Stats 210A, Fall 2020
Homework 6

Due date: Wednesday, Oct. 14

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).

If you need to write code to answer a question, show your code. If you need to include a plot,
 make sure the plot is readable, with appropriate axis labels and a legend if necessary. Points will
 be deducted for very hard-to-read code or plots.

1. Effective degrees of freedom

We can write a standard Gaussian sequence model in the form

\[ Y_i = \mu_i + \varepsilon_i, \quad \varepsilon_i \sim_{i.i.d.} N(0, \sigma^2), \quad i = 1, \ldots, n \]

with \( \mu \in \mathbb{R}^n \) and \( \sigma^2 > 0 \) possibly unknown. If we estimate \( \mu \) by some estimator \( \hat{\mu}(Y) \), we
can compute the residual sum of squares (RSS):

\[ \text{RSS}(\hat{\mu}, Y) = \| \hat{\mu}(Y) - Y \|^2 = \sum_{i=1}^{n} (\hat{\mu}_i(Y) - Y_i)^2. \]

If we were to observe the same signal with independent noise \( Y^* = \mu + \varepsilon^* \), the expected
prediction error (EPE) is defined as

\[ \text{EPE}(\mu, \hat{\mu}) = \mathbb{E}_\mu [\| \hat{\mu}(Y) - Y^* \|^2] = \mathbb{E}_\mu [\| \hat{\mu}(Y) - \mu \|^2] + n\sigma^2. \]

Because \( \hat{\mu} \) is typically chosen to make RSS small for the observed data \( Y \) (i.e., to fit \( Y \) well),
the RSS is usually an optimistic estimator of the EPE, especially if \( \hat{\mu} \) tends to overfit. To
quantify how much \( \hat{\mu} \) overfits, we can define the effective degrees of freedom (or simply the
degrees of freedom) of \( \hat{\mu} \) as

\[ \text{DF}(\mu, \hat{\mu}) = \frac{1}{2\sigma^2} \mathbb{E}[\text{EPE} - \text{RSS}], \]

which uses optimism as a proxy for overfitting.

For the following questions assume we also have a predictor matrix \( X \in \mathbb{R}^{n \times d} \), which is
simply a matrix of fixed real numbers. Suppose that \( d \leq n \) and \( X \) has full column rank.

(a) Show that if \( \hat{\mu} \) is differentiable with \( \mathbb{E}_\mu \| D\hat{\mu}(Y) \|_F < \infty \) then

\[ \sum_{i=1}^{n} \frac{\partial \hat{\mu}_i(Y)}{\partial Y_i} \]

is an unbiased estimator of the DF. (Recall \( D\hat{\mu}(Y) \) is the Jacobian matrix from class).

(b) Suppose \( \hat{\mu} = X\hat{\beta} \), where \( \hat{\beta} \) is the ordinary least squares estimator (i.e., chosen to
minimize the RSS). Show that the DF is \( d \). (This confirms that DF generalizes the
intuitive notion of degrees of freedom as “the number of free variables”).
(c) Suppose $\hat{\mu} = X\hat{\beta}$, where $\hat{\beta}$ minimizes the penalized least squares criterion:

$$\hat{\beta} = \arg \min_{\beta} \|Y - X\beta\|^2_2 + \rho\|\beta\|^2_2,$$

for some $\rho \geq 0$. Show that the DF is $\sum_{j=1}^{d} \frac{\lambda_j}{\lambda_j + \lambda_1}$, where $\lambda_1 \geq \cdots \geq \lambda_d > 0$ are the eigenvalues of $X'X$ (counted with multiplicity) (Hint: use the singular value decomposition of $X$).

2. Soft thresholding

Consider the soft thresholding operator with parameter $\lambda \geq 0$, defined as

$$\eta_\lambda(x) = \begin{cases} 
  x - \lambda & x > \lambda \\
  0 & |x| \leq \lambda \\
  x + \lambda & x < -\lambda 
\end{cases}$$

Note that, although we didn’t prove it in class, Stein’s lemma applies for continuous functions $h(x)$ which are differentiable except on a measure zero set; you can apply it here without worrying.

Assume $X \sim N_d(\theta, I_d)$ for $\theta \in \mathbb{R}^d$, which we will estimate via $\delta_\lambda(X) = (\eta_\lambda(X_1), \ldots, \eta_\lambda(X_d))$. Soft thresholding is sometimes used when we expect sparsity: a small number of relatively large $\theta_i$ values. $\lambda$ here is called a tuning parameter since it determines what version of the estimator we use, but doesn’t have an obvious statistical interpretation.

(a) Show that $|\{i : |X_i| > \lambda\}|$ is an unbiased estimator of the degrees of freedom of $\delta_\lambda$ (so, in a sense, the DF is the expected number of “free variables”).

