (not using the previous simulations to avoid bias) and use the control variates to finally compute the estimate of α .

4. You could certainly come up with all sorts of random variables for which the means are known. The point is that we want these random variables to be heavily dependent on each other. So if you have a complicated random variable, it's quite likely going to be difficult to find random variables to use as control variates.

Problem 4: The expected shortfall of a rv Z at the p 'th quantile is defined as

$$\mathbf{E}[Z|Z > q]$$

where q is the p 'th quantile value, that is $\mathbf{P}\{Z \leq q\} = p$.

This problem is concerned with estimating the expected shortfall of a so-called *Asian Option*. The option is based on an underlying asset V which evolves in discrete time according to

$$V_{n+1} = V_n R_{n+1}$$

where R_{n+1} is a log-Normal random variable, i.e.

$$\ln R_n \sim N(r, \sigma^2 \Delta t),$$

r is the risk free rate of return over one period, and σ^2 is the volatility of V . The price of a (simplified) Asian option with *expiration* N time units in the future is

$$X_N = \mathbf{E}_{V_0} \left(N^{-1} \sum_{i=1}^N V_i - K \right)_+$$

where K is the *strike price*. In general, an analytic expression is not known for X_N . (Note: $(x)_+ = x$ if $x \geq 0$ and $(x)_+ = 0$ if $x < 0$.)

We will treat the V_n 's as the weekly values of the asset. The annual risk-free rate of return is 5%, the volatility $\sigma^2 = 0.1$, $N = 50$ weeks¹, $V_0 = \$100$ and $K = \$115$. (Note: Since the risk-free rate and σ^2 were given as annual values, the generation of R_n must reflect that. The volatility in R_n was given as $\sigma^2 \Delta t$, where $\Delta t = \frac{1}{50}$, so it already adequately reflects this. For the weekly return, the effective rate is $r = \ln(1.05)/50 = 9.7580 \times 10^{-4}$.)

Using a bootstrap technique, compute the 95% confidence interval for the expected shortfall of the rv

$$\left(N^{-1} \sum_{i=1}^N V_i - K \right)_+,$$

with $p = 10\%$. Please include a description of the your method as well as the code.

Solution:

1. Generate the weekly returns of log-normal r.v.'s.
2. Compute the weekly prices of the asset as well as the terminal payout of the option.
3. Find q and compute the expected shortfall.
4. Bootstrap the values in the upper 10% and compute the expected shortfalls. (You could also bootstrap over all the prices, compute the payouts, the new q values, and the expected shortfalls.)
5. Use computed expected shortfalls to create 95% CI.

¹There are 50 working weeks in the financial year.

```

clear, clc
% parameters to the problem
n = 5000; m = 500;
r = 0.05; vol = .1; N = 50; v0 = 100; K = 115;
rweek = ((1+r)^(1/50)-1);
% Sample the weekly returns
Vsamples = exp(rweek*ones(N,n) + sqrt(vol/N)*randn(50,n));
% Create the price outcomes
V = v0*cumprod(Vsamples);
% Compute the option outcomes
P = sum(V)/N - K;
ind = find(P<0); P(ind) = 0;
% Compute the expected shortfall
q = quantile(P,.9);
ind = find(P>q); P90 = P(ind);
ExSF = mean(P90);
% Bootstrap to get the confidence interval
bootind = ceil(length(P90)*rand(m,length(P90)));
ExSFboot = mean(P90(bootind),2);
Z = quantile([ExSF;ExSFboot],[.025,.975]);
fprintf('Expected Shortfall: %f\n',ExSF);
fprintf('Bootstrapped 95% CI: [%f, %f]\n',Z(1),Z(2));

```

Problem 5: Suppose that $x(\cdot)$ is the solution to a deterministic differential equation

$$\frac{d}{dt}x(t) = \phi(\theta, x(t))$$

such that

$$x(0) = x_0$$

where ϕ is deterministic and θ represents a vector of parameters. (For example, $\phi(\theta, x) = \theta x$ in Example 13 on page 52.) Assume that x_0 and θ are measured with error, and that $(X_0, \hat{\theta})^T$ are multivariate normally distributed with mean $(x_0, \theta)^T$.

