FINAL PROJECT II: LOGISTIC REGRESSION
ECE565: Estimation, Filtering, and Detection

Nam Nguyen

School of EECE, Oregon State University
December 2023

1 Define the parameter vector θ and the observation vector for this problem.

The parameter vector is $\mathbf{w} = \begin{bmatrix} w_1 & ... & w_d \end{bmatrix}^T \in \mathbb{R}^d$. The observation vector is $\mathbf{x}_i = \begin{bmatrix} x_1 & ... & x_d \end{bmatrix}^T \in \mathbb{R}^d$, for $i = 1, ..., n$.

2 Find the probabilistic model for the observations given the parameters.

The logistic regression model is a standard approach for training a linear classifier with a dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$. Here, $\mathbf{x}_i$ is a d-dimensional feature vector, and $y_i \in \{0, 1\}$ is its label. The observations $(\mathbf{x}_i, y_i)$ for $i = 1, 2, ..., n$ are assumed to be independent and identically distributed $(i.i.d)$. The logistic function models the conditional probability:

$$p(y|\mathbf{x}; \mathbf{w}) = \frac{e^{(\mathbf{w}^T \mathbf{x})y}}{1 + e^{\mathbf{w}^T \mathbf{x}}}, \tag{1}$$

where $\mathbf{w} \in \mathbb{R}^d$ is an unknown parameter vector. Utilizing the $i.i.d$ assumption, the joint probability model of the observations is expressed as:

$$p(\mathbf{X}, \mathbf{y}|\mathbf{w}) = \prod_{i=1}^{n} \frac{e^{(\mathbf{w}^T \mathbf{x}_i)y_i}}{1 + e^{\mathbf{w}^T \mathbf{x}_i}}, \tag{2}$$

where $\mathbf{X} = [\mathbf{x}_1^T, \mathbf{x}_2^T, ..., \mathbf{x}_n^T]$ denotes the collection of all feature vectors, and $\mathbf{y} = [y_1, y_2, ..., y_n]^T$ denotes the collection of all instance labels. Consequently, the log-likelihood function is given by:

$$l(\mathbf{w}) = \sum_{i=1}^{n} \left[y_i(\mathbf{w}^T\mathbf{x}_i) - \log(1 + e^{\mathbf{w}^T\mathbf{x}_i}) \right], \tag{3}$$

Note that l_2-regularization [2] and l_1-regularization [3] are common in LR.

3 Find the Cramer-Rao Lower Bound (CRLB) for w

We use a method similar to [4] to compute the Fisher Information Matrix (FIM) via FIM $= -\mathbb{E}\left[\frac{d^2 l(\mathbf{w})}{d\mathbf{w}\mathbf{w}^T}\right]$, where the second derivative of the log-likelihood function is given by:

$$\frac{\partial^2 l(\mathbf{w})}{\partial \mathbf{w}\mathbf{w}^T} = \sum_{i=1}^{n} \frac{e^{\mathbf{w}^T\mathbf{x}_i}}{(1 + e^{\mathbf{w}^T\mathbf{x}_i})^2} \mathbf{x}_i\mathbf{x}_i^T, \tag{4}$$

Substituting (4) into the FIM equation yields:

$$\text{FIM} = n\mathbb{E}\left[\frac{e^{\mathbf{w}^T\mathbf{x}}}{(1 + e^{\mathbf{w}^T\mathbf{x}})^2} \mathbf{x}\mathbf{x}^T\right], \tag{5}$$

Finding a closed-form expression for the FIM for arbitrary $f(x)$ is challenging and is often approximated through empirical evaluation. To derive a non-singular closed-form FIM, we assume $\mathbf{x} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, resulting in:

$$\text{FIM} = n\sigma^2 \left[(\alpha_2 - \alpha_0)\mathbf{u}_1\mathbf{u}_1^T + \alpha_0\mathbf{I}\right], \tag{6}$$

where $\mathbf{u}_1 = \mathbf{w}/ \parallel \mathbf{w} \parallel$, $\alpha_k = \alpha_k(\sigma \parallel \mathbf{w} \parallel)$ and

$$\alpha_k(a) = \mathbb{E}\left[\frac{e^{az}}{(1 + e^{az})^2} z^k\right] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \frac{e^{az - \frac{1}{2}z^2}}{(1 + e^{az})^2} z^k dz, \tag{7}$$

where $z \sim \mathcal{N}(0, 1)$.

Next, we proceed to derive the Fisher Information Matrix (FIM) transformation outlined in Eq. (6). Let $\mathbf{u}_1 = \frac{\mathbf{w}}{\|\mathbf{w}\|}$, $c = \parallel \mathbf{w} \parallel$, and $\mathbf{U} = \begin{bmatrix}\mathbf{u}_1 & \mathbf{u}_2 & \dots & \mathbf{u}_d\end{bmatrix}$, where $\mathbf{u}_2, \mathbf{u}_3, \dots, \mathbf{u}_d$ can be arbitrarily chosen to satisfy $\mathbf{U}\mathbf{U}^T = \mathbf{I}$. We initiate by multiplying $\mathbf{U}\mathbf{U}^T$ on both sides of the FIM in Eq. (5), resulting in:

$$\text{FIM} = n\mathbf{U}\mathbb{E}\left[\frac{e^{c\mathbf{u}_1^T\mathbf{x}}}{(1 + e^{c\mathbf{u}_1^T\mathbf{x}})^2} \mathbf{U}^T\mathbf{x}\mathbf{x}^T\mathbf{U}\right]\mathbf{U}^T, \tag{8}$$

