MATH 472/572 Computational Statistics - Spring 2020

Final EXAM - Noncoding Part

May 5 (Tuesday), 2:30 PM- 4:30PM

Student Name: _____ Student ID: _____

EXAM rule: This is a close-book and close-notes exam. No calculator is allowed. No any form of discussion or communication with others is allowed.

Answers by themselves are not adequate without clearly indicating your reasoning.

Note: Please write your answer in a blank piece of paper or in a digital file. Upload your solution file (a single PDF file is preferred) to the blackboard course site under Assignment Final link.

---

1. (10 pts) Suppose that $E(Y|X) = s(x)$ for a smooth function s. For a given point x, let $\hat{s}(x)$ be an estimator of $s(x)$. The quality of $\hat{s}(x)$ as an estimator of $s(x)$ at x can be measured by the Mean Squared Error (MSE) at x:

$$\mathrm{MSE}\left[\hat{s}(x)\right] = \mathrm{E}\left\{\left[\hat{s}(x) - s(x)\right]^2\right\}.$$

Define:

$$\mathrm{bias}\left[\hat{s}(x)\right] = E\left[\hat{s}(x)\right] - s(x).$$

Prove:

$$\mathrm{MSE}\left[\hat{s}(x)\right] = \left\{\mathrm{bias}\left[\hat{s}(x)\right]\right\}^2 + \mathrm{Var}\left[\hat{s}(x)\right].$$

2. (20 pts)For a linear regression model:

$$Y = X\beta + \epsilon,$$

where $Y \in R^n$ is the response vector, $X \in R^{n\times p}$ (where $n > p$) is the given predictor matrix, $\beta \in R^p$ is the linear coefficient, $\epsilon \in R^n$ is the random noise with $E(\epsilon) = 0$.

   (a) Find $\hat{\beta}$: the ordinary least square estimator of β.

   (b) Prove $E(X\hat{\beta}) = X\beta$.

   (c) Under the assumption that $\epsilon \sim N_p(0, \sigma^2 I)$, that is ϵ is Gaussian noise with covariance matrix $\sigma^2 I$ where I is the $p \times p$ identity matrix, first find the Fisher information matrix, then find the covariance of the maximum likelihood estimator (MLE) of β.

3. (10 pts) Let X be a random variable with probability density function (PDF):

$$f(x) = \frac{e^x}{(1+e^x)^2}, \quad -\infty < x < \infty.$$

Define

$$Y = \frac{1}{1+e^{-X}}.$$

Find the PDF of Y.

4. (10 pts) Consider the model given by $X \sim$ Lonnormal(0,1) and $\log Y = 9 + 3 \log X + \epsilon$, where $\epsilon \sim N(0, 1)$. Derive

$$E\left(\frac{Y}{X}|X\right),$$

the conditional mean of $\frac{Y}{X}$ given X.

5. (10 pts) You are asked to generate random numbers from the following probability density function (PDF):

$$f(x) = 3x^2, \quad 0 < x < 2.$$

Suppose you know how to generate random number U from the standard uniform distribution. Find a function $g(\cdot)$, such that $X = g(U)$ is the random variable with PDF $f(x)$.

6. (10 pts) Write down an algorithm to generate random numbers from *Beta*(2, 3) distribution by Rejection Sampling.

References

1. Let $l(\theta)$ denote the log likelihood function, then the *Fisher Information Marix* is defined as

$$I(\theta) = E[l'(\theta)\, l'(\theta)^T] = -E[l''(\theta)].$$

2. Some Properties of random vector operation:

Let $Y \in R^n$ be a random vector, $A \in R^{m \times n}$ a given matrix, then

$$E(AY) = A\, E(Y)$$
$$Var(AY) = A\, Var(Y)\, A^T$$

3. Table 1.1, 1.2, 1.3

4. Rejection Sampling - Section 6.2.3

TABLE 1.1 Notation and description for common probability distributions of discrete random variables.

