Stats 210A, Fall 2023
Homework 4

Due date: Wednesday, Sep. 27

You may disregard measure-theoretic niceties about conditioning on measure-zero sets, almost-sure equality vs. actual equality, "all functions" vs. "all measurable functions," etc. (unless the problem is explicitly asking about such issues).

1. Bayesian law of large numbers

(a) Let $p(x)$ and $q(x)$ denote two strictly positive probability densities with respect to a common dominating measure μ. The *Kullback–Leibler divergence* between p and q is defined as

$$D(p\|q) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}\mu(x).$$

Show that $D(p\|q) \geq 0$, with equality only in the case that $p(X) = q(X)$ almost surely
Hint: recall that $\log(1 + x) \leq x$ for all $x > -1$.

(b) Consider a dominated likelihood model $\mathcal{P} = \{p_\theta(x) : \theta \in \Theta\}$, where the parameter space Θ is a finite set, and the densities are strictly positive on $\mathcal{X}$. Let λ denote a prior density w.r.t. the counting measure on Θ, and consider the Bayes posterior after observing a sample $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p_{\theta_0}(x)$ for some *fixed* value θ_0 (that is, we are doing a *frequentist* analysis of the *Bayesian* posterior distribution). Assume that all the densities are distinct; that is, $p_{\theta_1}(X) = p_{\theta_2}(X)$ almost surely if and only if $\theta_1 = \theta_2$.

If the prior λ puts positive mass on all values in Θ, show that as $n \to \infty$, the posterior density eventually concentrates nearly all its mass on the true value θ_0. That is,

$$\mathbb{P}_{\theta_0} \left[\lambda(\theta_0 \mid X_1, \ldots, X_n) \geq 1 - \varepsilon \right] \to 1, \quad \text{for all } \varepsilon > 0.$$

(**Hint:** use the law of large numbers).

Moral: At least for a finite parameter space, the Bayes estimator always converges to the right answer as long as we put positive mass on the right answer. This result can be generalized with more effort to continuous parameter spaces under some regularity conditions on the likelihood function, similar to the types of conditions we will use to guarantee the MLE is consistent.

The requirement that the prior density should be nonzero everywhere is sometimes called Cromwell's Rule, after Oliver Cromwell's famous plea to the Church of Scotland: "I beseech you, in the bowels of Christ, think it possible that you may be mistaken."

2. Fisher information for location and scale families

Consider a scale family

$$p_\theta(x) = \frac{1}{\theta} p_0\left(\frac{x}{\theta}\right), \quad \theta > 0.$$

where p_0 is some fixed probability density function with respect to the Lebesgue measure.

(a) Show that the Fisher information of a single observation X is given by

$$J(\theta) = \frac{1}{\theta^2} \int_{-\infty}^{\infty} \left[\frac{u p_0'(u)}{p_0(u)} + 1 \right]^2 p_0(u) du.$$

Try to explain in your own words why it makes sense that the Fisher information should be proportional to θ^{-2} (the verbal explanation will be graded leniently).

(b) If we instead parameterize the model using $\zeta = \log \theta$, show that the Fisher information $J(\zeta)$ of a single observation X does not depend on ζ. Explain in your own words why this makes sense.

3. Ridge regression

Consider the *Gaussian linear model* where

$$y_i = x_i'\beta + \varepsilon_i, \quad \text{with } \varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2) \text{ for } i = 1, \ldots n,$$

where $\beta \in \mathbb{R}^d$ is unknown, and the covariate vectors $x_i \in \mathbb{R}^d$ are fixed and known. Assume the error variance $\sigma^2 > 0$ is also known. We observe the response vector $y \in \mathbb{R}^n$.

(a) Assume that $d \leq n$, and the design matrix $\mathbf{X}$ (the $n \times d$ matrix whose ith row is x_i') has full column rank. Show that the OLS estimator $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y$ is the UMVU estimator of β.

Note: Remember that the design matrix $\mathbf{X}$ is not data in the same sense y is; it is more like a known parameter.

(b) Now consider Bayesian estimation with the prior $\beta \sim N(\mu, \tau^2 I_d)$. Under the same prior as in part (b), find the posterior distribution of β. Does it matter whether $d > n$, or whether $\mathbf{X}$ has full column rank?

(c) Suppose that $\mathbf{X}\gamma = 0$ for some nonzero $\gamma \in \mathbb{R}^d$. Show that no unbiased estimator exists for $g(\beta) = \beta'\gamma$. What is the posterior distribution for $g(\beta)$?

4. Other loss functions

Assume for each problem below that there exists an estimator with finite Bayes risk.

(a) Consider a Bayesian model with a discrete parameter θ. What is the Bayes estimator for the loss $L(\theta, d) = 1\{\theta \neq d\}$?

(b) Next consider a Bayesian model with a single real parameter θ, and assume that the posterior distribution of θ given $X = x$ is absolutely continuous (with respect to the Lebesgue measure) for all x. What is the Bayes estimator for the *absolute error loss* $L(\theta, d) = |\theta - d|$?

(c) Under the same assumptions as part (b), what loss function $L_\gamma(\theta, d)$ would give the posterior γ quantile as its Bayes estimator; that is, the estimator $\delta_\gamma(X)$ has $\mathbb{P}(\theta < \delta_\gamma(X) \mid X) = \gamma$.

5. Exponential-exponential model

Consider a Bayesian model with prior distribution $\lambda(\theta) = e^{-\theta} 1\{\theta > 0\}$ for θ (the standard exponential distribution) and whose likelihood is the exponential location family:

$$p_\theta(x) = e^{\theta - x} 1\{x > \theta\},$$

where we observe a sample $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p_\theta(x)$ given θ.

(a) Calculate the posterior distribution for θ for $n > 1$.

(b) For $n = 1$, calculate the posterior distribution and the Bayes estimator under squared error loss.

(c) Still for $n = 1$, calculate the MSE for the Bayes estimator and the UMVU estimator as a function of θ. Plot the risk function for $\theta \in [0, 5]$. For what values of θ does the Bayes estimator perform better?

(d) Still for $n = 1$, calculate the Bayes risk for the Bayes estimator, and for the UMVU estimator $X_1 - 1$, using squared error loss.

Moral: The Bayes estimator tends to have better risk in places where the prior is large, sometimes at the cost of performing very poorly where the prior puts very little mass.