Let $\mathbf{z} = \mathbf{U}^T\mathbf{x}/\sigma$, noting that $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. Substituting $\mathbf{U}^T\mathbf{x}$ with $\sigma\mathbf{z}$ and $\mathbf{u}_1^T\mathbf{x} = \sigma z_1$ into Eq. (8) yields:

$$\text{FIM} = n\sigma^2\mathbf{U}\mathbb{E}\left[\frac{e^{c\sigma z_1}}{(1 + e^{c\sigma z_1})^2} \mathbf{z}\mathbf{z}^T\right]\mathbf{U}^T, \tag{9}$$

This expression can be succinctly denoted as:

$$\text{FIM} = n\sigma^2 \mathbf{U}\mathbf{A}\mathbf{U}^T, \tag{10}$$

where $\mathbf{A} = \mathbb{E}\left[\frac{e^{c\sigma z_1}}{(1+e^{c\sigma z_1})^2}\mathbf{z}\mathbf{z}^T\right]$, $\mathbf{V} = \mathbb{E}\left[\frac{e^{c\sigma z_1}}{(1+e^{c\sigma z_1})^2}\mathbf{z}\right]$, and α_0 is as defined in Eq. (7). Utilizing the independence among the $\mathbf{z}_i$'s, we can further simplify $\mathbf{A}$ and $\mathbf{V}$ as $\mathbf{A} = \text{diag}\begin{bmatrix}\alpha_2 & \alpha_0 & \dots & \alpha_0\end{bmatrix}$ and $\mathbf{V} = \begin{bmatrix}\alpha_1 & 0 & \dots & 0\end{bmatrix}^T$. Let $\mathbf{e}_i$ be the canonical vector with all zero entries except for 1 at the ith entry. Replacing $\mathbf{A} = (\alpha_2 - \alpha_0)\mathbf{e}_1\mathbf{e}_1^T + \alpha_0\mathbf{I}$ and $\mathbf{V} = \mathbf{e}_1\alpha_1$ into Eq. (10), we obtain:

$$\text{FIM} = n\left[\sigma^2\mathbf{U}[(\alpha_2 - \alpha_0)\mathbf{e}_1\mathbf{e}_1^T + \alpha_0\mathbf{I}]\right]\mathbf{U}^T, \tag{11}$$

Substituting $\mathbf{U}\mathbf{e}_1 = \mathbf{u}_1$ into Eq. (11) yields Eq. (6):

$$\text{CRLB} = \text{FIM}^{-1} = \frac{1}{n\sigma^2}\left[(\alpha_2 - \alpha_0)\mathbf{u}_1\mathbf{u}_1^T + \alpha_0\mathbf{I}\right]^{-1}, \tag{12}$$

The Sherman-Morrison formula is employed as follows:

Lemma 3.1 *Consider $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ as an invertible matrix, and let $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$ be vectors. If $\boldsymbol{u}^T\boldsymbol{A}^{-1}\boldsymbol{v} \neq -1$, then $\boldsymbol{A} + \boldsymbol{u}\boldsymbol{v}^T$ is invertible, and the inverse is given by*

$$(\boldsymbol{A} + \boldsymbol{u}\boldsymbol{v}^T)^{-1} = \boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^T\boldsymbol{A}^{-1}}{1 + \boldsymbol{v}^T\boldsymbol{A}^{-1}\boldsymbol{u}}. \tag{13}$$

This formula is utilized to determine the inverse of the Fisher Information Matrix (FIM). Let $\mathbf{A} = \alpha_0\mathbf{I}$, $\mathbf{u} = (\alpha_2 - \alpha_0)\mathbf{u}_1$, and $\mathbf{v} = \mathbf{u}_1$. Substituting into Eq. (13) results in:

$$\text{CRLB} = \text{FIM}^{-1} = \frac{1}{n\sigma^2}\left(\frac{1}{\alpha_0}\mathbf{I} + \frac{\frac{\alpha_2 - \alpha_0}{\alpha_0^2}\mathbf{u}_1\mathbf{u}_1^T}{1 + \frac{\alpha_2 - \alpha_0}{\alpha_0}\mathbf{u}_1^T\mathbf{u}_1}\right), \tag{14}$$

$$\text{CRLB} = \frac{1}{n\sigma^2\alpha_0}\left[\mathbf{I} - \frac{\alpha_2 - \alpha_0}{\alpha_2}\mathbf{u}_1\mathbf{u}_1^T\right], \tag{15}$$

This Cramér-Rao Lower Bound (CRLB) is then employed to establish a lower bound on the mean squared error of an unbiased estimation of $\mathbf{w}$:

$$\mathbb{E}(\parallel \hat{\mathbf{w}} - \mathbf{w} \parallel^2) \geq \text{tr}(\text{CRLB}) = \sum_{i=1}^{d} \text{CRLB}_{ii}, \tag{16}$$

4 Derive the likelihood of ML estimation of w

Following (3), the likelihood of the ML estimation is given by

$$l(\mathbf{w}) = \sum_{i=1}^{n}\left(y_i(\mathbf{w}^T\mathbf{x}_i) - \log(1 + e^{\mathbf{w}^T\mathbf{x}_i})\right), \tag{17}$$

We have the following optimization problem:

$$\hat{\mathbf{w}}_{MLE} = \arg\max_{\mathbf{w}} l(\mathbf{w}) = \arg\max_{\mathbf{w}} \sum_{i=1}^{n} \left(y_i(\mathbf{w}^T \mathbf{x}_i) - \log(1 + e^{\mathbf{w}^T \mathbf{x}_i}) \right), \tag{18}$$

The gradient of the log-likelihood function is given by

$$\nabla_{\mathbf{w}} l(\mathbf{w}) = \sum_{i=1}^{n} \left(y_i - \frac{e^{\mathbf{w}^T \mathbf{x}_i}}{1 + e^{\mathbf{w}^T \mathbf{x}_i}} \right) \mathbf{x}_i, \tag{19}$$

5 Derive the iterative scaling iterations for estimating w as well as the gradient descent approach for estimating w.

5.1 Iterative scaling iterations approach

5.1.1 Version 1

The iterative scaling approach used in likelihood maximization is a lower-bound methodology. Each iteration involves the formulation of a simple lower bound on the likelihood, followed by a transition to its maximum. Notably, iterative scaling provides an additive lower bound concerning the parameters w_k, allowing for the flexibility to update either a singular parameter or all parameters in each step [5].