Name	Notation and Parameter Space	Density and Sample Space	Mean and Variance
Bernoulli	$X \sim \text{Bernoulli}(p)$ $0 \le p \le 1$	$f(x) = p^x(1-p)^{1-x}$ $x = 0$ or 1	$E\{X\} = p$ $\text{var}\{X\} = p(1-p)$
Binomial	$X \sim \text{Bin}(n, p)$ $0 \le p \le 1$ $n \in \{1, 2, \ldots\}$	$f(x) = \binom{n}{x} p^x(1-p)^{n-x}$ $x = 0, 1, \ldots n$	$E\{X\} = np$ $\text{var}\{X\} = np(1-p)$
Multinomial	$\mathbf{X} \sim \text{Multinomial}(n, \mathbf{p})$ $\mathbf{p} = (p_1, \ldots, p_k)$ $0 \le p_i \le 1$ and $n \in \{1, 2, \ldots\}$ $\sum_{i=1}^k p_i = 1$	$f(\mathbf{x}) = \binom{n}{x_1 \quad \ldots \quad x_k} \prod_{i=1}^k p_i^{x_i}$ $\mathbf{x} = (x_1, \ldots, x_k)$ and $x_i \in \{0, 1, \ldots n\}$ $\sum_{i=1}^k x_i = n$	$E\{\mathbf{X}\} = n\mathbf{p}$ $\text{var}\{X_i\} = np_i(1-p_i)$ $\text{cov}\{X_i, X_j\} = -np_ip_j$
Negative Binomial	$X \sim \text{NegBin}(r, p)$ $0 \le p \le 1$ $r \in \{1, 2, \ldots\}$	$f(x) = \binom{x+r-1}{r-1} p^r(1-p)^x$ $x \in \{0, 1, \ldots\}$	$E\{X\} = r(1-p)/p$ $\text{var}\{X\} = r(1-p)/p^2$
Poisson	$X \sim \text{Poisson}(\lambda)$ $\lambda > 0$	$f(x) = \frac{\lambda^x}{x!} \exp\{-\lambda\}$ $x \in \{0, 1, \ldots\}$	$E\{X\} = \lambda$ $\text{var}\{X\} = \lambda$

TABLE 1.2 Notation and description for some common probability distributions of continuous random variables.

Name	Notation and Parameter Space	Density and Sample Space	Mean and Variance
Beta	$X \sim$Beta(α, β) $\alpha > 0$ and $\beta > 0$	$f(x) = \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$ $0 \le x \le 1$	$E\{X\} = \dfrac{\alpha}{\alpha+\beta}$ $\text{var}\{X\} = \dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Cauchy	$X \sim$Cauchy(α, β) $\alpha \in \Re$ and $\beta > 0$	$f(x) = \dfrac{1}{\pi\beta\left[1+\left(\dfrac{x-\alpha}{\beta}\right)^2\right]}$ $x \in \Re$	$E\{X\}$ is nonexistent $\text{var}\{X\}$ is nonexistent
Chi-square	$X \sim \chi^2_\nu$ $\nu > 0$	$f(x) =$ Gamma$(\nu/2, 1/2)$ $x > 0$	$E\{X\} = \nu$ $\text{var}\{X\} = 2\nu$
Dirichlet	$\mathbf{X} \sim$Dirichlet$(\boldsymbol{\alpha})$ $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$ $\alpha_i > 0$ $\alpha_0 = \sum_{i=1}^k \alpha_i$	$f(\mathbf{x}) = \dfrac{\Gamma(\alpha_0)\prod_{i=1}^k x_i^{\alpha_i-1}}{\prod_{i=1}^k \Gamma(\alpha_i)}$ $\mathbf{x} = (x_1, \ldots, x_k)$ and $0 \le x_i \le 1$ $\sum_{i=1}^k x_i = 1$	$E\{\mathbf{X}\} = \boldsymbol{\alpha}/\alpha_0$ $\text{var}\{X_i\} = \dfrac{\alpha_i(\alpha_0-\alpha_i)}{\alpha_0^2(\alpha_0+1)}$ $\text{cov}\{X_i, X_j\} = \dfrac{-\alpha_i\alpha_j}{\alpha_0^2(\alpha_0+1)}$
Exponential	$X \sim$Exp(λ) $\lambda > 0$	$f(x) = \lambda \exp\{-\lambda x\}$ $x > 0$	$E\{X\} = 1/\lambda$ $\text{var}\{X\} = 1/\lambda^2$
Gamma	$X \sim$Gamma(r, λ) $\lambda > 0$ and $r > 0$	$f(x) = \dfrac{\lambda^r x^{r-1}}{\Gamma(r)} \exp\{-\lambda x\}$ $x > 0$	$E\{X\} = r/\lambda$ $\text{var}\{X\} = r/\lambda^2$

TABLE 1.3 Notation and description for more common probability distributions of continuous random variables.

