**Statistical Machine Learning**
**Statistics GR5241 — Spring 2022**

**Homework 1**

**Collaboration policy:** Collaboration on solving the homework is allowed, after you have thought about the problems on your own. It is also OK to get clarification (but not solutions) from books or online resources, again after you have thought about the problems on your own. There are two requirements: first, cite your collaborators fully and completely (e.g., "Jane explained to me what is asked in Question 3 (a)"). Second, write your solution *independently*: close the book and all of your notes, and send collaborators out of the room, so that the solution comes from you only.
**The following problems are due on Friday, February 11th, 11:59pm.**

1. **Bias-variance decomposition.**

   Consider a $p$-dimensional vector $\boldsymbol{x} \in \mathbb{R}^p$ drawn from a Gaussian distribution with an identity covariance matrix $\boldsymbol{\Sigma} = \mathbf{I}_p$ and an unknown mean $\boldsymbol{\mu}$, i.e. $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_p)$. Our goal is to evaluate the effectiveness of an estimator $\hat{\boldsymbol{\mu}} = \boldsymbol{f}(\boldsymbol{x})$ of the mean from only a single sample (i.e. $n = 1$) by measuring its mean squared error $\mathbb{E}[\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2]$, where $\|\cdot\|^2$ is the squared Euclidean norm and the expectation is taken over the data generating distribution.

   Note that for any estimator $\hat{\boldsymbol{\theta}}$ of a parameter vector $\boldsymbol{\theta}$, its mean squared error can be decomposed as:

$$\mathbb{E}\Big[\,\Big\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\Big\|^2\,\Big] = \Big\|\mathrm{Bias}[\hat{\boldsymbol{\theta}}]\Big\|^2 + \mathrm{Trace}(\mathrm{Var}[\hat{\boldsymbol{\theta}}]), \text{ where:}$$

$$\mathrm{Bias}[\hat{\boldsymbol{\theta}}] = \mathbb{E}[\hat{\boldsymbol{\theta}}] - \boldsymbol{\theta} \quad \text{and} \quad (\mathrm{Var}[\hat{\boldsymbol{\theta}}])_{j,j} = \mathrm{Var}[\hat{\theta}_j] = \mathbb{E}[(\hat{\theta}_j - \mathbb{E}[\hat{\theta}_j])^2]$$

   Here, $\mathrm{Trace}(\cdot)$ denotes the sum of the diagonal elements of a square matrix, $(\mathrm{Var}[\hat{\boldsymbol{\theta}}])_{j,j}$ denotes the $j$th diagonal element of $\mathrm{Var}[\hat{\boldsymbol{\theta}}]$, and $\hat{\theta}_j$ denotes the $j$th element of $\hat{\boldsymbol{\theta}}$.

   (a) (4 Points) Derive the maximum likelihood estimator:

$$\hat{\boldsymbol{\mu}}_{\mathrm{MLE}} = \arg\max_{\boldsymbol{\mu}} P(\boldsymbol{x}; \boldsymbol{\mu}).$$

   What is its mean squared error?

   (b) (4 Points) Derive the $\ell_2$-regularized maximum likelihood estimator:

$$\hat{\boldsymbol{\mu}}_{\mathrm{RMLE}} = \arg\max_{\boldsymbol{\mu}} \log P(\boldsymbol{x}; \boldsymbol{\mu}) - \lambda \|\boldsymbol{\mu}\|^2 .$$

   What is its mean squared error?

(c) (4 Points) Consider an estimator of the form $\hat{\boldsymbol{\mu}}_{\text{SCALE}} = c\boldsymbol{x}$ where $c \in \mathbb{R}$ is a constant scaling factor. Find the value $c^*$ that minimizes its mean squared error:

$$c^* = \arg\min_c \mathbb{E}[\|c\boldsymbol{x} - \boldsymbol{\mu}\|^2].$$

What is the corresponding minimum mean squared error?