1. Compute the small noise approximations for the solution $X(t)$ to

$$\frac{d}{dt}X(t) = \phi(\hat{\theta}, X(t))$$

such that

$$X(0) = X_0$$

2. Discuss the computational issues that arise in computing the variance of your small noise approximation

Solution:

1. Solution Method 1

First, let's express $x(t)$ in integral form as

$$x(t) = x_0 + \int_0^t \phi(\theta, x(s)) ds.$$

Let

$$g(\theta, x_0; t) = x_0 + \int_0^t \phi(\theta, x(s)) ds,$$

where we note that $\phi(\theta, x(s))$ can have a dependence on x_0 . Just consider Example 3.13 of the notes. In that case,

$$\phi(\theta, x(s)) = \theta x(s) = \theta x_0 e^{\theta t}.$$

The small noise approximation requires computing

$$\frac{\partial}{\partial \theta} g(\theta, x_0; t) = \int_0^t \frac{\partial}{\partial \theta} \phi(\theta, x(s)) ds$$

and

$$\frac{\partial}{\partial x_0} g(\theta, x_0; t) = 1 + \int_0^t \frac{\partial}{\partial x_0} \phi(\theta, x(s)) ds.$$

Thus, we have that $X(t) - x(t) \sim \mathcal{N}(0, \sigma(t)^2)$ where (for Σ the covariance matrix for $(\hat{\theta}, X_0)$)

$$\sigma(t)^2 = \left(\int_0^t \frac{\partial}{\partial \theta} \phi(\theta, x(s)) ds, 1 + \int_0^t \frac{\partial}{\partial x_0} \phi(\theta, x(s)) ds \right)^T \Sigma \left(\int_0^t \frac{\partial}{\partial \theta} \phi(\theta, x(s)) ds, 1 + \int_0^t \frac{\partial}{\partial x_0} \phi(\theta, x(s)) ds \right),$$

Solution Method 2

We are looking to say something about $X(t) - x(t)$ but we know only about $\frac{d}{dt}(X(t) - x(t))$. So let's start there with the small noise approximation:

$$\begin{aligned} \frac{d}{dt}(X(t) - x(t)) &= \phi(\hat{\theta}, X(t)) - \phi(\theta, x(t)) \\ &\approx \phi(\theta, x(t)) + \frac{\partial}{\partial \theta} \phi(\theta, x(t))(\hat{\theta} - \theta) + \frac{\partial}{\partial x_0} \phi(\theta, x(t))(X_0 - x_0) - \phi(\theta, x(t)) \\ &= \frac{\partial}{\partial \theta} \phi(\theta, x(t))(\hat{\theta} - \theta) + \frac{\partial}{\partial x_0} \phi(\theta, x(t))(X_0 - x_0) \end{aligned}$$

Integrating gives

$$X(t) - x(t) - X_0 - x_0 \approx (\hat{\theta} - \theta) \int_0^t \frac{\partial}{\partial \theta} \phi(\theta, x(s)) ds + (X_0 - x_0) \int_0^t \frac{\partial}{\partial x_0} \phi(\theta, x(s)) ds$$

Rearranging gives

$$X(t) - x(t) \approx (\hat{\theta} - \theta) \int_0^t \frac{\partial}{\partial \theta} \phi(\theta, x(s)) ds + (X_0 - x_0) \left(1 + \int_0^t \frac{\partial}{\partial x_0} \phi(\theta, x(s)) ds \right),$$

which is exactly what we saw above.

Alternative Solution from Tzu-Wei

Use the small noise approximation of ϕ at x_0 and θ . For small t , we have

$$\frac{d}{dt} x(t) = \phi(\theta, x(t)) \approx \phi(\theta, x_0) + \frac{\partial}{\partial x} \phi(\theta, x_0)(x(t) - x_0)$$

and

$$\frac{d}{dt} X(t) = \phi(\theta, X(t)) \approx \phi(\theta, x_0) + \frac{\partial}{\partial x} \phi(\theta, x_0)(x(t) - x_0) + \frac{\partial}{\partial \theta} \phi(\theta, x_0)(\hat{\theta} - \theta)$$

so

$$\frac{d}{dt}(X(t) - x(t)) \approx \frac{\partial}{\partial x} \phi(\theta, x_0)(X(t) - x(t)) + \frac{\partial}{\partial \theta} \phi(\theta, x_0)(\hat{\theta} - \theta)$$

We now have a first-order ODE in $X(t) - x(t)$. Solving this gives

$$X(t) - x(t) \approx (X_0 - x_0) e^{\frac{\partial}{\partial x} \phi(\theta, x_0)t} + \frac{\partial}{\partial \theta} \phi(\theta, x_0)(\hat{\theta} - \theta) \int_0^t e^{\frac{\partial}{\partial x} \phi(\theta, x_0)(t-s)} ds$$

