Stats 210A, Fall 2021
Homework 4

Due date: Wednesday, Sep. 29

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 \( \mu \). The Kullback–Leibler divergence between \( p \) and \( q \) is defined as
   \[
   D(p∥q) = \int_X p(x) \log \frac{p(x)}{q(x)} \, d\mu(x).
   \]
   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 \( P = \{ p_\theta(x) : \theta \in \Omega \} \), where the parameter space \( \Omega \) is a finite set, and the densities are strictly positive on \( X \). Let \( \lambda \) denote a prior density w.r.t. the counting measure on \( \Omega \), and consider the Bayes posterior after observing a sample \( X_1, \ldots, X_n \) i.i.d. \( p_{\theta_0}(x) \) for some fixed value \( \theta_0 \) (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 \( \lambda \) puts positive mass on all values in \( \Omega \), show that as \( n \to \infty \), the posterior density eventually concentrates nearly all its mass on the true value \( \theta_0 \). That is,
   \[
   \mathbb{P}_{\theta_0} [\lambda(\theta_0 \mid X_1, \ldots, X_n) \geq 1 - \varepsilon] \to 1, \quad \text{for all } \varepsilon > 0.
   \]
   
   **(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. Mean parameterization of exponential families

   Consider an \( s \)-parameter exponential family \( P \) with densities wrt some measure on \( X \):
   \[
   p_\eta(x) = e^{\eta^T(x) - A(\eta)}h(x).
   \]

   Assume the parameter space \( \Xi \) is an open subset of \( \mathbb{R}^s \), and that \( T_1(X), \ldots, T_s(X) \) are linearly independent in the sense that there are no \( \alpha \in \mathbb{R} \) and \( \beta \in \mathbb{R}^s \) such that \( \alpha + \beta^T T(X) = 0 \) almost surely.

   Define the function \( \psi(\eta) = \mathbb{E}_\eta T(X) \) and consider parameterizing the family instead according to \( \theta = \psi(\eta) \), called the mean parameter.
(a) Show that the mapping \( \psi : \mathbb{R}^s \to \mathbb{R}^s \) is one-to-one: that is, for any \( \eta^{(0)} \neq \eta^{(1)} \), \( \psi(\eta^{(0)}) \neq \psi(\eta^{(1)}) \) (otherwise \( \theta = \psi(\eta) \) would not be a valid parameterization of the entire family \( \mathcal{P} \)).

Hint: Consider varying \( \eta \) on a line segment between \( \eta^{(0)} \) and \( \eta^{(1)} \).

(b) For the remainder of the problem, assume \( s = 1 \). If we parameterize in terms of the mean parameter \( \theta \), we have

\[
p_{\theta}(x) = e^{\eta(\theta) T(x) - B(\theta) h(x)},
\]

where \( \eta(\theta) = \psi^{-1}(\theta) \). Show that the first derivative of \( B \) is given by

\[
\dot{B}(\theta) = \theta / \text{Var}_\theta(T(X)).
\]

Hint: It may be useful to use the identity that if \( y = f(x) \), and \( f \) is invertible, then \( \frac{d(f^{-1})(y)}{dy} = \frac{1}{f'(f^{-1}(y))} \).

(c) Show that the Fisher information for \( \theta \) is

\[
J(\theta) = \text{Var}_\theta(T(X))^{-1}.
\]

(d) Use the above to find the Jeffreys prior for \( \theta \) in the Poisson family with \( \theta > 0 \):

\[
\text{Pois}(\theta) = \frac{\theta^x e^{-\theta}}{x!}, \quad x = 0, 1, \ldots
\]

and in the binomial family with \( \theta \in (0, 1) \):

\[
\text{Binom}(n, \theta) = \theta^n (1 - \theta)^{n-x} \binom{n}{x}, \quad x = 0, 1, \ldots, n.
\]

For each, either give the density of the prior explicitly or give an expression proportional to the prior and show it is improper.

Moral: The mean parameter gives a valid parameterization of the model, and can be more convenient to work with than the natural parameter.

3. Ridge regression

Consider the Gaussian linear model where

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

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 \( X \) (the \( n \times d \) matrix whose \( i \)th row is \( x'_i \)) has full column rank. Show that the OLS estimator \( \hat{\beta} = (X'X)^{-1}X'y \) is the UMVU estimator of \( \beta \).

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

(b) Now consider Bayesian estimation with the prior \( \beta \sim N(\mu, \tau^2 I_d) \). Find the posterior distribution of \( \beta \). Does it matter whether \( d > n \), or whether \( X \) has full column rank?

(c) Suppose that \( X\gamma = 0 \) 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 \( \Theta \). 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 \( \Theta \), and assume that the posterior distribution of \( \Theta \) 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 \( \gamma \) quantile as its Bayes estimator; that is, the estimator \( \delta_\gamma(X) \) has \( P(\Theta < \delta_\gamma(X) | X) = \gamma \).

5. Exponential-exponential model

Consider a Bayesian model with prior distribution \( \lambda(\theta) = e^{-\theta}1\{\theta > 0\} \) for \( \Theta \) (the standard exponential distribution) and whose likelihood is the exponential location family from homework 3:

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

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

(a) Calculate the posterior distribution for \( \theta \) 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 \( \theta \). Plot the risk function for \( \theta \in [0, 5] \). For what values of \( \theta \) 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.