Let $s = \max_i \sum_k |x_{ik}|$. The foundation of iterative scaling relies on the consideration of the following two bounds:

Lemma 5.1

$$-\log(x) \geq 1 - \frac{x}{x_0} - \log(x_0), \tag{20}$$

for any x_0

Lemma 5.2

$$-\exp(-\sum_k q_k w_k) \geq -\sum_k q_k \exp(-w_k) - (1 - \sum_k q_k), \tag{21}$$

for any $q_k > 0$ *satisfying* $\sum_k q_k \leq 1$

The second lemma is derived through the application of Jensen's inequality to the function e^{-x}:

$$\begin{cases} \exp(-\sum_k q_k w_k) \leq \sum_k q_k \exp(-w_k), \\ \qquad\qquad \text{if } \ \sum_k q_k = 1, \end{cases} \tag{22}$$

Let some of the parameters w_k be set to 0, yielding

$$\begin{cases} \exp(-\sum_k q_k w_k) \leq \sum_k q_k \exp(-w_k) + (1 - \sum_k q_k), \\ \qquad\qquad \text{if } \ \sum_k q_k \leq 1, \end{cases} \tag{23}$$

Commence by expressing the symmetric likelihood function as:

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \prod_i \frac{1}{1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i)}, \tag{24}$$

Upon applying the first lemma (20) at the current parameter values $\mathbf{w}_0$, it follows that the log-likelihood function is bounded by

$$\begin{aligned} p(\mathbf{y}|\mathbf{X}, \mathbf{w}) &= -\sum_i \log(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i)), \\ &\geq \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}_0) + \sum_i \left(1 - \frac{1 + \exp(-y_i \mathbf{w})^T \mathbf{x}_i}{1 + \exp(-y_i \mathbf{w}_0^T \mathbf{x}_i)}\right), \\ &= \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}_0) + \sum_i (1 - \sigma(y_i \mathbf{w}_0^T \mathbf{x}_i))(1 - \exp(-y_i (\mathbf{w} - \mathbf{w}_0)^T \mathbf{x}_i)), \end{aligned} \tag{25}$$

where $\sigma(m) = \frac{1}{1+\exp(-m)}$.

Maximizing this bound over the parameter vector $\mathbf{w}$ remains a formidable task. Consequently, we employ the second lemma (21) by setting $q_k = |x_{ik}|/s$, yielding:

$$g(w_k) = -\sum_i (1 - \sigma(y_i \mathbf{w}_0^T x_i)) \sum_k \frac{|x_i k|}{x} \exp(-y_i \text{sign}(x_{ik}) s (w_k - w_{0k})), \tag{26}$$

$$\frac{dg(w_k)}{dw_k} = \sum_i (1 - \sigma(y_i \mathbf{w}_0^T x_i)) y_i x_{ik} \exp(-y_i \text{sign}(x_{ik}) s (w_k - w_{0k})) = 0, \tag{27}$$

$$\exp(2s(w_k - w_{0k})) = \frac{\sum_{i|y_i x_{ik>0}} (1 - \sigma(y_i \mathbf{w}_0^T \mathbf{x}_i)) |x_{ik}|}{\sum_{i|y_i x_{ik<0}} (1 - \sigma(y_i \mathbf{w}_0^T \mathbf{x}_i)) |x_{ik}|}, \tag{28}$$

$$w_k = w_{0k} + \frac{1}{2s} \log \frac{\sum_{i|y_i x_{ik>0}} (1 - \sigma(y_i \mathbf{w}_0^T \mathbf{x}_i)) |x_{ik}|}{\sum_{i|y_i x_{ik<0}} (1 - \sigma(y_i \mathbf{w}_0^T \mathbf{x}_i)) |x_{ik}|}, \tag{29}$$

Hence, the iterative scaling update is given by

$$\hat{w}_k^{new} = \hat{w}_k^{old} + \frac{1}{2s} \log \frac{\sum_{i|y_i x_{ik>0}} (1 - \sigma(y_i \left(\hat{\mathbf{w}}^{old}\right)^T \mathbf{x}_i)) |x_{ik}|}{\sum_{i|y_i x_{ik<0}} (1 - \sigma(y_i \left(\hat{\mathbf{w}}^{old}\right)^T \mathbf{x}_i)) |x_{ik}|}, \tag{30}$$

The implementation algorithm of the iterative scaling method is summarized in Algorithm 1. The cost of this algorithm is $O(nd)$ per iteration.