Therefore $X(t)$ is approximated by

$$X(t) \approx x(t) + a(t)(X_0 - x_0) + b(t)(\hat{\theta} - \theta)$$

with $a(t) = e^{\frac{\partial}{\partial x}\phi(\theta, x_0)t}$ and $b(t) = \frac{\partial}{\partial \theta}\phi(\theta, x_0) \int_0^t e^{\frac{\partial}{\partial x}\phi(\theta, x_0)(t-s)} ds$. Assume the original covariance matrix is C . Finally we conclude that

$$X(t) \approx x(t) + N(0, BCB^T), B = \begin{pmatrix} a(t) & 0 \\ 0 & b(t) \end{pmatrix}$$

We can then repeatedly solve this problem by taking time steps from 0 to t_1 , to t_2 and so forth. At each time, the problem looks essentially identical to the one we just showed. The changes will be the initial condition we use each time and the covariance matrix used in the small noise approximation.

2. If we use the first SNA, computing the variance is going to require dealing with those integrals. If ϕ takes a particularly nasty form, we may run into trouble if we have to do this a lot. The second approach doesn't require the same integration steps. We just need the derivative values. However, the possibility of error propagation is a concern.

Problem 6: Develop a corresponding approximation confidence interval for $q(p)$, where $q(p)$ is the " p^{th} quantile" of the random variable Y defined as the smaller root of the equation

$$P\{Y \leq q(p)\} = p$$

Such quantile computations are of interest on "value at risk" calculations in the finance setting.

Solution:

We proceed similarly to Example 9 on page 53 of the notes. For any n , define the empirical distribution as

$$F_n(x) = n^{-1} \sum_{j=1}^n \mathbb{1}_{\{Y_j \leq x\}}.$$

The p -th quantile estimator $q_n(p)$ approximately satisfies the equation

$$F_n(q_n(p)) = p.$$

For a well chosen n , this equation is satisfied exactly. Namely, when pn is an integer. From here on, we suppose a well chosen n . Continuing, we have

$$F_n(q_n(p)) - F_n(q(p)) = p - F_n(q(p)).$$

Let's now invoke the CLT. We see that $F_n(q(p))$ is a random value. It doesn't necessarily give us p . It returns a random value that depends on our sample $Y_j, j = 1, \dots, n$. But we see that $F_n(q(p))$ has mean

$$\mathbf{E}[F_n(q(p))] = \frac{1}{n} \sum_{j=1}^n \mathbf{E}[\mathbb{1}_{\{Y_j \leq q(p)\}}] = p.$$

and variance, by independence of the Y_j ,

$$\text{Var}(F_n(q(p))) = \frac{1}{n} \text{Var}(\mathbb{1}_{\{Y \leq q(p)\}}) = \frac{p(1-p)}{n}.$$

Thus, the CLT gives

$$F_n(q_n(p)) - F_n(q(p)) = p - F_n(q(p)) \stackrel{\mathcal{D}}{\approx} \mathcal{N}\left(0, \frac{p(1-p)}{n}\right). \quad (1)$$

But we don't want to consider $F_n(q(p))$ since that's not something we can work with. We're actually interested in $q_n(p)$. To resolve this, we use the small noise approximation technique. If the distribution

function $F(\cdot) = \mathbf{P}(Y \leq \cdot)$ is strictly increasing and continuous, then $q_n(p) \xrightarrow{\text{a.s.}} q(p)$ as $n \rightarrow \infty$. So for large n , we say (without refined rigor)

$$F_n(q_n(p)) - F_n(q(p)) \approx F(q_n(p)) - F(q(p)) \approx F'(q(p))(q_n(p) - q(p)) = f(q(p))(q_n(p) - q(p)). \quad (2)$$

We then combine (1) and (2) to get

$$q_n(p) - q(p) \approx \frac{1}{f(q(p))} (p - F_n(q(p))) \stackrel{\mathcal{D}}{\approx} \mathcal{N}\left(0, \frac{p(1-p)}{nf(q(p))^2}\right).$$

Concluding, the approximate $100(1 - \delta)\%$ confidence interval we would give for the p -quantile value $q(p)$ is

$$\left[q_n(p) - z \frac{\sqrt{p(1-p)}}{\sqrt{nf(q(p))}}, q_n(p) + z \frac{\sqrt{p(1-p)}}{\sqrt{nf(q(p))}} \right],$$

where z is selected such that $\mathbf{P}(-z \leq \mathcal{N}(0, 1) \leq z) = 1 - \delta$. Of course, you can see that we require knowledge of $q(p)$ so we would naturally extend this result to using the estimators for unknown quantities like we do with the sample variance for the true variance.