(d) Consider the James-Stein estimator:

$$\hat{\boldsymbol{\mu}}_{\text{JS}} = \left(1 - \frac{p-2}{\|\boldsymbol{x}\|^2}\right)\boldsymbol{x}.$$

Note that $\hat{\boldsymbol{\mu}}_{\text{JS}}$ can be written as $\boldsymbol{x} - \boldsymbol{g}(\boldsymbol{x})$ where $\boldsymbol{g}(\boldsymbol{x}) = \frac{p-2}{\|\boldsymbol{x}\|^2}\boldsymbol{x}$. This allows us to separate the mean squared error into three parts:

$$\begin{aligned}
\mathbb{E}[\|\hat{\boldsymbol{\mu}}_{\text{JS}} - \boldsymbol{\mu}\|^2] &= \mathbb{E}[\|\boldsymbol{x} - \boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{\mu}\|^2] \\
&= \mathbb{E}[\boldsymbol{x}^\intercal\boldsymbol{x} - 2\boldsymbol{x}^\intercal\boldsymbol{\mu} + \boldsymbol{\mu}^\intercal\boldsymbol{\mu} + \boldsymbol{g}(\boldsymbol{x})^\intercal\boldsymbol{g}(\boldsymbol{x}) - 2\boldsymbol{x}^\intercal\boldsymbol{g}(\boldsymbol{x}) + 2\boldsymbol{\mu}^\intercal\boldsymbol{g}(\boldsymbol{x})] \\
&= \mathbb{E}[\|\boldsymbol{x} - \boldsymbol{\mu}\|^2] + \mathbb{E}[\|\boldsymbol{g}(\boldsymbol{x})\|^2] - 2\mathbb{E}[(\boldsymbol{x} - \boldsymbol{\mu})^\intercal\,\boldsymbol{g}(\boldsymbol{x})]
\end{aligned}$$

Furthermore, from Stein's lemma, we know that:

$$\mathbb{E}[(\boldsymbol{x} - \boldsymbol{\mu})^\intercal\,\boldsymbol{g}(\boldsymbol{x})] = \mathbb{E}\Big[\sum_{j=1}^{p} \frac{\partial}{\partial x_j} g_j(\boldsymbol{x})\Big]$$

where $x_j$ is the $j$th element of $\boldsymbol{x}$ and $g_j(\boldsymbol{x})$ is the $j$th element of $\boldsymbol{g}(\boldsymbol{x})$.

i. (1 Points) Find $\mathbb{E}[\|\boldsymbol{x} - \boldsymbol{\mu}\|^2]$.

ii. (1 Points) Find $\mathbb{E}[\|\boldsymbol{g}(\boldsymbol{x})\|^2]$. (Hint: your answer will include $\mathbb{E}[\|\boldsymbol{x}\|^{-2}]$.)

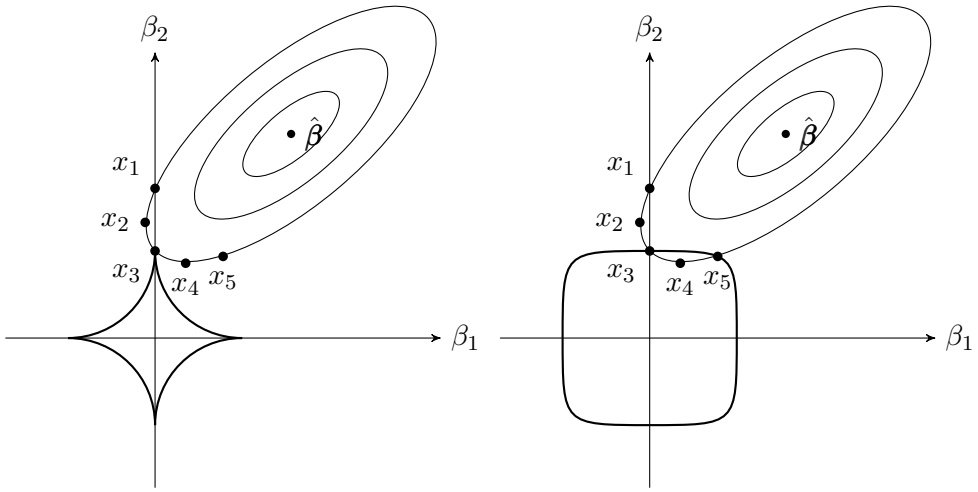
iii. (1 Points) Show that:

$$\frac{\partial}{\partial x_j} g_j(\boldsymbol{x}) = (p-2)\frac{\|\boldsymbol{x}\|^2 - 2x_j^2}{\|\boldsymbol{x}\|^4}$$

iv. (1 Points) What is the resulting mean squared error. (Hint: your answer will include $\mathbb{E}[\|\boldsymbol{x}\|^{-2}]$.)

(e) (4 Points) Qualitatively compare these estimators, noting any similarities between them. How does regularization affect an estimator's bias and variance? Which estimator would you choose to approximate $\boldsymbol{\mu}$ from real data about which you have no prior knowledge? How does the data dimensionality $p$ affect your answer?

2. $\ell_q$ **regression.**




The figures show the cost function components of the $\ell_q$-regression problems with $q = 0.5$ (left) and $q = 4$ (right).

(a) (5 Points) Does one/none/both of the cost functions encourage sparse estimates? If so, which one? Explain your answer.

(b) (5 Points) Which of the points $x_1, \ldots, x_5$ would achieve the smallest cost under the $\ell_q$-constrained least squares cost function? For each of the two cases, name the respective point and give a brief explanation for your answer.

(c) (5 Points) Write down the loss function with $q = 0.5$ and $q = 4$ respectively. Which one you think is easier to solve numerically and why?

3. **Logistic regression.**

In class, we will learn about MLE of parameters in logistic regression. For a given data $x \in \mathbb{R}^p$, the probability of $Y$ being 1 in logistic regression is

$$P(Y = 1|X = x) = \frac{\exp(w_0 + x^T w)}{1 + \exp(w_0 + x^T w)}, \tag{1}$$

where $w_0$ and $w = (w_1, w_2, ..., w_p)^T$ are model parameters. In this problem, we consider the maximum a posteriori setting, where we put a Gaussian prior on the parameters:

$$w_i \sim \mathcal{N}(\mu, 1)$$

for $i = 0, 1, 2, ..., p$.

(a) (10 Points) Assuming you are given a dataset with $n$ training examples and $p$ features, write down a formula for the conditional log posterior likelihood of the training data in terms of $\mu$, the class labels $y^{(i)}$, the features $x_1^{(i)}, \ldots, x_p^{(i)}$, and the parameters $w_0, w_1, \ldots, w_p$, where the superscript $(i)$ denotes the sample index. This will be your objective function for gradient ascent.

(b) (10 Points) Compute the partial derivative of the objective with respect to $w_0$, to an arbitrary $w_i$ and $\mu$, i.e. derive $\partial f/\partial w_0$, $\partial f/\partial w_i$, $\partial f/\partial \mu$ where $f$ is the objective that you provided above. Use (1) to simplify the formula. What is the MAP estimation of $w_0$ and $w$ with $\mu$ given?

4. **MLE and MAP, another interpretation of the penalty term in ridge regression**
Suppose we observe N data samples $\{(x_i, y_i)\}_{i=1}^N$, where $y_i$ is generated by the following rule:

$$y_i = x_i^\mathsf{T}\beta + \epsilon_i,$$

where $x_i, \beta \in \mathbb{R}^d$, $y_i \in \mathbb{R}$, and $\epsilon_i$ is an i.i.d random noise drawn from the Gaussian Distribution:

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

with a known constant $\sigma$. We further denote $Y = [y_1, y_2, ..., y_N]^\mathsf{T}$ and $X = [x_1, x_2, ..., x_N]^\mathsf{T}$.