Algorithm 1 Iterative Scaling Approach Version 1.

```
Input: $\hat{\mathbf{w}}^{(0)}$, $\mathbf{X}$, $\mathbf{y}$
    Initialisation: Set $t = 0$
repeat
    $s = \max_i \sum_k |x_{ik}|$;
    $\hat{w}_k^{(t+1)} = \hat{w}_k^{(t)} + \frac{1}{2s} \log \frac{\sum_{i|y_i x_{ik}>0} (1-\sigma(y_i (\hat{\mathbf{w}}^{(t)})^T \mathbf{x}_i))|x_{ik}|}{\sum_{i|y_i x_{ik}<0} (1-\sigma(y_i (\hat{\mathbf{w}}^{(t)})^T \mathbf{x}_i))|x_{ik}|}$;
    $t = t + 1$;
until convergence
Output: $\hat{\mathbf{w}}^*$
```

5.1.2 Version 2

The parameter $\mathbf{w}$ is derived by maximizing the log-likelihood function $l(\mathbf{w})$ or minimizing $\min_{\mathbf{w}} -l(\mathbf{w})$, expressed by the equation:

$$f(\mathbf{w}) = -l(\mathbf{w}) = \sum_{i=1}^{n} (\log(1 + e^{\mathbf{w}^T \mathbf{x}_i}) - \mathbf{w}^T \mathbf{x}_i y_i), \tag{31}$$

To construct a surrogate for $f(\mathbf{w})$, we initiate by establishing an upper bound g. Initially, we provide an upper bound for $\log(1 + e^{\mathbf{x}^T \mathbf{w}})$ as follows:

$$\begin{aligned} \log(1 + e^{\mathbf{x}_i^T \mathbf{w}}) &= \log(1 + e^{\mathbf{x}_i^T \mathbf{w}'} e^{\mathbf{x}_i^T (\mathbf{w} - \mathbf{w}')}), \\ &= \log(1 + e^{\mathbf{x}_i^T \mathbf{w}'}) + \log\left(\frac{1}{1 + e^{\mathbf{x}_i^T \mathbf{w}'}} + \frac{e^{\mathbf{x}_i^T \mathbf{w}'}}{1 + e^{\mathbf{x}_i^T \mathbf{w}'}} e^{\mathbf{x}_i^T (\mathrm{w} - \mathbf{w}')}\right), \\ &= \log(1 + e^{\mathbf{x}_i^T \mathbf{w}'}) + \log(1 - p_i(\mathbf{w}')) + p_i(\mathbf{w}') e^{\mathbf{x}_i^T (\mathbf{w} - \mathbf{w}')}, \\ &= \log(1 + e^{\mathbf{x}_i^T \mathbf{w}'}) + \log(1 + p(\mathbf{w}'))(e^{\mathbf{w}_i^T (\mathbf{w} - \mathbf{w}')} - 1), \\ &\leq \log(1 + e^{\mathbf{x}_i^T \mathbf{w}'} + p_i(\mathbf{w}'))(e^{\mathbf{x}_i^T (\mathbf{w} - \mathbf{w}')} - 1), \\ &= \log(1 + e^{\mathbf{x}_i^T \mathbf{w}'}) - p_i(\mathbf{w}') + p_i(\mathbf{w}') e^{\mathbf{x}_i^T (\mathbf{w} - \mathbf{w}')}, \end{aligned} \tag{32}$$

where $p_i(\mathbf{w}) = \frac{e^{\mathbf{x} i^T \mathbf{w}}}{1 + e^{\mathbf{x} i^T \mathbf{w}}}$. It is important to note that this bound lacks a closed-form solution, necessitating further bounding. To achieve this, we proceed to bound $e^{\mathbf{x}_i^T (\mathbf{w} - \mathbf{w}')}$. Define $s(\mathbf{x}_i) = \sum_k |x_{ik}|$ and let $S = \max_i s(\mathbf{x}_i)$. Additionally, denote $x_{ik} = |x_{ik}| \mathrm{sgn}(x_{ik})$, and $q_{ik} = \frac{|x_{ik}|}{S}$. It is worth noting that $q_{ik} \geq 0$ and $\sum_k q_{ik} \leq 1$. As a result, we obtain:

$$
\begin{aligned}
e^{\mathbf{x}_i^T(\mathbf{w}-\mathbf{w}^{'})} &= e^{\sum_k x_{ik}(\mathbf{w}-\mathbf{w}^{'})}, \\
&= e^{\sum_k |x_{ik}| \mathrm{sgn}(x_{ik})(\mathbf{w}_k-\mathbf{w}_k^{'})}, \\
&= e^{\sum_k \frac{|x_{ik}|}{S} S\mathrm{sgn}(x_{ik})(\mathbf{w}_k-\mathbf{w}_k^{'})}, \\
&= e^{\sum_k q_{ik} S\mathrm{sgn}(x_{ik})(\mathbf{w}_k-\mathbf{w}_k^{'})} + (1-\sum_k q_{ik})0, \qquad (33) \\
&\leq \sum_k q_{ik} e^{s(\mathbf{x})\mathrm{sgn}(\mathbf{x}_k)(\mathbf{w}_k-\mathbf{w}_k^{'})} + (1-\sum_k q_{ik}e^0), \\
&= \sum_k \frac{|x_{ik}|}{S} e^{S\mathrm{sgn}(x_{ik})(\mathbf{w}_k-\mathbf{w}_k^{'})} + 1 - \sum_k q_{ik},
\end{aligned}
$$

The inequality in the above step relies on the convexity of the exponential function, or equivalently, Jensen's inequality. Specifically, it can be expressed as $e^{\sum_i \alpha_i t_i} \leq \alpha_i e^{t_i}$, where the probabilities α_i adhere to non-negativity ($\alpha_i \geq 0$) and the sum-to-one constraint ($\sum_i \alpha_i = 1$).

$$
\log(1+e^{\mathbf{x}_i^T\mathbf{w}}) \leq const + p_i(\mathbf{w}^{'}) \sum_k \frac{|x_{ik}|}{S} e^{S\mathrm{sgn}(x_{ik})(\mathbf{w}_k-\mathbf{w}_k^{'})}, \qquad (34)
$$

where *const* represents a term that is independent of $\mathbf{w}$. In conclusion, we can constrain $f(\mathbf{w})$ as follows:

$$
f(\mathbf{w}) \leq g(\mathbf{w},\mathbf{w}^{'}), \qquad (35)
$$

and

$$
g(\mathbf{w},\mathbf{w}^{'}) = const^{'} + \sum_{i=1}^{n}\sum_{k=1}^{d}\left(p_i(\mathbf{w}^{'})\frac{|x_{ik}|}{S} e^{S\mathrm{sgn}(x_{ik})(\mathbf{w}_k-\mathbf{w}_k^{'})} - \mathbf{w}_k x_{ik} y_i\right), \qquad (36)
$$

