Problem Sheet 2

1 Maximum Likelihood Estimation of σ

As presented in Lecture 4, we consider a discriminative framework, where the input datapoints $\mathbf{x}_1, \dots, \mathbf{x}_N$ are fixed (we will not consider these as being generated by a random process). Let $\mathbf{w}$ and σ be the parameters defining the linear model with Gaussian noise, *i.e.,*

$$y_i \sim \mathcal{N}(\mathbf{x}_i^\mathsf{T}\mathbf{w}, \sigma^2). \tag{1.1}$$

In Lecture 4 we showed that the maximum likelihood estimate for $\mathbf{w}$ is the same as the least square estimator, $\mathbf{w}_{\mathrm{ML}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$. Show that the MLE for σ^2 is given by

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}(\mathbf{y} - \mathbf{X}\mathbf{w}_{\mathrm{ML}})^\mathsf{T}(\mathbf{y} - \mathbf{X}\mathbf{w}_{\mathrm{ML}}). \tag{1.2}$$

2 Centering and Ridge Regression

Assume that $\frac{1}{N}\sum_{i=1}^N \mathbf{x}_i = \mathbf{0}$, *i.e.,* the data is centered. (In this question we will treat the constant term separately, as centering this would give us a column of 0s.) Let us denote the parameter for the leading constant term as b (for "bias"). So the linear model is $\widehat{y} = b + \mathbf{x}^\mathsf{T}\mathbf{w}$. Consider minimizing the ridge objective:

$$\mathcal{L}_{\text{ridge}}(\mathbf{w}, b) = (\mathbf{X}\mathbf{w} + b\mathbf{1} - \mathbf{y})^\mathsf{T}(\mathbf{X}\mathbf{w} + b\mathbf{1} - \mathbf{y}) + \lambda\mathbf{w}^\mathsf{T}\mathbf{w} \tag{2.1}$$

Here $\mathbf{1}$ is the vector of all ones and note that b^2 is not regularized. Show that if $\widehat{b}$ and $\widehat{\mathbf{w}}$ are the resulting solutions obtained by minimising the above objective, then

$$\widehat{b} = \frac{1}{N}\sum_{i=1}^N y_i$$
$$\widehat{\mathbf{w}} = (\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I}_D)^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$$

What happens if you also center $\mathbf{y}$?

3 Bias of the Least Squared Estimator

Suppose that the data $\mathcal{D} = \langle(\mathbf{x}_i, y_i)\rangle_{i=1}^N$ is truly generated from a linear model, *i.e.,*

$$\mathbb{E}\left[y \mid \mathbf{x}, \mathbf{w}^*\right] = \mathbf{x}^\mathsf{T}\mathbf{w}^* \tag{3.1}$$

for some fixed (but unknown) parameter vector $\mathbf{w}^*$. Recall that the least squares estimator is

$$\widehat{\mathbf{w}}_{\mathrm{LS}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}. \tag{3.2}$$

1. Assume that $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are fixed and that $\mathbf{X}^\mathsf{T}\mathbf{X}$ is invertible. You can think of the data $\mathcal{D}$ as a random variable (because of the possible noise in the y_is). Thus, $\widehat{\mathbf{w}}_{\mathrm{LS}}(\mathcal{D})$ is itself a random variable. Show that the expectation of the estimator $\widehat{\mathbf{w}}_{\mathrm{LS}}(\mathcal{D})$ (with respect to $\mathcal{D}$) is $\mathbf{w}^*$. Such an estimator is called an *unbiased* estimator, as its expectation equals the true parameter value.

2. Now suppose we have some other estimator $\widehat{\mathbf{w}}$ which may not be unbiased. The bias is defined as
$$\mathrm{Bias}(\widehat{\mathbf{w}}) = \|\mathop{\mathbb{E}}_{\mathcal{D}}\left[\widehat{\mathbf{w}}(\mathcal{D})\right] - \mathbf{w}^*\|. \tag{3.3}$$
Thus, the bias is the Euclidean distance between the expectation of the estimator and the true parameter. Suppose you are interested in minimizing the squared distance between the estimated parameters and true parameters, *i.e.,* to minimize $\|\widehat{\mathbf{w}}(\mathcal{D}) - \mathbf{w}^*\|^2$. Show that the expected (with respect to $\mathcal{D}$) squared distance can be decomposed as follows:
$$\mathop{\mathbb{E}}_{\mathcal{D}}\left[\|\widehat{\mathbf{w}}(\mathcal{D}) - \mathbf{w}^*\|^2\right] = \|\mathop{\mathbb{E}}_{\mathcal{D}}\left[\widehat{\mathbf{w}}(\mathcal{D})\right] - \mathbf{w}^*\|^2 + \mathop{\mathbb{E}}_{\mathcal{D}}\left[\|\widehat{\mathbf{w}}(D) - \mathop{\mathbb{E}}_{\mathcal{D}}\left[\widehat{\mathbf{w}}(\mathcal{D})\right]\|^2\right] \tag{3.4}$$
The first term above is just the squared bias and the second term above is the variance of the estimator. Thus, while being unbiased looks like a natural property to demand of estimators, it might sometimes be preferable to have a *biased* estimator if it has a much lower variance. This is what ridge regression or LASSO does.