Problem 7: Let X be a Cauchy r.v. so that its density is given by

$$f(x) = \frac{1}{\pi(1 + (x - b)^2)}$$

1. Compute the distribution of $X_1 + X_2$ where X_1 , and X_2 are independent copies of X .
2. If X_1, X_2, \dots, X_n is an i.i.d. sample from X , what does $n^{-1}(X_1 + \dots + X_n)$ converge to?
3. How might you estimate the parameter b (in lieu of part 2)?
4. Explain how to generate the r.v. X using inversion.
5. Consider the following algorithm:
 - i. Generate independent r.v.'s V_1 and V_2 that are uniform on $[-1, 1]$.
 - ii. If $V_1^2 + V_2^2 \leq 1$, return $X = V_2/V_1$; else, return to step i.

Show that X is Cauchy distributed. (This can be faster than inversion when computing the arc tangent is expensive.)

Solution:

1. Appeal to the characteristic functions:

$$\mathbf{E}[\exp(i\theta(X_1 + X_2))] = \mathbf{E}[\exp(i\theta X_1)]^2 = \exp(2bi\theta - 2|\theta|)$$

which is the characteristic function for a Cauchy r.v. with median $2b$ and shape parameter 2 . That is

$$f_{X_1+X_2}(x) = \frac{1}{2\pi(1 + (\frac{x-2b}{2})^2)}$$

2. Redo the above computation but for $\frac{1}{2}(X_1 + X_2)$. You get the original distribution. You can extend this to $\frac{1}{n}(X_1 + \dots + X_n)$. This shows that the sample mean isn't converging to anything; it's always Cauchy distributed. Thus, the sample mean isn't all that helpful for this distribution.
3. Though the expectation of the Cauchy distribution is undefined, the median is not. The sample median does in fact converge to the median, b , and this will give us an estimator for the parameter b .

4. A quick integration (or even quicker look at Wikipedia) gives the cdf

$$F_X(x) = \frac{1}{\pi} \arctan(x - b) + \frac{1}{2}.$$

Inverting this gives the generation scheme:

$$X = b + \tan(\pi(U - \frac{1}{2}))$$

where $U \sim \text{Uniform}[0, 1]$.

5. This algorithm is generating points uniformly distributed on the unit disc and then returning the ratio of the coordinates. In terms of polar coordinates (R, θ) , this is $X = \tan(\theta)$. And if we wanted to generate points uniformly on the unit disc using polar coordinates, we would take $\theta \sim \text{Uniform}[-\pi/2, \pi/2]$ ($\tan(\theta)$ returns the same values for the first and third quadrants and second and fourth quadrants so we only need to distribute θ in the first and fourth quadrants). So $X = \tan(\theta) = \tan(\pi(U - 1/2))$ where $U \sim \text{Uniform}[0, 1]$. But that's precisely the inversion formula for the case $b = 0$. So to generate a Cauchy r.v. with location b , we just take $X = b + V_2/V_1$ in the original algorithm.

Problem 8: Consider a linear model in which

$$Y_i = a^* x_i + b^* + \epsilon_i, \quad 1 \leq i \leq n,$$

where the ϵ_i 's are iid rvs with density

$$f(x) = \frac{\lambda^*}{2} e^{-\lambda^* |x|}.$$

Note: From a terminology standpoint, it is important to recognize that this is different from a linear regression problem. A linear regression problem aims to find a least squares solution for the model. In this problem, we are looking for maximum likelihood estimates for the model. These are distinct problems, except when the error is normally distributed. In that case, they are equivalent.

1. The MLE for a^* , b^* and λ^* solves an optimization problem. What is it?
2. Show that the MLE can be computed as the solution to a linear program.

Solution:

1. Since ϵ_i have the density f above, $Y_i - a^* x_i - b^*$ has this density. Thus, we define the likelihood function $L_n(a, b, \lambda)$ to be

$$\begin{aligned} L_n(a, b, \lambda) &= \prod_{i=1}^n \frac{\lambda}{2} e^{-\lambda |Y_i - ax_i - b|} \\ &= \left(\frac{\lambda}{2}\right)^n \exp\left(-\lambda \sum_{i=1}^n |Y_i - ax_i - b|\right) \end{aligned}$$

Thus, the optimization problem to solve is

$$\begin{aligned} &\text{maximize} \quad \left(\frac{\lambda}{2}\right)^n \exp\left(-\lambda \sum_{i=1}^n |Y_i - ax_i - b|\right) \\ & \quad a, b, \lambda \end{aligned}$$