Drawing from the inequalities employed in constructing the surrogate function, it is evident that $f(\mathbf{w}) \leq g(\mathbf{w},\mathbf{w}^{'})$. Furthermore, by setting $\mathbf{w} = \mathbf{w}^{'}$, we can confirm that the two inequalities hold with equality, implying $f(\mathbf{w}^{'}) = g(\mathbf{w}^{'},\mathbf{w}^{'})$. Consequently, g fulfills the criteria for a surrogate function. The resulting surrogate $g(\mathbf{w},\mathbf{w}^{'})$ is separable in the elements of the vector $\mathbf{w}$, enabling us to address d distinct minimization problems (one for each element of $\mathbf{w}$) to minimize g. The outcome of the minimization concerning $\mathbf{w}_k$ is expressed as:

$$
w_k^{t+1} = w_k^{(t)} + \frac{1}{S}\log\left(\frac{B_k + \sqrt{B_k^2 + 4A_{1k}A_{2k}}}{2A_{1k}}\right), \qquad (37)
$$

where

$$\begin{cases} A_{1k} = \frac{1}{2}\sum_i(|x_{ik}| + x_{ik})p_i(\mathbf{w}^{'}), \\ A_{2k} = \frac{1}{2}\sum_i(|x_{ik}| - x_{ik})p_i(\mathbf{w}^{'}), \\ B_j = \sum_i x_{ik}y_i, \end{cases} \tag{38}$$

The convergence of the algorithm occurs when the argument of the logarithm equals 1, specifically when $B_k = A_{1k} - A_{2k}$, i.e., when $\sum_i x_{ik}p_i(\mathbf{w}^{'}) = \sum_i x_{ik}y_i$. Remarkably, this aligns with the maximum likelihood equation derived by equating the derivative of the log-likelihood to zero. The algorithm is succinctly outlined in Algorithm 2.

Hence, the iterative scaling update is given by

$$\hat{w}_k^{new} = \hat{w}_k^{old} + \frac{1}{S}\log\left(\frac{B_k + \sqrt{B_k^2 + 4A_{1k}A_{2k}}}{2A_{1k}}\right), \tag{39}$$

where

$$\begin{cases} A_{1k} = \frac{1}{2}\sum_i(|x_{ik}| + x_{ik})p_i(\hat{\mathbf{w}}^{old}), \\ A_{2k} = \frac{1}{2}\sum_i(|x_{ik}| - x_{ik})p_i(\hat{\mathbf{w}}^{old}), \\ B_j = \sum_i x_{ik}y_i, \end{cases} \tag{40}$$

and $p_i(\hat{\mathbf{w}}^{old}) = \frac{e^{\mathbf{x}_i{}^T\hat{\mathbf{w}}^{old}}}{1+e^{\mathbf{x}_i{}^T\hat{\mathbf{w}}^{old}}}$

Algorithm 2 Iterative Scaling Approach Version 2.

```
Input: $\hat{\mathbf{w}}^{(0)}$, $\mathbf{X}$, $\mathbf{y}$
   Initialisation: Set $t = 0$
repeat
    $S = \max_i \sum_k |x_{ik}|$;
    $p_i(\hat{\mathbf{w}}^{(t)}) = \frac{e^{\mathbf{x}_i{}^T\hat{\mathbf{w}}^{(t)}}}{1+e^{\mathbf{x}_i{}^T\hat{\mathbf{w}}^{(t)}}}$
    $A_{1k} = \frac{1}{2}\sum_i(|x_{ik}| + x_{ik})p_i(\hat{\mathbf{w}}^{(t)})$;
    $A_{2k} = \frac{1}{2}\sum_i(|x_{ik}| - x_{ik})p_i(\hat{\mathbf{w}}^{(t)})$;
    $B_j = \sum_i x_{ik}y_i$;
    $\hat{w}_k^{t+1} = \hat{w}_k^{(t)} + \frac{1}{S}\log\left(\frac{B_k+\sqrt{B_k^2+4A_{1k}A_{2k}}}{2A_{1k}}\right)$;
    $t = t + 1$;
until convergence
Output: $\hat{\mathbf{w}}^*$
```

5.2 Gradient descent approach

We can use the gradient descent approach to find the ML estimation of $\mathbf{w}$. The gradient descent update is given by

$$\begin{aligned}\hat{\mathbf{w}}^{new} &= \hat{\mathbf{w}}^{old} + \eta \nabla_{\mathbf{w}} l(\hat{\mathbf{w}}^{old}), \\ &= \hat{\mathbf{w}}^{old} + \eta \sum_{i=1}^{n} \left(y_i - \frac{e^{\hat{\mathbf{w}}^{old^T} \mathbf{x}_i}}{1 + e^{\hat{\mathbf{w}}^{old^T} \mathbf{x}_i}} \right) \mathbf{x}_i,\end{aligned} \tag{41}$$

where η is the step size.

The gradient descent algorithm, described in Algorithm 3, is guaranteed to converge to the global maximum of the log-likelihood function if η is sufficiently small. However, the gradient descent algorithm is computationally expensive since it requires the computation of the gradient at each iteration.