Now, we are interested in estimating $\beta$ from the observed data.

(a) (5 Points) What are the dimensions of $Y$ and $X$? of Derive the likelihood function $\mathcal{L}(\beta)$.

(b) (5 Points) Show that the MLE estimator $\hat{\beta}_{\text{MLE}}$ of $\beta$ is equivalent to the solution of the following linear regression problem:

$$\min_\beta \frac{1}{2}\|Y - X\beta\|_2^2 \tag{2}$$

(c) (5 Points) Now we suppose $\beta$ is not a deterministic parameter, but a random variable having a Gaussian prior distribution:

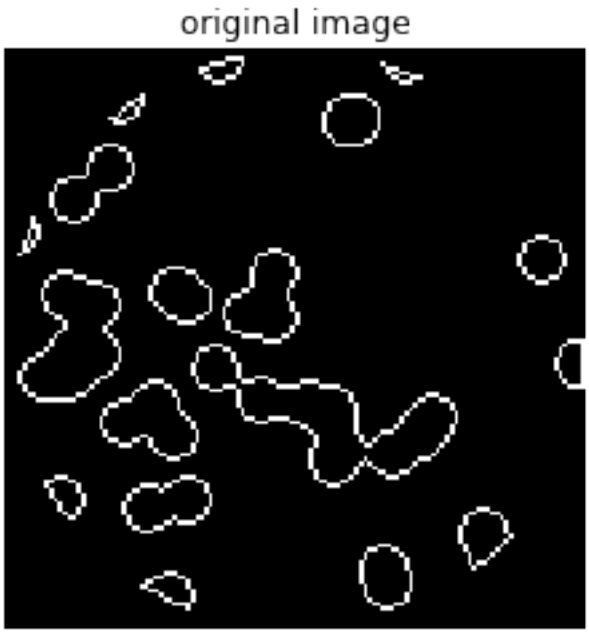
$$p(\beta) \sim \mathcal{N}(0, \frac{\sigma^2}{2\lambda}I),$$

where $I$ is a $d \times d$ identity matrix and $\lambda > 0$ is a known parameter. Show that the MAP estimation $\hat{\beta}_{\text{MAP}}$ of $\beta$ is equivalent to the solution of the following ridge regression problem:

$$\min_\beta \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2 \tag{3}$$

(d) (5 Points) Refer to the closed form solutions of (2) and (3) in the lecture slides, what might be an issue of $\hat{\beta}_{\text{ml}}$ if $d \gg N$? How can $\hat{\beta}_{\text{map}}$ possibly address it?

5. **Lasso vs ridge regression, compressed sensing** In this coding exercise, we will see how to lasso and ridge regression methods to reconstruct an image from a set of corrupted data. The dataset contains of 2 txt file, `hw1_Q5_X.txt` and `hw1_Q5_Y.txt`. The original image is shown as the following:




The dimension of the figure is $128 \times 128 = 16384$, but we only have 2304 of them available. In class, this is identified as the "high-dimensional" problem: we are using 2304 samples to predict the value of 16384. We will experiment with different values of $\lambda \in \{0.000001, 0.0001, 0.01, 0.1, 1\}$.

(a) (10 Points) Using the 5 penalty values of $\lambda$ mentioned above to fit the ridge regression model. You will have the coefficient $\beta \in \mathbb{R}^{16384}$. Reshape $\beta$ into $128 \times 128$ matrix, and plot the matrix as in the original image. Which coefficient yields the best result?

(b) (10 Points) Using the 5 penalty values of $\lambda$ mentioned above to fit the lasso model. You will have the coefficient $\beta \in \mathbb{R}^{16384}$. Reshape $\beta$ into $128 \times 128$ matrix, and plot the matrix as in the original image. Which coefficient yields the best result?

(c) (5 Points) Comparing the best results from lasso and ridge regression respectively, what do you find?