

- Do not open the exam before you are instructed to do so.
- The exam is closed book, closed notes except your one-page cheat sheet.
- Usage of electronic devices is forbidden. If we see you using an electronic device (phone, laptop, etc.) you will get a zero.
- You have 1 hour and 20 minutes.
- Write your initials at the top right of each page (e.g., write “BR” if you are Ben Recht).
- Mark your answers on the exam itself in the space provided. Do not attach any extra sheets.

First name	
Last name	
SID	
First and last name of student to your left	
First and last name of student to your right	

## Q1. [25 pts] Clip Loss

Let  $S = \{(x_1, y_1), \dots, (x_n, y_n)\}$  be a set of  $n$  points sampled i.i.d. from a distribution  $\mathcal{D}$ . This is the training set with  $x_i \in \mathbb{R}^d$  being the features and  $y_i \in \{-1, 1\}$  being the labels. Define the *clip loss* of a linear classifier  $w \in \mathbb{R}^d$  as

$$\text{loss}(w^\top x, y) = \text{clip}(yw^\top x)$$

Where clip is the function

$$\text{clip}(z) = \begin{cases} 1 & \text{if } z < 0 \\ 0 & \text{if } z \geq 1 \\ 1 - z & \text{otherwise.} \end{cases}$$

For any  $d$ -dimensional vector  $w$ , define the *risk* of  $w$  as

$$R[w] = \mathbb{E}_{\mathcal{D}}[\text{loss}(w^\top x, y)],$$

and the *empirical risk* of  $w$  as

$$R_S[w] = \frac{1}{n} \sum_{i=1}^n \text{loss}(w^\top x_i, y_i).$$

**(a)** [5 pts] Is the function clip convex? If you would like, you can justify your answer by drawing a picture.

**(b)** [5 pts] Show that if  $R_S[w] = 0$  and  $\|w\|_2^2 < 1$ , then the margin of the hyperplane defined by  $w$  is greater than 1.

(c) [5 pts] Prove that  $\mathbb{E}_S[R_S[w]] = R[w]$ .

(d) [5 pts] Prove that  $\text{Var}(R_S[w]) \leq \frac{1}{n}$ .

(e) [5 pts] Is it possible to have a  $w$  such that  $R_S[w] = 0$ , but  $R[w] > 0$ ? Justify your answer.

## Q2. [25 pts] Regularization

We consider here a discriminative approach for solving the classification problem illustrated in Figure 1.

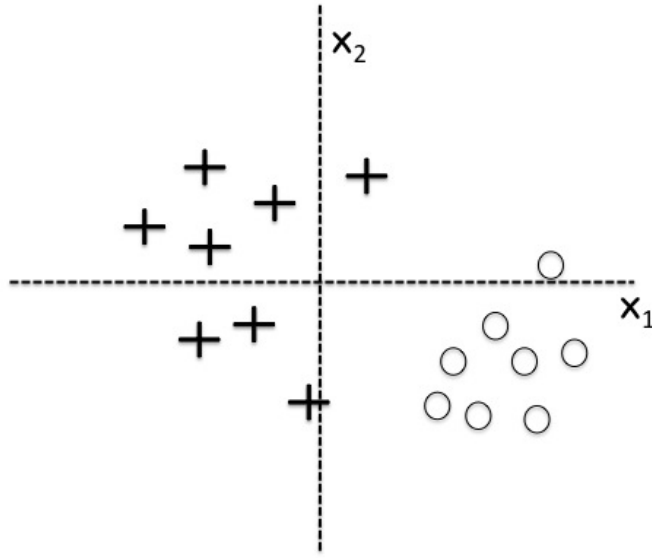


Figure 1: The two-dimensional labeled training set, where ‘+’ corresponds to class  $y = 1$  and ‘O’ corresponds to class  $y = 0$ .

Suppose we attempt to solve the binary classification task depicted in Figure 1 with the simple linear logistic regression model

$$P(y = 1|x, \mathbf{w}) = g(w_0 + w_1x_1 + w_2x_2) = \frac{1}{1 + \exp(-w_0 - w_1x_1 - w_2x_2)}$$

Notice that training data can be separated with *zero* training error with a linear separator.

