1. Laplace Prior & $\ell^1$-Regularization

Suppose you draw $n$ i.i.d. data points $(x_1, y_1), \ldots, (x_n, y_n)$, where $n$ is a positive integer and the true relationship is $Y = WX + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. (That is, $Y$ has a linear dependence on $X$, with additive Gaussian noise.) Further suppose that $W$ has a prior distribution with density $f_W(w) = \frac{1}{2\beta} e^{-|w|/\beta}$, $\beta > 0$. (This is known as the Laplace distribution.) Show that finding the MAP estimate of $W$ given the data points $\{(x_i, y_i) : i = 1, \ldots, n\}$ is equivalent to minimizing the cost function

$$J(w) = \sum_{i=1}^{n} (y_i - wx_i)^2 + \lambda |w|$$

(you should determine what $\lambda$ is). This is interpreted as a one-dimensional $\ell^1$-regularized least-squares criterion, also known as LASSO.

**Solution:**

The likelihood for $W$ is

$$\mathcal{L}(w \mid (x_1, y_1), \ldots, (x_n, y_n)) \propto \mathcal{L}( (x_1, y_1), \ldots, (x_n, y_n) \mid W = w) f_W(w)$$

(technically, the expression on the right should be divided by the likelihood of the data, but this has no dependence on $w$, so we omit the denominator for simplicity)

$$= \prod_{i=1}^{n} \mathcal{L}(x_i, y_i \mid W = w) f_W(w)$$

(the data points are conditionally independent given $W$)

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y_i - wx_i)^2}{2\sigma^2}} \cdot \frac{1}{2\beta} e^{-|w|/\beta}$$

(here we say that the likelihood of $(x_i, y_i)$ given $W$ is the density of $\varepsilon_i$, which is $\mathcal{N}(0, \sigma^2)$, evaluated at $y_i - wx_i$)
(again, we throw out constant factors that do not depend on the data points or \( w \)).

We wish to maximize this expression \( w.r.t. \ w \), but we will find it more convenient to take the log-likelihood instead.

\[
\ell(w \mid (x_1, y_1), \ldots, (x_n, y_n)) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - wx_i)^2 - \frac{1}{\beta} |w|.
\]

Since we want to maximize the log-likelihood, this is equivalent to \( \text{minimizing} \) the cost function

\[
J(w) = \sum_{i=1}^{n} (y_i - wx_i)^2 + \lambda |w|
\]

where \( \lambda = \frac{2\sigma^2}{\beta} \).

2. Flipping Coins and Hypothesizing

You flip a coin until you see heads. Let

\[
X = \begin{cases} 
1 & \text{if the bias of the coin is } q > p, \\
0 & \text{if the bias of the coin is } p.
\end{cases}
\]

Find a decision rule \( \hat{X}(Y) \) that maximizes \( P[\hat{X} = 1 \mid X = 1] \) subject to \( P[\hat{X} = 1 \mid X = 0] \leq \beta \) for \( \beta \in [0,1] \). Remember to calculate the randomization constant \( \gamma \).

**Solution:**

Let \( Y \) be the number of flips until we see heads. Write the likelihood ratio.

\[
L(y) = \frac{P[Y = y \mid X = 1]}{P[Y = y \mid X = 0]} = \frac{(1 - q)^{y-1} q}{(1 - p)^{y-1} p},
\]

which is strictly decreasing in \( y \) since \( q > p \). Hence, the hypothesis testing rule is of the form \( \hat{X} = 1 \) if \( Y < \alpha \) for some \( \alpha \). Observe that

\[
P[Y < \alpha \mid X = 0] = \sum_{y=1}^{\alpha-1} p(1 - p)^{y-1} = 1 - (1 - p)^{\alpha-1}.
\]

Therefore, we should choose \( \alpha \) such that \( 1 - (1 - p)^{\alpha-1} \leq \beta \), i.e.

\[
\alpha \leq 1 + \frac{\log(1 - \beta)}{\log(1 - p)}.
\]