Algorithm 3 Gradient Descent Approach.

```
Input: $\hat{\mathbf{w}}^{(0)}$, $\eta$, $\mathbf{X}$, $\mathbf{y}$
    Initialisation: Set $t = 0$
repeat
    $\hat{\mathbf{w}}^{(t+1)} = \hat{\mathbf{w}}^{(t)} + \eta \nabla_{\mathbf{w}} l(\hat{\mathbf{w}}^{(t)}) = \hat{\mathbf{w}}^{(t)} + \eta \sum_{i=1}^{n} \left( y_i - \frac{e^{\left(\hat{\mathbf{w}}^{(t)}\right)^T \mathbf{x}_i}}{1+e^{\left(\hat{\mathbf{w}}^{(t)}\right)^T \mathbf{x}_i}} \right) \mathbf{x}_i$;
    $t = t + 1$;
until convergence
Output: $\hat{\mathbf{w}}^{*}$
```

6 Jeffreys Prior for MAP Estimation

In **MAP** estimation, the model parameters are determined by solving:

$$\hat{\mathbf{w}}MAP = \arg\max \mathbf{w} \left(l(\mathbf{w}) + \log(p(\mathbf{w})) \right), \tag{42}$$

Our objective is to address the maximization problem in (42) specifically in the context of Jeffreys prior. According to Jeffreys prior, the probability of the prior is proportional to the square root of the determinant of the **FIM**:

$$p(\mathbf{w}) \propto \sqrt{\det(\mathbf{FIM})}, \tag{43}$$

The determinant of **FIM** is given by

$$\det(\mathbf{FIM}) = \det(n\sigma^2 \left[(\alpha_2 - \alpha_0)\mathbf{u}_1\mathbf{u}_1^T + \alpha_0 \mathbf{I} \right]) = (n\sigma^2\alpha_0)^d \det\left(\frac{\alpha_2 - \alpha_0}{\alpha_0} \mathbf{u}_1\mathbf{u}_1^T + \mathbf{I} \right), \tag{44}$$

Lemma 6.1

$$\det(\boldsymbol{I}_m + \boldsymbol{u}\boldsymbol{v}^T) = 1 + \boldsymbol{v}^T \boldsymbol{u}. \tag{45}$$

where $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^m$

Algorithm 4 Gradient Ascent Approach.

```
Input: $\hat{\mathbf{w}}^{(0)}$, $\eta$, $\mathbf{X}$, $\mathbf{y}$
    Initialisation: Set $t = 0$
repeat
    $\frac{\partial \log p(\mathbf{X},\mathbf{y}|\hat{\mathbf{w}}^{(t)})}{\partial \hat{\mathbf{w}}^{(t)}} = \sum_{i=1}^{n} \left( y_i \mathbf{x}_i - \frac{e^{\left(\hat{\mathbf{w}}^{(t)}\right)^T \mathbf{x}_i} \mathbf{x}_i}{1+e^{\left(\hat{\mathbf{w}}^{(t)}\right)^T \mathbf{x}_i}} \right) + \frac{1}{2}\left[(d-1)\frac{\partial \alpha_0}{\partial \left\|\hat{\mathbf{w}}^{(t)}\right\|} + \frac{\partial \alpha_2}{\partial \left\|\hat{\mathbf{w}}^{(t)}\right\|}\right] \frac{\hat{\mathbf{w}}^{(t)}}{\left\|\hat{\mathbf{w}}^{(t)}\right\|}$;
    where $\frac{\partial \alpha_k}{\partial \left\|\hat{\mathbf{w}}^{(t)}\right\|} = \sigma \mathbb{E}\left[z^{k+1}(p - 3p^2 + 2p^3)\right]$ and $p = \frac{e^{\left\|\hat{\mathbf{w}}^{(t)}\right\|\sigma z}}{1+e^{\left\|\hat{\mathbf{w}}^{(t)}\right\|\sigma z}}$ with $z \sim \mathcal{N}(0,1)$
    $\hat{\mathbf{w}}^{(t+1)} = \hat{\mathbf{w}}^{(t)} + \eta \frac{\partial \log p(\mathbf{X},\mathbf{y}|\hat{\mathbf{w}}^{(t)})}{\partial \hat{\mathbf{w}}^{(t)}}$
    $t = t + 1$;
until convergence
Output: $\hat{\mathbf{w}}^{*}$
```

Apply the Lemma (45) to the determinant of **FIM** yields

$$\begin{aligned}\det(\mathbf{FIM}) &= (n\sigma^2\alpha_0)^d \left(1 + \frac{\alpha_2 - \alpha_0}{\alpha_0}\mathbf{u}_1^T\mathbf{u}_1\right) = (n\sigma^2\alpha_0)^d \left(1 + \frac{\alpha_2 - \alpha_0}{\alpha_0}\right) \\ &= (n\sigma^2)^d \alpha_0^{d-1}\alpha_2,\end{aligned} \tag{46}$$

The optimization problem (42) is equivalent to the following optimization problem:

$$\begin{aligned}\hat{\mathbf{w}}_{MAP} &= \arg\max_{\mathbf{w}} l(\mathbf{w}) + \log(\sqrt{(n\sigma^2)^d \alpha_0^{d-1}\alpha_2}), \\ &= \arg\max_{\mathbf{w}} \quad const + l(\mathbf{w}) + \frac{1}{2}(d-1)\log(\alpha_0) + \frac{1}{2}\log(\alpha_2),\end{aligned} \tag{47}$$