Consider training regularized logistic regression model where we try to maximize

$$\sum_{i=1}^n \log P(y_i|x_i, w_0, w_1, w_2) - Cw_j^2$$

for very large  $C$ . The regularization penalties used in penalized conditional log-likelihood estimation are  $-Cw_j^2$  where  $j \in \{0, 1, 2\}$ . In other words, only one of the parameters is regularized in each case. Given the training data in Figure 1, how does the training error change with regularization of each parameter  $w_j$ ? State whether the training error increases or stays the same (zero) for each  $w_j$  for large  $C$ . Provide a brief justification for each of your answers.

(a) [5 pts] By regularizing  $w_2$

(b) [5 pts] By regularizing  $w_1$

(c) [5 pts] By regularizing  $w_0$

Now suppose we want to regularize *both*  $w_1$  and  $w_2$ . This means we want to maximize the penalized log-likelihood

$$\sum_{i=1}^n \log P(y_i|x_i, w_0, w_1, w_2) - C(w_1^2 + w_2^2)$$

Consider again the problem in Figure 1 and the same linear logistic regression model  $P(y = 1|x, \mathbf{w}) = g(w_0 + w_1x_1 + w_2x_2)$ .

(d) [5 pts] For very large  $C$ , which value(s) do you expect  $w_0$  to take? Explain briefly. (Note that the number of points from each class is the same.) (You can give a range of values for  $w_0$  if you deem necessary).

(e) [5 pts] Assume that we obtain more data points from the '+' class that corresponds to  $y = 1$  so that the class labels become unbalanced. Again for very large  $C$ , with the same regularization for  $w_1$  and  $w_2$  as above, which value(s) do you expect  $w_0$  to take? Explain briefly. (You can give a range of values for  $w_0$  if you deem necessary).

### Q3. [25 pts] Bias-variance tradeoff in linear regression

Recall the statistical model for linear regression from lecture. Fix a set of points  $x_1, x_2, \dots, x_n \in \mathbb{R}^d$  and an unknown regressor  $\theta_* \in \mathbb{R}^d$ . Suppose we observe  $y_1, y_2, \dots, y_n \in \mathbb{R}$  via the process

$$y_i = x_i^\top \theta_* + \varepsilon_i,$$

where the noise vector  $\varepsilon := \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \in \mathbb{R}^n$  satisfies

$$\mathbb{E}\varepsilon = 0, \quad \text{Cov}(\varepsilon) = \sigma^2 I_n.$$

Using the convention from lecture, we write

$$X := \begin{bmatrix} -x_1^\top - \\ -x_2^\top - \\ \vdots \\ -x_n^\top - \end{bmatrix} \in \mathbb{R}^{n \times d}, \quad Y := \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n.$$

With this notation, our statistical model is equivalent to

$$Y = X\theta_* + \varepsilon.$$

You may assume throughout this problem that the matrix  $X^\top X$  is invertible. Recall the two least-squares estimators we studied in lecture

$$\hat{\theta}_{\text{ols}} = \arg \min_{\theta \in \mathbb{R}^d} \frac{1}{2} \|X\theta - Y\|_2^2 \quad (\text{OLS})$$

$$\hat{\theta}_{\text{ridge}} = \arg \min_{\theta \in \mathbb{R}^d} \frac{1}{2} \|X\theta - Y\|_2^2 + \frac{\lambda}{2} \|\theta\|_2^2 \quad (\text{Ridge}).$$

For the Ridge estimator, you can assume that  $\lambda > 0$  is known and fixed throughout the problem.

(a) [5 pts] Write down the closed form solutions for  $\hat{\theta}_{\text{ols}}$  and  $\hat{\theta}_{\text{ridge}}$ . Simply state the answer, no need to rederive it.

