1 Maximum Likelihood Review

Suppose you are collecting data on the relative rates of different types of twins, and you obtain the following observations:

- there are $m_i$ pairs of identical male twins and $f_i$ pairs of identical female twins
- there are $m_f$ pairs of fraternal male twins and $f_f$ pairs of fraternal female twins
- there are $b$ pairs of fraternal opposite gender twins

To model this data, we choose these distributions and parameters:

- Given that a pair of siblings are twins, they are identical with probability $\theta$, and non-identical with probability $1 - \theta$.
- Given they are identical twins, the twins are both male with probability $p$ and both female with probability $1 - p$.
- Given they are twins and not identical (and thus are fraternal twins), the probability of both male twins is $q^2$, probability of both female twins is $(1 - q)^2$ and probability of opposite gender twins is $2q(1 - q)$.

(a) Write expressions for the likelihood and the log-likelihood of the data as functions of the parameters $\theta$, $p$, and $q$ for the observations $m_i$, $f_i$, $m_f$, $f_f$, $b$.

Likelihood $L(\theta, p, q) =$

Log likelihood $l(\theta, p, q) =$
(b) What are the maximum likelihood estimates for $\theta$, $p$ and $q$? Scratch space is provided to you here, which you may find useful.
2 MAP Estimation Review

Suppose we have a data set of $n$ data points $D = \{x_1, \ldots, x_n\}$, with each point drawn independently from a Gaussian with mean $\mu$ and variance $\sigma^2$.

$$p(x) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

We will place the following prior on $\mu$:

$$\mu \sim N(\mu_0, \sigma_0^2)$$

The MAP estimate of $\mu$ is defined as

$$\mu_{MAP} = \arg \max_{\mu} p(\mu|D)$$

(a) Write an expression for the MAP estimate of $\mu$.

(b) What happens to the MAP estimate as $\sigma_0^2 \to \infty$, and how is this estimate related to the ML estimate? Interpret this result.

(c) What happens to the MAP estimate as $\sigma^2 \to \infty$?
3 Prediction Error of Ridge Regression

(a) Let $A$ be a $d \times n$ matrix and $B$ be a $n \times d$ matrix. For any $\mu > 0$, show that $(AB + \mu I)^{-1}A = A(BA + \mu I)^{-1}$, if $AB + \mu I$ and $BA + \mu I$ are invertible.

(b) Let $X \in \mathbb{R}^{n \times d}$ be $n$ samples of $d$ features, and $y \in \mathbb{R}^n$ be the corresponding $n$ samples of the quantity that you would like to predict with regression. Let

$$\hat{\theta}_\lambda = \arg\min_\theta \|X\theta - y\|^2_2 + \lambda\|\theta\|^2_2,$$

for $\lambda > 0$, be the solution to the ridge regression problem. Using part (a), show that $\hat{\theta}_\lambda = X^\top(XX^\top + \lambda I)^{-1}y$.

(c) Suppose $X$ has the singular value decomposition $U\Sigma V^\top$, where $\Sigma = \text{diag}(s_1, \cdots, s_d)$, $s_i \geq 0$. Show that $\hat{\theta}_\lambda = VD U^\top y$, where $D$ is a diagonal matrix to be determined.
(d) Let \( \hat{y}_\lambda = X \hat{\theta}_\lambda \) be the predictions made by the ridge regressor \( \hat{\theta}_\lambda \). Suppose we have \( y = X \theta_* + z \), where \( \theta_* \in \mathbb{R}^d \) and \( z = \mathcal{N}(0, \sigma^2 I) \in \mathbb{R}^n \) (\( \sigma > 0 \)). Further suppose that \( X \) is an orthogonal matrix, that is, \( X^\top X = I \).

\[ \mathbb{E} \| X (\hat{\theta}_\lambda - \theta_*) \|^2 \]

is the expected squared difference between the predictions made by the ridge regressor \( \hat{y}_\lambda \) and \( X \theta_* \), where the expectation is taken with respect to \( z \) (\( \| \cdot \| \) denotes \( \ell_2 \) norm).

Show that \( \mathbb{E} \| X (\hat{\theta}_\lambda - \theta_*) \|^2 = \frac{1}{(1+\lambda)^2} \left( \lambda^2 \| \theta_* \|^2 + d \sigma^2 \right) \).

(e) What is the \( \lambda^* \) that you should pick to minimize the prediction error you computed in part (e)? Comment on how \( d, \sigma^2 \), and \( \theta_* \) affect the optimal choice of the regularization parameter \( \lambda \).