2. To convert this optimization problem to a linear program, we first convert the likelihood function to the log-likelihood function

$$\mathcal{L}_n(a, b, \lambda) = n \ln(\lambda) - \lambda \sum_{i=1}^n |Y_i - ax_i - b|.$$

Now, its quite evident this is not a linear program because of the term $\ln(\lambda)$ among other things. So now we optimize individually for λ by taking $\partial\mathcal{L}_n/\partial\lambda = 0$ to get,

$$\frac{n}{\lambda} - \sum_{i=1}^n |Y_i - ax_i - b| = 0.$$

Thus, we have $\lambda^* = n/\sum_{i=1}^n |Y_i - ax_i - b|$. Substitution yields,

$$\begin{aligned} \mathcal{L}_n(a, b, \lambda^*) &= n \ln \left(\frac{n}{\sum_{i=1}^n |Y_i - ax_i - b|} \right) - n \\ &= n \ln n - n \ln \left(\sum_{i=1}^n |Y_i - ax_i - b| \right) - n \end{aligned}$$

To conclude, we recognize that maximizing this log-likelihood function completely boils down to minimizing the summation. Thus, we are finally left with the linear problem,

$$\underset{a, b}{\text{minimize}} \quad \sum_{i=1}^n |Y_i - ax_i - b|.$$

While this technically isn't in proper linear program format, it can easily be turned into a linear program by adding some constraints.

Problem 9: LEAST SQUARES ESTIMATION

Name	Height (in)	Weight (lbs)	Age
Alfred	69.0	112.5	14
Alice	56.5	84.0	13
Barbara	65.3	98.0	13
Carol	62.8	102.5	14
Henry	63.5	102.5	14
James	57.3	83.0	12
Jane	59.8	84.5	12
Janet	62.5	112.5	15
Jeffrey	62.5	84.0	13
John	59.0	99.5	12
Joyce	51.3	50.5	11
Judy	64.3	90.0	14
Louise	56.3	77.0	12
Mary	66.5	112.0	15
Philip	72.0	150.0	16
Robert	64.8	128.0	12
Ronald	67.0	133.0	15
Thomas	57.5	85.0	11
William	66.5	112.0	15

Table 1: Data For Problem 9

1. Use the first two columns of data (Height and Weight) in Table 1 to build various regression models that attempt to explain weight as a function of height. Which of your models does the best job of predicting the weights of individuals in the third column? (There is no need to use the age here. If you wanted to, there is a way to factor in the age.)
2. Another possibility is to build separate regression models by gender. Does this improve the predictions?

- Now use the full data set to recompute the coefficients for the type of regression model that worked best in part a. Construct a confidence interval for the slope based on bootstrapping. (Take into account the fact that both weight and height should be viewed as random variables in this setting. In other words, this is not a setting where the levels of the explanatory variable (in this case, height) is carefully set by the experimenter at various predetermined levels; the height values that are observed are determined by the particular random sample that is selected.)
- Suppose that the height of a student is 71 inches. Use the bootstrap to construct a 95% prediction interval for that student's weight.

Solution:

- Let y be weight. We consider four regression models: height vs. y , height² vs. y (this is like looking at surface area vs. weight), height³ vs. y (this is like looking at volume vs. weight) and weighted least squares of x vs. y using ages as weights. The least squares estimators were computed using the normal equations $A^T Ax = A^T b$ (and $A^T W Ax = A^T W b$ for the weighted case), which is practical since the data set is small and the matrix A is not ill-conditioned. The results are summarized in the following table, where the general regression has the form:

$$y = ax + b$$

The quality of the model was measured by the MSE ($= \|r\|_2^2/n$ where r is the residual vector and $n = 19$ is the sample size). Based on the MSE, the second model does the best job predicting the

x	a	b	MSE
height	3.8990	-143.0269	112.76
height ²	0.0315	-23.3440	112.52
height ³	0.0003	16.8331	113.62
height (weighted LS)	3.929	-144.8399	112.79

weights of the individuals. This can be interpreted as measuring weight as a function of body "surface area".

- We consider the same regression models as in part 1 first for the men and then for the women. The results are summarized in the following table.

Men:

One can see that the MSE is smaller than when the regression models were run on the entire popu-

x	a	b	MSE
height	3.9125	-141.1010	128.94
height ²	0.0307	-17.0132	126.24
height ³	0.0003	24.3945	124.32
height (weighted LS)	3.9936	-146.3970	129.10

lation. This indicates that a separate regression model for the men improves the predictions. Observe also that when the men are considered separately, the best model is height³ vs. weight.

