1 Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) is a method of estimating the parameters of a statistical model given observations, by finding the parameters that maximize the likelihood of the observations. Concretely, given observations \( y_1, y_2, \ldots, y_n \) distributed according to \( p_\theta(y_1, y_2, \ldots, y_n) \) (here \( p_\theta \) can be a probability mass function for discrete observations or a density for continuous observations), the likelihood function is defined as \( L(\theta) = p_\theta(y_1, y_2, \ldots, y_n) \) and the MLE is

\[
\hat{\theta}_{\text{MLE}} = \arg \max_{\theta} L(\theta).
\]

We often make the assumption that the observations are \textit{independent and identically distributed} or iid, in which case \( p_\theta(y_1, y_2, \ldots, y_n) = p_\theta(y_1) \cdot p_\theta(y_2) \cdots \cdot p_\theta(y_n) \).

(a) Your friendly TA recommends maximizing the log-likelihood \( \ell(\theta) = \log L(\theta) \) instead of \( L(\theta) \). Why does this yield the same solution \( \hat{\theta}_{\text{MLE}} \)? Why is it easier to solve the optimization problem for \( \ell(\theta) \) in the iid case? Write down both \( L(\theta) \) and \( \ell(\theta) \) for the Gaussian \( f_\theta(y) = \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \) with \( \theta = (\mu, \sigma) \).
(b) What is \( \int p_\theta(y_1, y_2, \ldots, y_n) \, dy_1 \cdots dy_n \)? Can we say anything about \( \int p_\theta(y_1, y_2, \ldots, y_n) \, d\theta \)?

(c) The Poisson distribution is \( f_\lambda(y) = \frac{\lambda^y e^{-\lambda}}{y!} \). Let \( Y_1, Y_2, \ldots, Y_n \) be a set of independent and identically distributed random variables with Poisson distribution with parameter \( \lambda \). Find the joint distribution of \( Y_1, Y_2, \ldots, Y_n \). Find the maximum likelihood estimator of \( \lambda \) as a function of observations \( y_1, y_2, \ldots, y_n \).
Consider sample points $X_1, X_2, \ldots, X_n \in \mathbb{R}^d$ and associated values $y_1, y_2, \ldots, y_n \in \mathbb{R}$, an $n \times d$ design matrix $X = [X_1 \ldots X_n]^\top$ and an $n$-vector $y = [y_1 \ldots y_n]^\top$.

For the sake of simplicity, assume (1) that the sample data have been centered (i.e., each feature has mean 0) and (2) that the sample data have been whitened, meaning a linear transformation of is applied to the original data matrix so that the resulting features have variance 1 and the features are uncorrelated; i.e., $X^\top X = nI$.

For this question, we will not use a fictitious dimension nor a bias term; our linear regression function will output zero for $x = 0$.

Consider linear least-squares regression with regularization in the $\ell_1$-norm, also known as Lasso. The Lasso cost function is

$$J(w) = |Xw - y|^2 + \lambda \|w\|_1$$

where $w \in \mathbb{R}^d$ and $\lambda > 0$ is the regularization parameter. Let $w^* = \arg \min_{w \in \mathbb{R}^d} J(w)$ denote the weights that minimize the cost function.

In the following steps, we will explore the sparsity-promoting property of the $\ell_1$-norm and compare this with the $\ell_2$-norm.

1. We use the notation $X_i$ to denote column $i$ of the design matrix $X$, which represents the $i^{th}$ feature. Write $J(w)$ in the following form for appropriate functions $g$ and $f$.

$$J(w) = g(y) + \sum_{i=1}^d f(X_{i}, w_i, y, \lambda)$$
2. If \( w_i^* > 0 \), what is the value of \( w_i^* \)?

3. If \( w_i^* < 0 \), what is the value of \( w_i^* \)?

4. Considering parts 2 and 3, what is the condition for \( w_i^* \) to be zero?

5. Now consider ridge regression, which uses the \( \ell_2 \) regularization term \( \lambda |w|^2 \). How does this change the function \( f(\cdot) \) from part 1? What is the new condition in which \( w_i^* = 0 \)? How does
it differ from the condition you obtained in part 4?
3 Probabilistic Interpretation of Lasso

Let’s start with the probabilistic interpretation of least squares. Start with labels $y \in \mathbb{R}$, data $x \in \mathbb{R}^d$, and noise $z \sim N(0, \sigma^2)$, where $y = w^T x + z$. Recall from a previous homework that we then have

$$P(y|x, \sigma^2) \sim N(w^T x, \sigma^2)$$

However, maximum likelihood estimates (MLE) can overfit by picking parameters that mirror the training data. To ameliorate this issue, we can assume a Laplace prior on $w_j \sim \text{Laplace}(0, t)$, i.e.

$$P(w_j) = \frac{1}{2t} e^{-|w_j|/t}$$

$$P(w) = \prod_{j=1}^{D} P(w_j) = \left(\frac{1}{2t}\right)^D \cdot e^{-\sum |w_j| / t}$$

Here, we will see that this modification results in a new objective function, called Lasso.

Recall that the MLE objective finds the parameters that maximize the likelihood of the data,

$$w^* = \arg \max_w L(w)$$

$$= \arg \max_w P(Y_1, \ldots, Y_n, w, X_1, \ldots, X_n, \sigma^2)$$

$$= \arg \max_w \prod_{i=1}^{n} P(Y_i|X_i, w, \sigma^2).$$

When working in a Bayesian framework, we instead focus on the posterior distribution of the parameters conditioned on the data, $P(w|Y_1, \ldots, Y_n, X_1, \ldots, X_n, \sigma^2)$. To pick a single model, we can choose the $w$ that is most likely according to the posterior,

$$w^* = \arg \max_w P(w|Y_1, \ldots, Y_n, X_1, \ldots, X_n, \sigma^2)$$

$$= \arg \max_w \frac{P(w, Y_1, \ldots, Y_n|X_1, \ldots, X_n, \sigma^2)}{P(Y_1, \ldots, Y_n|X_1, \ldots, X_n, \sigma^2)}$$

$$= \arg \max_w \frac{P(Y_1, \ldots, Y_n|w, X_1, \ldots, X_n, \sigma^2)P(w)}{P(Y_1, \ldots, Y_n|X_1, \ldots, X_n, \sigma^2)}$$

$$= \arg \max_w \frac{L(w)P(w)}{P(Y_1, \ldots, Y_n)}$$

$$= \arg \max_w L(w)P(w) \quad \text{since } P(Y_1, \ldots, Y_n|X_1, \ldots, X_n, \sigma^2) \text{ does not depend on } w.$$
(a) Write the log-likelihood for this MAP estimate.

(b) We already have the log-likelihood for MAP. Show that MAP—in this case, Gaussian noise with a Laplace prior—is equivalent to minimizing the following. Additionally, identify the constant \( \lambda \). Note that \( ||w||_1 = \sum_{j=1}^{D} |w_j| \).

\[
J(w) = \sum_{i=1}^{n} (y_i - w^T x_i)^2 + \lambda ||w||_1
\]