(b) Show that

$$d + \sum_i \min(X_i^2, \lambda^2) - 2|\{i : |X_i| \leq \lambda\}|$$

is an unbiased estimator for the MSE of $\delta_\lambda$.

(c) Show that the risk-minimizing value $\lambda^*$ solves

$$\lambda \sum_i P_{\theta_i}(|X_i| > \lambda) = \sum_i \phi(\lambda - \theta_i) + \phi(\lambda + \theta_i),$$

where $\phi(z) = e^{-z^2/2} \sqrt{2\pi}$ is the standard normal density.

(d) Consider a problem with $\theta_1 = \cdots = \theta_{20} = 10$ and $\theta_{21} = \cdots = \theta_{500} = 0$. Compute $\lambda^*$ numerically. Then simulate a vector $X$ from the model and use it to automatically tune the value of $\lambda$ by minimizing SURE. Call the automatically tuned value $\hat{\lambda}(X)$ and report both $\lambda^*$ and $\hat{\lambda}(X)$. Finally plot the true MSE of $\delta_\lambda$ along with its SURE estimate against $\lambda$ for a reasonable range of $\lambda$ values. Add a horizontal line for the risk of the UMVU estimator.

(e) Compute and report the squared error loss $\|\delta(X) - \theta\|^2$ for the following four estimators:

(i) the UMVU estimator $\delta_0(X) = X$,

(ii) the optimally tuned soft-thresholding estimator $\delta_{\lambda^*}(X)$,

(iii) the automatically tuned soft-thresholding estimator $\delta_{\hat{\lambda}(X)}(X)$, and

(iv) the James-Stein estimator.

You do not need to compute the MSE. Intuitively, what do you think accounts for the good performance of soft-thresholding in this example?

3. Mean estimation
(a) Suppose $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N_d(\theta, I_d)$ and consider estimating $\theta \in \mathbb{R}^d$. Show that $\overline{X} = \frac{1}{n} \sum_i X_i$ is the minimax estimator of $\theta$ under squared error loss.

**Hint:** Find a least favorable sequence of priors.

(b) Suppose $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$ where $P$ is any distribution over the real numbers such that $\text{Var}_P(X) \leq 1$. Show that $\overline{X} = \frac{1}{n} \sum_i X_i$ is minimax for estimating $\theta(P) = \mathbb{E}_P X$ under the squared error loss.

**Hint:** Try to relate this problem to the Gaussian problem with $d = 1$.

(c) Assume $X \sim N(\theta, 1)$ with the constraint that $|\theta| \leq 1$. Show that the minimax estimator for squared error loss is $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$.

Plot its risk function.

**Hint:** Plot the risk function first. For this problem if you need to show that a function is maximized or minimized somewhere, you may do it numerically or by inspecting a graph if it is obvious enough.

### 4. Binomial minimax estimator

Consider estimating $\theta$ under squared error loss, for the model $X \sim \text{Binom}(n, \theta)$.

(a) Plot the risk function of the UMVU estimator and that of the minimax estimator for $n = 1, 10, 100$, and $1000$.

(b) Plot a histogram of the distribution of the UMVU estimator and that of the minimax estimator for $n = 10$ and $n = 1000$, when $\theta = 0.1$ and $\theta = 0.5$. Indicate with a vertical line where the true value of $\theta$ is.

(c) Let $M_n$ denote the interval on which the minimax estimator outperforms the UMVU estimator. Find the endpoints of $M_n$. What happens as $n \to \infty$?

### 5. Upper-bounding $\theta$

(a) Let $X \sim N(\theta, 1)$ for $\theta \in \mathbb{R}$, and consider the loss function

\[ L(\theta, d) = 1\{d < \theta\}; \]

that is, we observe $X$ and try to come up with an upper bound $\delta(x) \in \mathbb{R}$ for $\theta$. Show that the minimax risk is 0 (note you may not be able to find a minimax estimator).

(b) Now, consider a problem with the same loss function but without observing any data. Show the minimax risk (considering both randomized and non-randomized estimators) is 1, but the Bayes risk $r_\Lambda = 0$ for any prior $\Lambda$ (note there may be no estimator $\delta_\Lambda$ that attains the minimum Bayes risk).

**Note:** This problem exhibits a “duality gap” where the lower bounds we can get by trying different priors will always fall short of the minimax risk.

(c) **Optional** (not graded, no extra points): Now consider the same loss function, but now $X \sim N(\theta, \sigma^2)$ and $\sigma^2$ is unknown too. Find the minimax risk.

**Hint:** consider estimators of the form $\delta(X) = c|X|$.  

3