Women:

The MSE has decreased drastically for the women data. Again, we conclude that separate regression models by gender improve the predictions. For the women, it appears as though height vs. weight is the best model.

- The regression model that worked best in part 1) was x vs. y where $x = \text{height}^2$ and $y = \text{weight}$. The coefficients computed were $\hat{a} = 0.0315$ and $\hat{b} = -23.344$ where the model was given by $y = \hat{a}x + \hat{b}$. Here, we construct a CI for the slope of the regression line, namely \hat{a} .

x	a	b	MSE
height	3.4244	-117.3698	71.48
height ²	0.0286	-15.6465	76.54
height ³	0.0003	18.2995	82.06
height (weighted LS)	3.3789	-114.2258	71.68

Begin by generating $n = 19$ i.i.d. $\mathcal{N}(0, s^2)$ rvs, where s^2 is the sample variance of $y_i - \hat{a}x_i - \hat{b}$. We then use these ϵ_i to compute $y_{i1} = \hat{a}x_i + \hat{b} + \epsilon_i$ using the given values of x_i . This “synthetic” data set is used to recompute the slope, call it $\hat{\alpha}_*^1$. The procedure is repeated many times to get 1000 independent slopes $\hat{\alpha}_*^i$. To find a $1 - \delta\%$ CI, z_1 and z_2 were computed such that

$$\mathbf{P}(z_1 \leq \hat{a} - a^* \leq z_2) = 1 - \delta$$

From this it follows that a confidence interval for the slope is

$$[\hat{a} - z_2, \hat{a} - z_1]$$

Suppose we want a 95% CI. In our implementation, we found the z_i to be: $z_1 = -0.0055$, $z_2 = 0.0132$. Using these values, a 95% CI is $[0.0183, 0.0370]$.

- To construct the bootstrap prediction interval, we apply a similar procedure as before. Again, we generate a vector of ϵ_i which are normally distributed with mean 0 and variance equal to the sample variance of the residuals using the parameters \hat{a} and \hat{b} computed in part 3. We then re-estimate the parameters \hat{a} 1000 times and use each of these models to predict the weight of someone who is 71 inches tall. Let \hat{y}_{71}^i be the predictors generated by the bootstrapping procedure. To get a 95% PI for the weight, we first find z_1 and z_2 such that

$$\mathbf{P}(z_1 \leq \hat{y}_{71} - y_{71}^* \leq z_2) = 0.95$$

from which it follows that a 95% CI is $[\hat{y}_{71} - z_2, \hat{y}_{71} - z_1]$. Implementing this procedure in MATLAB gave $z_1 = -17.5795$ and $z_2 = 8.1544$, giving the 95% PI for y^* as $[123.8299, 142.1349]$.

Problem 10: When the arrival process to a single-server queue follows a so-called Poisson process having rate λ (i.e. the inter-arrival times χ_1, χ_2, \dots are i.i.d. exponential random variables having parameter $\lambda > 0$) with $\rho = \lambda \mathbf{E}V_0 < 1$, the steady-state random variable W_∞ corresponding to the time spent waiting in the queue can be represented as

$$W_\infty = \sum_{i=1}^N Z_i$$

where N is a geometric random variable having probability mass function $\mathbf{P}(N = k) = (1 - \rho)\rho^{k-1}$ and $(Z_i : i \geq 1)$ is an i.i.d. sequence of random variables independent of N satisfying

$$\mathbf{P}(Z_1 > x) = \frac{1}{\mathbf{E}V_0} \int_x^\infty \mathbf{P}(V_0 > y) dy$$

- Suppose that V_0 is exponential with parameter $\mu > 0$. Compute the distribution of W_∞ .
- Abandoning the assumption in part (a), write W_∞ as $W_\infty(\lambda)$ (reflecting its dependence on λ). Prove that if $\mathbf{E}V_0^2 < \infty$, $(1 - \rho)W_\infty(\lambda) \Rightarrow \Gamma$ as $\lambda \rightarrow \frac{1}{\mathbf{E}V_0}$ and compute the distribution of Γ (Γ is not necessarily a Γ distributed random variable. We just needed a letter and it seemed as good as any other).

Hint: Convergence in distribution is equivalent to point-wise convergence of characteristic functions. Use that fact to show the desired convergence as well as to derive the limiting characteristic function, and thus the distribution of Γ .