We adopt a gradient ascent approach to solve the optimization problem

$$\hat{\mathbf{w}}^{new} = \hat{\mathbf{w}}^{old} + \eta \frac{\partial \log p(\mathbf{X}, \mathbf{y}|\mathbf{w})}{\partial \mathbf{w}}, \tag{48}$$

where η is the step size. The gradient of the objective function is given by

$$\frac{\partial \log p(\mathbf{X}, \mathbf{y}|\mathbf{w})}{\partial \mathbf{w}} = \frac{\partial l(\mathbf{w})}{\mathbf{w}} + \frac{\partial \log p(\mathbf{w})}{\partial \mathbf{w}}, \tag{49}$$

$$\frac{\partial l(\mathbf{w})}{\mathbf{w}} = \sum_{i=1}^{n} \left(y_i \mathbf{x}_i - \frac{e^{\mathbf{w}^T \mathbf{x}_i} \mathbf{x}_i}{1 + e^{\mathbf{w}^T \mathbf{x}_i}} \right), \tag{50}$$

$$\frac{\partial \log p(\mathbf{w})}{\partial \mathbf{w}} = \frac{1}{2}\left[(d-1)\frac{\partial \alpha_0}{\partial \|\mathbf{w}\|} + \frac{\partial \alpha_2}{\partial \|\mathbf{w}\|}\right] \frac{\mathbf{w}}{\|\mathbf{w}\|}, \tag{51}$$

$$\frac{\partial \alpha_k}{\partial \|\mathbf{w}\|} = \sigma \mathbb{E}\left[z^{k+1}(p - 3p^2 + 2p^3) \right], \tag{52}$$

where $p = \frac{e^{\|\mathbf{w}\|\sigma z}}{1+e^{\|\mathbf{w}\|\sigma z}}$ and $z \sim \mathcal{N}(0,1)$.

7 Experimental Results

In this section, we present numerical simulations to assess our expression for the CRLB, comparing it with the MSE of ML estimators under various conditions, including no regularization, l_1-regularization, l_2-regularization, and iterative scaling approach over a range of parameter values.

General settings: We employ Monte Carlo simulations to compute the **MSE** of all considered estimators. For each of the 200 Monte Carlo runs, we generate n feature vectors $\mathbf{x} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ with $\sigma^2 = 1$.

7.1 CRLB evaluation

For the assessment of the CRLB, we systematically vary the values of n and $|\mathbf{w}|$ and subsequently compare the CRLB with the MSE of Maximum Likelihood (ML) estimators.

7.1.1 CRLB and MSE as a function of the data size n

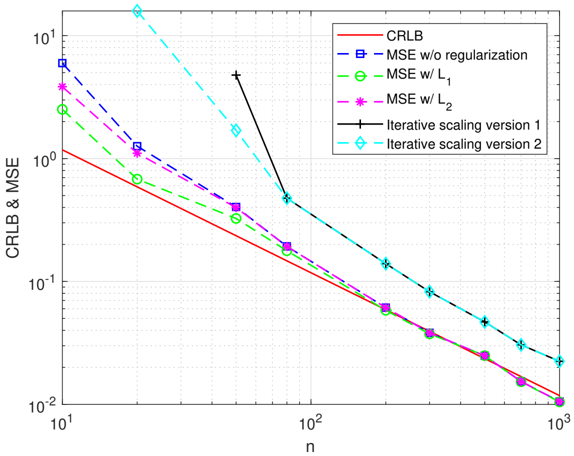


Figure 1: **CRLB** and **MSE** of **ML** estimations as a function of n for $\mathbf{w} = [1,1]^T\sqrt{2}$.

To investigate the influence of the number of data points, n, on both the Cramér-Rao Lower Bound (**CRLB**) and the Mean Squared Error (**MSE**), we

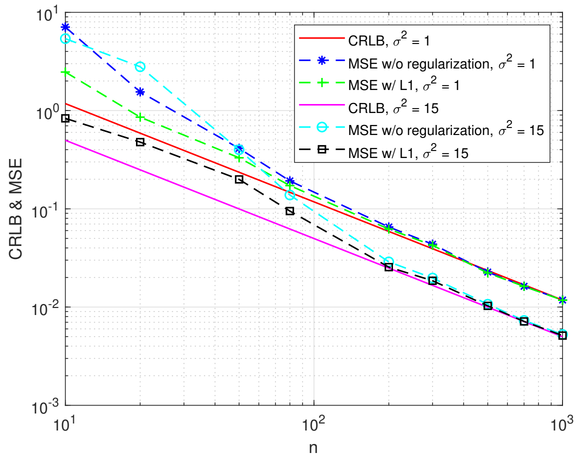


Figure 2: **CRLB** and **MSE** of **ML** estimations (no regularization, l_1-regularization) as a function of n with $d = 2$ for different values of σ^2.

vary n across the following values: [10, 20, 50, 80, 200, 300, 500, 700, 1000], while fixing $\mathbf{w} = [1, 1]^T\sqrt{2}$.

Figure 1 illustrates a decreasing trend for both the CRLB and the MSE with respect to n. Initially, with a small value of n, there is a substantial gap between the CRLB and MSE. However, as n increases, this gap gradually diminishes. For sufficiently large values of n, the difference between the CRLB and MSE becomes negligible. This observation aligns with the asymptotic property of the CRLB. As anticipated, the ML estimator demonstrates asymptotic efficiency. We also observe that the MSE of the ML estimator without regularization, l_1 and l_2 regularization are much better than the MSE of iterative scaling approach.

Figure 2 presents the CRLB and MSE as functions of n for various values of σ^2. It is observed that as σ^2 rises from 1 to 15 (resulting in an increase in Signal-to-Noise Ratio (SNR) since $\mathbf{x} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$), both the CRLB and MSE decrease. Additionally, the gap between the CRLB and MSE slightly widens with the increase in σ^2.