Name	Notation and Parameter Space	Density and Sample Space	Mean and Variance
Lognormal	$X \sim$Lognormal(μ, σ^2) $\mu \in \Re$ and $\sigma > 0$	$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\left(\frac{\log\{x\} - \mu}{\sigma}\right)^2\right\}$ $x \in \Re$	$E\{X\} = \exp\{\mu + \sigma^2/2\}$ $\text{var}\{X\} = \exp\{2\mu + 2\sigma^2\} - \exp\{2\mu + \sigma^2\}$
Multivariate Normal	$\mathbf{X} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k) \in \Re^k$ $\boldsymbol{\Sigma}$ positive definite	$f(\mathbf{x}) = \frac{\exp\{-(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})/2\}}{(2\pi)^{k/2}\lvert\boldsymbol{\Sigma}\rvert^{1/2}}$ $\mathbf{x} = (x_1, \ldots, x_k) \in \Re^k$	$E\{\mathbf{X}\} = \boldsymbol{\mu}$ $\text{var}\{\mathbf{X}\} = \boldsymbol{\Sigma}$
Normal	$X \sim$N(μ, σ^2) $\mu \in \Re$ and $\sigma > 0$	$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2\right\}$ $x \in \Re$	$E\{X\} = \mu$ $\text{var}\{X\} = \sigma^2$
Student's t	$X \sim t_\nu$ $\nu > 0$	$f(x) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\pi\nu}} \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2}$ $x \in \Re$	$E\{X\} = 0$ if $\nu > 1$ $\text{var}\{X\} = \frac{\nu}{\nu+2}$ if $\nu > 2$
Uniform	$X \sim$Unif(a, b) $a, b \in \Re$ and $a < b$	$f(x) = \frac{1}{b-a}$ $x \in [a, b]$	$E\{X\} = (a+b)/2$ $\text{var}\{X\} = (b-a)^2/12$
Weibull	$X \sim$Weibull(a, b) $a > 0$ and $b > 0$	$f(x) = abx^{b-1} \exp\{-ax^b\}$ $x > 0$	$E\{X\} = \frac{\Gamma(1 + 1/b)}{a^{1/b}}$ $\text{var}\{X\} = \frac{\Gamma(1 + 2/b) - \Gamma(1 + 1/b)^2}{a^{2/b}}$

Although this approach is not exact, we include it in this section because the degree of approximation is deterministic and can be reduced to any desired level by increasing m sufficiently. Compared to the alternatives, this simulation method is not appealing because it requires a complete approximation to F regardless of the desired sample size, it does not generalize to multiple dimensions, and it is less efficient than other approaches.

6.2.3 Rejection Sampling

If $f(x)$ can be calculated, at least up to a proportionality constant, then we can use *rejection sampling* to obtain a random draw from exactly the target distribution. This strategy relies on sampling candidates from an easier distribution and then correcting the sampling probability through random rejection of some candidates.

Let g denote another density from which we know how to sample and for which we can easily calculate $g(x)$. Let $e(\cdot)$ denote an *envelope*, having the property $e(x) = g(x)/\alpha \geq f(x)$ for all x for which $f(x) > 0$ for a given constant $\alpha \leq 1$. Rejection sampling proceeds as follows:

1. Sample $Y \sim g$.
2. Sample $U \sim \text{Unif}(0, 1)$.
3. Reject Y if $U > f(Y)/e(Y)$. In this case, do not record the value of Y as an element in the target random sample. Instead, return to step 1.
4. Otherwise, keep the value of Y. Set $X = Y$, and consider X to be an element of the target random sample. Return to step 1 until you have accumulated a sample of the desired size.