(This is known in the performance engineering literature as the “heavy traffic” theorem for queues.)

3. Suppose that V_0 is gamma distributed with shape parameter $\alpha = 2$ and scale parameter 1. What is an approximation to $P(W_\infty > 2)$ if $\lambda = 0.45$?

Solution:

1. First note that

$$\mathbf{P}(V_0 > y) = e^{-\mu y}$$

Thus

$$\mathbf{P}(Z > x) = \mu \int_x^\infty P(V_0 > y) dy = \mu \int_x^\infty e^{-\mu y} dy = e^{-\mu x}.$$

Thus $Z \sim \text{Exp}(\mu)$. We appeal to characteristic functions and iterated expectation to find the distribution of W_∞ :

$$\begin{aligned} \Phi_{W_\infty}(\theta) &= \sum_{n=1}^{\infty} \mathbf{P}(N = n) \Phi_Z(\theta)^n && \text{Iter. Ex. \& Ind. of } Z_i \\ &= \sum_{n=1}^{\infty} (1 - \rho) \rho^{n-1} \Phi_Z(\theta)^n \\ &= (1 - \rho) \Phi_Z(\theta) \sum_{n=0}^{\infty} \rho^n \Phi_Z(\theta)^n \\ &= \frac{(1 - \rho) \Phi_Z(\theta)}{1 - \rho \Phi_Z(\theta)} && \text{Geom. Ser.: } |\rho \Phi_Z| < 1 \\ &= \frac{(1 - \rho)(1 - i\theta/\mu)^{-1}}{1 - \rho(1 - i\theta/\mu)^{-1}} \\ &= \frac{1 - \rho}{(1 - \rho) - i\theta/\mu} \\ &= \frac{1}{1 - i\theta/(\mu - \lambda)} && (1 - \rho) = 1 - \lambda/\mu \end{aligned}$$

So $W_\infty \sim \text{Exp}(\mu - \lambda)$.

2. First note that $\lambda \nearrow \frac{1}{\mathbf{E}V_0}$ is equivalent to $\rho \nearrow 1$. Now appealing to the characteristic function and iterated expectation, we have

$$\begin{aligned} \Phi_{(1-\rho)W_\infty(\lambda)}(\theta) &= \sum_{n=1}^{\infty} \mathbf{P}(N = n) \Phi_{(1-\rho)Z}(\theta)^n && \text{Iter. Ex. \& Ind. of } Z_i \\ &= \sum_{n=1}^{\infty} (1 - \rho) \rho^{n-1} \mathbf{E}[e^{i\theta(1-\rho)Z}]^n \\ &= (1 - \rho) \mathbf{E}[e^{i\theta(1-\rho)Z}] \sum_{n=0}^{\infty} \rho^n \mathbf{E}[e^{i\theta(1-\rho)Z}]^n \\ &= \frac{(1 - \rho) \mathbf{E}[e^{i\theta(1-\rho)Z}]}{1 - \rho \mathbf{E}[e^{i\theta(1-\rho)Z}]} && \text{Geom. Ser.: } |\rho \Phi_Z| < 1 \end{aligned}$$

Since $|e^{i\theta(1-\rho)Z}| \leq 1$, we are allowed to bring the limit $\lim_{\rho \rightarrow 1}$ under the expectation. However, we quickly note that we have the indeterminate limit $\frac{0}{0}$ because

$$\lim_{\rho \rightarrow 1} \mathbf{E}[e^{i\theta(1-\rho)Z}] = 1.$$

However, we can apply L'Hospital's rule here because a wonderful property of characteristic functions is this:

$$\lim_{\theta \searrow 0} \frac{d}{d\theta} \mathbf{E}[e^{i\theta X}] = \mathbf{E}[iX e^{i\theta X}] = i\mathbf{E}[X]$$

So after applying this to our usage of L'Hospital's rule, we have

$$\frac{-\mathbf{E}[e^{i\theta(1-\rho)Z}] - i\theta(1-\rho)\mathbf{E}[Ze^{i\theta(1-\rho)Z}]}{-\mathbf{E}[e^{i\theta(1-\rho)Z}] + i\theta\rho\mathbf{E}[Ze^{i\theta(1-\rho)Z}]} \xrightarrow{\rho \nearrow 1} \frac{1}{1 - i\theta\mathbf{E}[Z]}$$