- (b) [5 pts] Let  $\hat{\theta} \in \mathbb{R}^d$  denote any estimator of  $\theta_*$ . In the context of this problem, an estimator  $\hat{\theta} = \hat{\theta}(X, Y)$  is any function which takes the data  $X$  and a realization of  $Y$ , and computes a guess of  $\theta_*$ .

Define the MSE (mean squared error) of the estimator  $\hat{\theta}$  as

$$\text{MSE}(\hat{\theta}) := \mathbb{E} \|\hat{\theta} - \theta_*\|_2^2.$$

Above, the expectation is taken w.r.t. the randomness inherent in  $\varepsilon$ . Define  $\hat{\mu} := \mathbb{E}\hat{\theta}$ . Show that, as we did in lecture, the MSE decomposes as such

$$\text{MSE}(\hat{\theta}) = \|\hat{\mu} - \theta_*\|_2^2 + \mathbf{Tr}(\text{Cov}(\hat{\theta})).$$

*Hint:* Expectation and trace commute, so  $\mathbb{E} \mathbf{Tr}(A) = \mathbf{Tr}(\mathbb{E}A)$  for any square matrix  $A$ .

- (c) [5 pts] Show that

$$\mathbb{E}\hat{\theta}_{\text{ols}} = \theta_*, \quad \mathbb{E}\hat{\theta}_{\text{ridge}} = (X^\top X + \lambda I_d)^{-1} X^\top X \theta_*.$$

That is,  $\hat{\theta}_{\text{ols}}$  is an *unbiased* estimator of  $\theta_*$ , whereas  $\hat{\theta}_{\text{ridge}}$  is a *biased* estimator of  $\theta_*$ .

- (d) [10 pts] Let  $\gamma_1 \geq \gamma_2 \geq \dots \geq \gamma_d$  denote the  $d$  eigenvalues of the matrix  $X^\top X$  arranged in non-increasing order. First, argue that the smallest eigenvalue,  $\gamma_d$ , is positive (i.e.  $\gamma_d > 0$ ). Then, show that

$$\mathbf{Tr}(\text{Cov}(\hat{\theta}_{\text{ols}})) = \sigma^2 \sum_{i=1}^d \frac{1}{\gamma_i}, \quad \mathbf{Tr}(\text{Cov}(\hat{\theta}_{\text{ridge}})) = \sigma^2 \sum_{i=1}^d \frac{\gamma_i}{(\gamma_i + \lambda)^2}.$$

Finally, use these formulas to conclude that

$$\mathbf{Tr}(\text{Cov}(\hat{\theta}_{\text{ridge}})) < \mathbf{Tr}(\text{Cov}(\hat{\theta}_{\text{ols}})).$$

**NOTE:** The inequality above was *incorrectly* stated on the exam. It is fixed in the solutions. Because of this, we are awarding one free point for every student regardless of whether or not they attempted this question. *Hint:* For the Ridge variance, consider writing  $X^\top X$  in terms of its eigen-decomposition  $U\Sigma U^\top$ .



## Q4. [15 pts] Nonlinear regression

In this problem, we'll come up with a method for estimating the acceleration on an object from noisy measurements of its position. We have noisy observations of the position of an object ( $p(t)$ ) (in 1D) at  $n$  points in time:

$$p_v = [p_0, \dots, p_n]$$
$$t_v = [t_0, \dots, t_n].$$

$p_i$  is a noisy measurement of  $p(t_i)$ . We believe the object is undergoing constant acceleration ( $a = p''(t)$ ).

Note that this means that  $p(t) = \frac{1}{2}at^2 + v(0)t + p(0)$  where  $v(0)$  is the initial velocity and  $p(0)$  is the initial position of the object.

- (a) [5 pts] Suppose that  $v(0) = 0$  and  $p(0) = 0$ . Write down a least squares problem to estimate  $a$  from the noisy measurements  $p_v$ .
- (b) [5 pts] Solve the optimization problem and give your answer in terms of  $p_v$  and  $t_v$ .
- (c) [5 pts] Now suppose the initial position and the velocity of the object are unknown; write down (but don't solve) a least squares problem to estimate the initial position and velocity along with the acceleration of the object.