7.1.2 CRLB and MSE as a function of the norm of $\| \mathbf{w} \|$

In this context, we systematically vary the norm of the vector $\mathbf{w}$ in an incremental manner, considering the following values: [0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10, 15, 20]. Additionally, we explore different data sizes with $n \in [50, 100, 1000]$. The objective of this experimental setup is to examine the impact of the classifier's sharpness, as denoted by $||\mathbf{w}||$, on the MSE of the ML estimator.

Figure 3 illustrates the MSE of the ML estimator (without regularization)

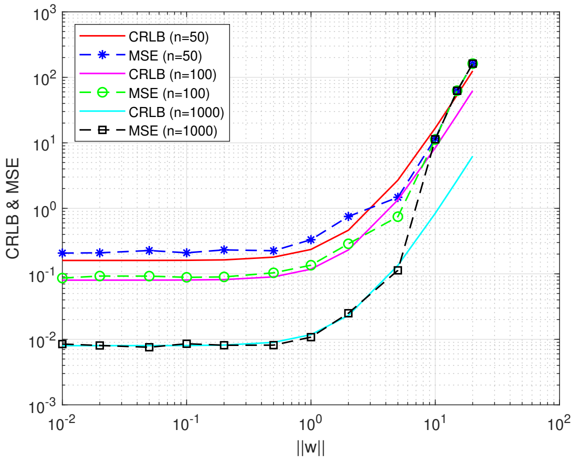


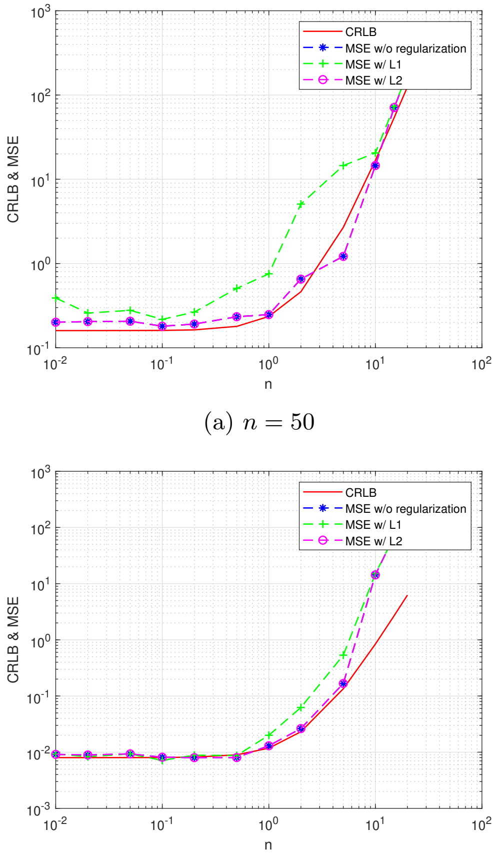
Figure 3: **CRLB** and **MSE** as a function of $\parallel \mathbf{w} \parallel$ for $n \in \{50, 100, 1000\}$.

and the CRLB as functions of $\parallel \mathbf{w} \parallel$. It is observed that with an increase in the value of $\parallel \mathbf{w} \parallel$, both the CRLB and MSE exhibit an increase. A comparison among curves associated with different values of n reveals that as n increases, the MSE of the ML estimator decreases, indicating increased efficiency. Furthermore, the MSE of the ML estimator consistently exceeds the CRLB, and the gap between them widens with the growing $\parallel \mathbf{w} \parallel$. As $\parallel \mathbf{w} \parallel$ increases, the ML estimator becomes less efficient. In general, achieving accurate parameter estimates for the Logistic Regression (LR) model when $\parallel \mathbf{w} \parallel$ is large necessitates a substantial increase in the number of data points n.

7.2 CRLB and MSE of ML estimation

In this section, we assess the CRLB and MSE of ML estimations under different regularization approaches, including no regularization, l_1-regularization, and l_2-regularization. The evaluation is conducted under the same conditions as in Section 7.1.2, except for the variation in n within the range $[50, 1000]$. For l_1 regularization, we introduce $\lambda \parallel \mathbf{w} \parallel_1$ to $l(\mathbf{w})$ in (3), and for l_2 regularization, we add $\lambda \parallel \mathbf{w} \parallel_2^2$.

Figures 4a and 4b depict the CRLB and MSE of ML estimations under various regularization approaches for $n = 50$ and $n = 1000$, respectively. For a smaller data size, $n = 50$, the MSE of the ML estimator exhibits a rapid increase with $\parallel \mathbf{w} \parallel$, resulting in a substantial gap with the CRLB. Conversely, with a larger data size, $n = 1000$, the gap between the MSE and CRLB is significantly smaller.



(a) $n = 50$

(b) $n = 1000$

Figure 4: **CRLB** and **MSE** of **ML** estimations as a function of $\parallel \mathbf{w} \parallel$ with $d = 2$ for (a) $n = 50$ and (b) $n = 1000$.

8 References

References

[1] T. Nguyen, R. Raich and P. Lai, "Jeffreys prior regularization for logistic regression," *2016 IEEE Statistical Signal Processing Workshop (SSP),* Palma de Mallorca, Spain, 2016, pp. 1-5.

[2] T. Zhang and F. Oles, "Text categorization based on regularized linear classifiers," *Information Retrieval,* vol. 4, pp. 531, 2001.

[3] Su-In Lee, Honglak Lee, Pieter Abbeel, and Andrew Y. Ng, "Efficient l 1 regularized logistic regression," *Proceedings of the National Conference on*

Artificial Intelligence, vol. 21, 2006.

[4] T. Zhang and F. J. Oles, “EA probability analysis on the value of unlabeled data for classification problems,” *ICML,* pp. 1191–1198, 2000.

[5] Thomas P Minka, ““Algorithms for maximum-likelihood logistic regression,” 2003.