Therefore, take \( \alpha = [1 + \log(1 - \beta)/\log(1 - p)] \). For the randomization, let \( P[\hat{X} = 1 \mid Y = \alpha] = \gamma \). The probability of false detection is

\[
P[\hat{X} = 1 \mid X = 0] = P[Y < \alpha \mid X = 0] + \gamma P[Y = \alpha \mid X = 0]
\]

\[
= 1 - (1 - p)^{\alpha-1} + \gamma p(1 - p)^{\alpha-1} \leq \beta.
\]
so we take
\[ \gamma = \frac{\beta - 1 + (1 - p)^{a - 1}}{p(1 - p)^{a - 1}}. \]

Hence, for the values of \( \alpha \) and \( \gamma \) described above,
\[ \hat{X} = \begin{cases} 1, & Y < \alpha, \\ Z, & Y = \alpha, \\ 0, & Y > \alpha, \end{cases} \]

where \( Z \) is 1 with probability \( \gamma \) and 0 otherwise.

3. BSC Hypothesis Testing

Consider a BSC with some error probability \( \epsilon \in [0.1, 0.5) \). Given \( n \) inputs and outputs \((x_i, y_i)\) of the BSC, solve a hypothesis problem to detect that \( \epsilon > 0.1 \) with a probability of false alarm at most equal to 0.05. Assume that \( n \) is very large and use the CLT.

*Hint*: The null hypothesis is \( \epsilon = 0.1 \). The alternate hypothesis is \( \epsilon > 0.1 \), which is a composite hypothesis (this means that under the alternate hypothesis, the probability distribution of the observation is not completely determined; compare this to a simple hypothesis such as \( \epsilon = 0.3 \), which does completely determine the probability distribution of the observation). The Neyman-Pearson Lemma we learned in class applies for the case of a simple null hypothesis and a simple alternate hypothesis, so it does not directly apply here.

To fix this, fix some specific \( \epsilon' > 0.1 \) and use the Neyman-Pearson Lemma to find the optimal hypothesis test for the hypotheses \( \epsilon = 0.1 \) vs. \( \epsilon = \epsilon' \). Then, argue that the optimal decision rule does not depend on the specific choice of \( \epsilon' \); thus, the decision rule you derive will be simultaneously optimal for testing \( \epsilon = 0.1 \) vs. \( \epsilon = \epsilon' \) for all \( \epsilon' > 0.1 \).

**Solution:**

We observe \( x_1, \ldots, x_n \) and \( y_1, \ldots, y_n \). Let \( x \) and \( y \) be the vectors of these observations. The likelihood is
\[ P(X = x, Y = y \mid \epsilon) = P(X = x \mid \epsilon) \cdot P(Y = y \mid X = x, \epsilon) \]

We can ignore \( P(X = x \mid \epsilon) \) since in the final likelihood ratio, it’ll cancel out in the numerator and denominator.

\[ P(Y = y \mid X = x, \epsilon) = \epsilon \sum_{i=1}^{n} 1\{y_i \neq x_i\} (1 - \epsilon)^{\sum_{i=1}^{n} 1\{y_i = x_i\}} \propto \left( \frac{\epsilon}{1 - \epsilon} \right) \sum_{i=1}^{n} 1\{y_i \neq x_i\} \]

What matters for estimating \( \epsilon \) is \( t := \sum_{i=1}^{n} 1\{x_i \neq y_i\} \). Therefore we can rewrite the likelihoods as follows. Under the null hypothesis, the likelihood is
\[ P(Y = y \mid X = x, \epsilon = 0.1) \propto \left( \frac{0.1}{0.9} \right)^t = \left( \frac{1}{9} \right)^t. \]
Fix some $\epsilon' > 0.1$; under the alternate hypothesis $\epsilon = \epsilon'$, then the likelihood is

$$P(Y = y \mid X = x, \epsilon = \epsilon') \propto \left( \frac{\epsilon'}{1 - \epsilon'} \right)^t.$$