The draws kept using this algorithm constitute an i.i.d. sample from the target density f; there is no approximation involved. To see this, note that the probability that a kept draw falls at or below a value y is

$$\begin{aligned} P[X \leq y] &= P\left[Y \leq y \,\middle|\, U \leq \frac{f(Y)}{e(Y)}\right] \\ &= P\left[Y \leq y \text{ and } U \leq \frac{f(Y)}{e(Y)}\right] \Big/ P\left[U \leq \frac{f(Y)}{e(Y)}\right] \\ &= \int_{-\infty}^{y} \int_{0}^{f(z)/e(z)} du\ g(z)\, dz \Big/ \int_{-\infty}^{\infty} \int_{0}^{f(z)/e(z)} du\ g(z)\, dz \qquad (6.4) \\ &= \int_{-\infty}^{y} f(z)\, dz, \qquad (6.5) \end{aligned}$$

which is the desired probability. Thus, the sampling distribution is exact, and α can be interpreted as the expected proportion of candidates that are accepted. Hence α is a measure of the efficiency of the algorithm. We may continue the rejection sampling procedure until it yields exactly the desired number of sampled points, but this requires a random total number of iterations that will depend on the proportion of rejections.

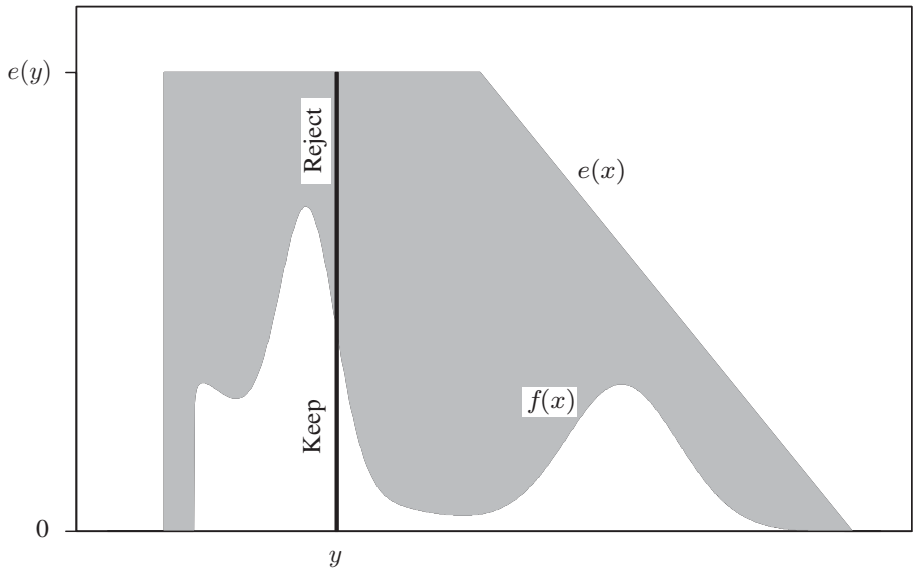

FIGURE 6.1 Illustration of rejection sampling for a target distribution f using a rejection sampling envelope e.

Recall the rejection rule in step 3 for determining the fate of a candidate draw, $Y = y$. Sampling $U \sim \text{Unif}(0, 1)$ and obeying this rule is equivalent to sampling $U|y \sim \text{Unif}(0, e(y))$ and keeping the value y if $U < f(y)$. Consider Figure 6.1. Suppose the value y falls at the point indicated by the vertical bar. Then imagine sampling $U|Y = y$ uniformly along the vertical bar. The rejection rule eliminates this Y draw with probability proportional to the length of the bar above $f(y)$ relative to the overall bar length. Therefore, one can view rejection sampling as sampling uniformly from the two-dimensional region under the curve e and then throwing away any draws falling above f and below e. Since sampling from f is equivalent to sampling uniformly from the two-dimensional region under the curve labeled $f(x)$ and then ignoring the vertical coordinate, rejection sampling provides draws exactly from f.

The shaded region in Figure 6.1 above f and below e indicates the waste. The draw $Y = y$ is very likely to be rejected when $e(y)$ is far larger than $f(y)$. Envelopes that exceed f everywhere by at most a slim margin produce fewer wasted (i.e., rejected) draws and correspond to α values near 1.