4 Maximum Likelihood and Model Selection

Let the random variable $x \in \{0, 1\}$ model the outcome of an experiment, such that the event $x = 1$ occurs with probability θ_1. Suppose that someone else observes the experiment and reports to you the outcome, y. But this person is unreliable and only reports the result correctly with probability θ_2. That is, $p(y \mid x, \theta_2)$ is given by

	$y = 0$	$y = 1$
$x = 0$	θ_2	$1 - \theta_2$
$x = 1$	$1 - \theta_2$	θ_2

Assume that θ_2 is independent of x and θ_1.

1. Write down the joint probability distribution $p(x, y \mid \boldsymbol{\theta})$ as a 2×2 table, in terms of $\boldsymbol{\theta} = (\theta_1, \theta_2)$.

2. Given the following dataset: $\mathbf{x} = (1, 1, 0, 1, 1, 0, 0)$, $\mathbf{y} = (1, 0, 0, 0, 1, 0, 1)$. What are the numerical values of the MLEs for θ_1 and θ_2? What is the numerical value $p(\mathcal{D} \mid \widehat{\boldsymbol{\theta}}, M_2)$ where M_2 denotes this 2-parameter model? Justify your answer by including the derivations.

3. Now consider a model with 4 parameters, $\boldsymbol{\theta} = (\theta_{0,0}, \theta_{0,1}, \theta_{1,0}, \theta_{1,1})$, representing $p(x, y \mid \boldsymbol{\theta}) = \theta_{x,y}$. (Only 3 of these parameters are free to vary, since they must sum to one.) What is the MLE of $\boldsymbol{\theta}$? What is $p(\mathcal{D} \mid \widehat{\boldsymbol{\theta}}, M_4)$ where M_4 denotes this 4-parameter model?

4. Suppose we are not sure which model is correct. We compute the leave-one-out cross-validated log-likelihood of the 2-parameter model and the 4-parameter model as follows:

$$L(M) = \sum_{i=1}^{N} \log p(x_i, y_i \mid M, \widehat{\theta}(\mathcal{D}_{-i}))$$

and $\widehat{\theta}(\mathcal{D}_{-i})$ denotes the MLE computed on $\mathcal{D}$ excluding the i^{th} observation. Which model will CV pick and why?

5 The *Huber loss* in a linear regression setting

In this question, we will investigate the *Huber loss* in a linear regression setting. Given arbitrary but fixed parameters $\lambda, \mu \in \mathbb{R}$ such that $\lambda, \mu > 0$, the Huber loss is given by the function $h_{\lambda,\mu} : \mathbb{R} \to \mathbb{R}$ such that

$$h_{\lambda,\mu}(z) = \begin{cases} \lambda \left(|z| - \frac{\lambda}{4\mu} \right) & \text{if } |z| \geq \frac{\lambda}{2\mu}, \\ \mu z^2 & \text{otherwise.} \end{cases}$$

Given a vector $\mathbf{z} = (z_1, \ldots, z_D) \in \mathbb{R}^D$, we extend $h_{\lambda,\mu}$ such that $h_{\lambda,\mu}(\mathbf{z}) = \sum_{i=1}^{D} h_{\lambda,\mu}(z_i)$. Recall that when dealing with absolute values, the sign function defined as follows is often helpful:

$$\operatorname{sign}(z) = \begin{cases} 1 & \text{if } z > 0, \\ -1 & \text{otherwise.} \end{cases}$$

1. Let us fix $\lambda = 4$ and $\mu = 1$. Draw three graphs plotting $h_{4,1}(z)$, and the absolute and the square loss functions. Briefly compare the Huber loss to the absolute and the square loss functions. What can you say about outliers?

2. Given a training set $\mathcal{D} = \langle (\mathbf{x}_i, y_i) \rangle_{i=1}^{N}$, where $\mathbf{x}_i \in \mathbb{R}^D$ and $y_i \in \mathbb{R}$, we ignore any noise terms and define the loss function as

$$\mathcal{L}(\mathbf{w}; \mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} h_{\lambda,\mu}(y_i - \mathbf{w}^\mathsf{T} \cdot \mathbf{x}_i),$$

where $\mathbf{w} \in \mathbb{R}^D$ is the model parameter.

Compute $\nabla_{\mathbf{w}} \mathcal{L}$.

3. For the remainder of this question, we will discuss using the Huber loss as a regulariser. Let $\ell : \mathbb{R} \to \mathbb{R}$ be an arbitrary loss function, and consider the following regularised loss functions:

$$\mathcal{H}(\mathbf{z}; \mathcal{D}) = h_{\lambda,\mu}(\mathbf{z}) + \frac{1}{N} \sum_{i=1}^{N} \ell(y_i - \mathbf{z}^\mathsf{T} \cdot \mathbf{x}_i),$$

$$\mathcal{S}(\mathbf{v}, \mathbf{w}; \mathcal{D}) = \lambda ||\mathbf{v}||_1 + \mu ||\mathbf{w}||_2^2 + \frac{1}{N} \sum_{i=1}^{N} \ell(y_i - (\mathbf{v} + \mathbf{w})^\mathsf{T} \cdot \mathbf{x}_i).$$

Suppose we let $\lambda \to \infty$ in $\mathcal{H}(\mathbf{z}; \mathcal{D})$ and $\mathcal{S}(\mathbf{v}, \mathbf{w}; \mathcal{D})$. Which types of regularised regression do we obtain? What happens when $\mu \to \infty$?