The likelihood ratio is therefore

$$L(t) = \left( \frac{9\epsilon'}{1 - \epsilon'} \right)^t.$$

The likelihood ratio is a non-decreasing function of $T := \sum_{i=1}^{n} 1\{X_i \neq Y_i\}$, so a threshold on the likelihood ratio corresponds to a threshold on $T$. According to the Neyman-Pearson Lemma, the optimal decision rule is to declare the alternate hypothesis when $T > \lambda(\epsilon')$ where $\lambda(\epsilon')$ is the threshold that is determined by setting the PFA exactly equal to the constraint, i.e.,

$$PFA = P(T > \lambda(\epsilon') \mid \epsilon = 0.1) = 0.05.$$

Since $n$ is very large, $T = \sum_{i=1}^{n} 1\{X_i \neq Y_i\}$ is approximately a normal random variable. Note that without the approximation $T$ is a binomial since input-output pairs of the channel are independent. Now we calculate the following.

$$P(X_1 \neq Y_1) = P(Y_1 = 1 \mid X_1 = 0)P(X_1 = 0) + P(Y_1 = 0 \mid X_1 = 1)P(X_1 = 1) = \epsilon$$

Thus, $T \sim \mathcal{N}(n\epsilon, n\epsilon(1 - \epsilon))$ and

$$P(\mathcal{N}(0.1n, 0.09n) > \lambda(\epsilon')) = Q \left( \frac{\lambda(\epsilon') - 0.1n}{\sqrt{0.09n}} \right) = Q(1.67) = 0.05.$$

Thus, $\lambda(\epsilon') = 0.1n + 1.67\sqrt{0.09n}$. This does not depend on the choice of $\epsilon'$, so the decision rule is the same for all $\epsilon' > 0.1$ and we are done.

4. Exam Difficulties

The difficulty of an EECS 126 exam, $\Theta$, is uniformly distributed on $[0, 100]$, and Alice gets a score $X$ that is uniformly distributed on $[0, \Theta]$. Alice gets her score back and wants to estimate the difficulty of the exam.

(a) What is the LLSE for $\Theta$?

(b) What is the MAP of $\Theta$?

Solution:

(a) We need $E[\Theta]$, var $X$, and cov($X, \Theta$). First, $E[\Theta] = 50$. By the **Law of Total Variance**, 

$$\text{var } X = E[\text{var } (X \mid \Theta)] + \text{var } E[X \mid \Theta].$$

The first term can be found as follows.

$$\text{var } (X \mid \Theta) = \frac{\Theta^2}{12} \implies E[\text{var } (X \mid \Theta)] = \int_{0}^{100} \frac{\theta^2}{12} \cdot \frac{1}{100} \, d\theta = \frac{10000}{36}.$$  

4
By noting \( E[X | \Theta] = \Theta/2 \), the second term is
\[
\frac{1}{4} \left( \frac{10000}{12} \right) = \frac{10000}{48}.
\]
Thus,
\[
\text{var} X = \frac{70000}{144}.
\]

Now, the covariance can be found as follows.

\[\text{cov}(X, \Theta) = E[\Theta X] - E[\Theta] E[X]\]

We found \( E[\Theta] \) above, and \( E[X] = E[E[X | \Theta]] = E[\Theta/2] = 25 \). Also,
\[
E[\Theta X] = E[E[\Theta X | \Theta]] = E\left[ \frac{\Theta^2}{2} \right] = \frac{\text{var} \Theta + E[\Theta]^2}{2} = \frac{10000/12 + 2500}{2}
\]
\[
= \frac{5000}{3}.
\]
Thus,
\[
\text{cov} (X, \Theta) = \frac{1250}{3}.
\]

Then, the LLSE of \( \Theta \) is
\[
L[\Theta | X] = E[\Theta] + \frac{\text{cov}(X, \Theta)}{\text{var} X} (X - E[X]) = 50 + \frac{6}{7} (X - 25).
\]

(b) Given \( X, X \leq \Theta \leq 100 \). In order to maximize
\[
f_{X|\Theta}(x | \theta) = \frac{1}{\theta},
\]
one should choose \( \hat{\Theta} = X \).