Suppose now that the target distribution f is only known up to a proportionality constant c. That is, suppose we are only able to compute easily $q(x) = f(x)/c$, where c is unknown. Such densities arise, for example, in Bayesian inference when f is a posterior distribution known to equal the product of the prior and the likelihood scaled by some normalizing constant. Fortunately, rejection sampling can be applied in such cases. We find an envelope e such that $e(x) \geq q(x)$ for all x for which $q(x) > 0$. A draw $Y = y$ is rejected when $U > q(y)/e(y)$. The sampling probability remains correct because the unknown constant c cancels out in the numerator and denominator of (6.4) when f is replaced by q. The proportion of kept draws is α/c.

Multivariate targets can also be sampled using rejection sampling, provided that a suitable multivariate envelope can be constructed. The rejection sampling algorithm is conceptually unchanged.

To produce an envelope we must know enough about the target to bound it. This may require optimization or a clever approximation to f or q in order to ensure that e can be constructed to exceed the target everywhere. Note that when the target is continuous and log-concave, it is unimodal. If we select x_1 and x_2 on opposite sides of that mode, then the function obtained by connecting the line segments that are tangent to $\log f$ or $\log q$ at x_1 and x_2 yields a piecewise exponential envelope with exponential tails. Deriving this envelope does not require knowing the maximum of the target density; it merely requires checking that x_1 and x_2 lie on opposite sides of it. The adaptive rejection sampling method described in Section 6.2.3.2 exploits this idea to generate good envelopes.

To summarize, good rejection sampling envelopes have three properties: They are easily constructed or confirmed to exceed the target everywhere, they are easy to sample, and they generate few rejected draws.

Example 6.1 (Gamma Deviates) Consider the problem of generating a Gamma(r, 1) random variable when $r \geq 1$. When Y is generated according to the density

$$f(y) = \frac{t(y)^{r-1} t'(y) \exp\{-t(y)\}}{\Gamma(r)} \tag{6.6}$$

for $t(y) = a(1 + by)^3$ for $-1/b < y < \infty, a = r - \frac{1}{3}$, and $b = 1/\sqrt{9a}$, then $X = t(Y)$ will have a Gamma(r, 1) distribution [443]. Marsaglia and Tsang describe how to use this fact in a rejection sampling framework [444]. Adopt (6.6) as the target distribution because transforming draws from f gives the desired gamma draws.

Simplifying f and ignoring the normalizing constant, we wish to generate from the density that is proportional to $q(y) = \exp\{a \log\{t(y)/a\} - t(y) + a\}$. Conveniently, q fits snugly under the function $e(y) = \exp\{-y^2/2\}$, which is the unscaled standard normal density. Therefore, rejection sampling amounts to sampling a standard normal random variable, Z, and a standard uniform random variable, U, then setting $X = t(Z)$ if

$$U \leq \frac{q(Z)}{e(Z)} = \exp\left\{ \frac{Z^2}{2} + a \log\left\{ \frac{t(Z)}{a} \right\} - t(Z) + a \right\} \tag{6.7}$$

and $t(Z) > 0$. Otherwise, the draw is rejected and the process begun anew. An accepted draw has density Gamma(r, 1). Draws from Gamma(r, 1) can be rescaled to obtain draws from Gamma(r, λ).

In a simulation when $r = 4$, over 99% of candidate draws are accepted and a plot of $e(y)$ and $q(y)$ against y shows that the two curves are nearly superimposed. Even in the worst case ($r = 1$), the envelope is excellent, with less than 5% waste.□

Example 6.2 (Sampling a Bayesian Posterior) Suppose 10 independent observations (8, 3, 4, 3, 1, 7, 2, 6, 2, 7) are observed from the model $X_i|\lambda \sim \text{Poisson}(\lambda)$. A lognormal prior distribution for λ is assumed: $\log \lambda \sim N(\log 4, 0.5^2)$. Denote the likelihood as $L(\lambda|\mathbf{x})$ and the prior as $f(\lambda)$. We know that $\hat{\lambda} = \bar{x} = 4.3$ maximizes $L(\lambda|\mathbf{x})$ with respect to λ; therefore the unnormalized posterior, $q(\lambda|\mathbf{x}) = f(\lambda)L(\lambda|\mathbf{x})$