Thus, no matter the distribution of Z , so long as it is integrable, we have that $(1-\rho)W_\infty(\lambda) \Rightarrow \Gamma \sim \text{Exp}(1/\mathbf{E}[Z])$. But we can go further with this, namely computing $\mathbf{E}[Z]$ from V_0 . We use the following:

$$\mathbf{E}[Z] = \mathbf{E}\left[\int_0^Z dt\right] = \mathbf{E}\left[\int_0^\infty I(Z > t)dt\right] \tag{3}$$

Using Fubini's Theorem, we can interchange the expectation and the integral since the integrand is positive and plug in the distribution given in the problem, giving:

$$\mathbf{E}[Z] = \int_0^\infty \mathbf{E}[I(Z > t)] dt = \int_0^\infty P(Z > t)dt = \int_0^\infty \int_t^\infty \frac{P(V_0 > y)}{\mathbf{E}[V_0]} dy dt \tag{4}$$

We expand the domain of integration in t by adding two indicator functions: $I(y > t)$ to maintain the same range of integration, and $I(V_0 > t)$ to make sure that V_0 is in the proper range as well, giving:

$$\begin{aligned} \mathbf{E}[Z] &= \int_0^\infty \int_0^\infty I(y > t)I(V_0 > t) \frac{P(V_0 > y)}{\mathbf{E}[V_0]} dy dt \\ &= \int_0^\infty \int_0^\infty I(y > t)I(V_0 > t) \frac{\mathbf{E}[I(V_0 > y)]}{\mathbf{E}[V_0]} dy dt \\ &= \int_0^\infty \frac{\mathbf{E}[I(V_0 > y)]}{\mathbf{E}[V_0]} \int_0^\infty I(y > t)I(V_0 > t) dt dy \end{aligned} \tag{5}$$

This the integrand with respect to time becomes: $I(t < y)I(t < V_0)$ which has support only on the interval $[0, \min\{y, V_0\}]$, and takes on the value of 1 there. Thus:

$$\begin{aligned} \mathbf{E}[Z] &= \int_0^\infty \frac{\mathbf{E}[I(V_0 > y)]}{\mathbf{E}[V_0]} \int_0^{\min\{y, V_0\}} dt dy \\ &= \frac{1}{\mathbf{E}[V_0]} \int_0^\infty \mathbf{E}[I(y < V_0)] \min\{y, V_0\} dy \end{aligned} \tag{6}$$

Again, applying Fubini's Theorem, we can interchange the integral and the expectation, obtaining:

$$\mathbf{E}[Z] = \frac{\mathbf{E}\left[\int_0^\infty I(y < V_0) \min\{y, V_0\} dy\right]}{\mathbf{E}[V_0]} \tag{7}$$

On the domain in which the indicator variable takes on a non-zero value ($y < V_0$) y is the minimum of $\{y, V_0\}$, thus:

$$\mathbf{E}[Z] = \frac{\mathbf{E}\left[\int_0^{V_0} y dy\right]}{\mathbf{E}[V_0]} = \frac{\mathbf{E}[V_0^2]}{2\mathbf{E}[V_0]} \tag{8}$$

This suggests that we can model the distribution of W_∞ with:

$$F_{W_\infty}(x) = 1 - e^{-\frac{1}{\mathbf{E}[Z]}x} = 1 - e^{-\frac{2\mathbf{E}[V_0]}{\mathbf{E}[V_0^2]}x}$$

3. We are given a gamma distribution with shape parameter $\alpha = 2$ and scale parameter $\mu = 1$. Thus:

$$W_0 \sim \frac{\mu}{\Gamma(\alpha)} (\mu x)^{\alpha-1} e^{-\mu x} = x e^{-x} \tag{9}$$

The mean and the variance for a gamma distribution are well known:

$$\mathbf{E}[V_0] = \frac{\alpha}{\mu} = 2 \quad \text{and} \quad \mathbf{E}[V_0^2] = \text{Var}[V_0] + \mathbf{E}[V_0]^2 = 2 + 4 = 6$$

Thus

$$\rho = \lambda \mathbf{E}[V_0] = 0.45 \times 2 = 0.9 \approx 1$$

and we see that the heavy traffic limit result applies. Therefore we can approximate the distribution of W_∞ as:

$$F_{W_\infty} \sim 1 - e^{-\frac{2 \times 2}{6}x} = 1 - e^{-\frac{2}{3}x}$$

Therefore we can approximate the probability as:

$$\mathbf{P}(W_\infty > 2) = P((1 - \rho)W_\infty > 2(1 - \rho)) \approx 1 - (1 - e^{-\frac{2}{3} \times 2(1 - \rho)}) = e^{-\frac{4}{3} \times 0.1} = e^{-\frac{4}{30}}$$