5. **Photodetector LLSE**

Consider a photodetector in an optical communications system that counts the number of photons arriving during a certain interval. A user conveys information by switching a photon transmitter on or off. Assume that the probability of the transmitter being on is \( p \). If the transmitter is on, the number of photons transmitted over the interval of interest is a Poisson random variable \( \Theta \) with mean \( \lambda \), and if it is off, the number of photons transmitted is 0. Unfortunately, regardless of whether or not the transmitter is on or off, photons may be detected due to “shot noise”. The number \( N \) of detected shot noise photons is a Poisson random variable \( N \) with mean \( \mu \), independent of the transmitted photons. Let \( T \) be the number of transmitted photons and \( D \) be the number of detected photons. Find \( L[T | D] \).

**Solution:**
\[
L[T | D] = E[T] + \frac{\text{cov}(T, D)}{\text{var} D} (D - E[D])
\]
We find each of these terms separately. We can see by the law of total expectation that $E[T] = p\lambda$. Now, we have:

$$\text{cov}(T, D) = E[(T - E[T])(D - E[D])]$$

$$= E[(T - E[T])(T - E[T]) + N - E[N])]$$

$$= E[(T - E[T])^2] + E[(T - E[T])(N - E[N])] = \text{var} T$$

$$= p(\lambda^2 + \lambda) - (p\lambda)^2$$

where the second to last equality follows since $T$ and $N$ are independent. We now find:

$$\text{var} D = \text{var}(T + N) = \text{var} T + \text{var} N = p(\lambda^2 + \lambda) - (p\lambda)^2 + \mu$$

Finally, we have $E[D] = E[T] + E[N] = p\lambda + \mu$. Putting these together, we have the LLSE (no need to simplify the equation).

6. [Bonus] $p$-Value

The bonus question is just for fun. You are not required to submit the bonus question, but do give it a try and write down your progress.

Let us define what the $p$-value of a hypothesis test is. Given an observation $Y$ and a constraint of $\beta$ on the PFA, the Neyman-Pearson rule will either declare that the alternate hypothesis is true or not. The constraint on the PFA controls the trade-off between declaring the alternate hypothesis to be true when it is not (false alarm), and declaring the alternate hypothesis to be true when it is (correct detection). Therefore, for very high values of $\beta$, the hypothesis test will declare that the alternate hypothesis is true, and for very low values of $\beta$, the hypothesis test will declare that the null hypothesis is true.

(Intuitively, the smaller the value of $\beta$, the more conservative the resulting hypothesis test is, i.e., it will be more reluctant to declare that the alternate hypothesis is true.)

The $p$-value of the observation is the smallest value of $\beta$ such that the alternate hypothesis is declared true.

Think about this carefully, and explain why the $p$-value is not the probability that the alternate hypothesis is true.

**Solution:**

First of all, it is not possible to talk about the “probability” that the alternate hypothesis is true in frequentist statistics, because there is no probability distribution on the hypotheses.

The $p$-value asks the following question: *if* the null hypothesis is true, then what is the probability of seeing data which is at least as extreme as the data that you already saw?

Since the $p$-value is interpreted only when the null hypothesis is true, then the $p$-value is *meaningless* if the alternate hypothesis is true. However, if the null hypothesis is true, then the $p$-value is uniformly distributed on $[0, 1]$. (Why is this true? Well, what is the probability of seeing data that is 5% likely? It is
exactly 5%! In other words, if \( P \) is a random variable representing the \( p \)-value, then \( \mathbb{P}(P \leq 0.05) = 0.05 \). What a puzzling concept!

In fact, frequentist procedures (such as hypothesis testing) are often criticized because they are sometimes hard to interpret.