Homework #2

CSE 446/546: Machine Learning
Profs. Jamie Morgenstern and Simon Du
Due: **Wednesday** November 3, 2021 11:59pm
A: 96 points, **B:** 29 points

Please review all homework guidance posted on the website before submitting to GradeScope. Reminders:

- Make sure to read the "What to Submit" section following each question and include all items.
- Please provide succinct answers and supporting reasoning for each question. Similarly, when discussing experimental results, concisely create tables and/or figures when appropriate to organize the experimental results. All explanations, tables, and figures for any particular part of a question must be grouped together.
- For every problem involving generating plots, please include the plots as part of your PDF submission.
- When submitting to Gradescope, please link each question from the homework in Gradescope to the location of its answer in your homework PDF. Failure to do so may result in deductions of up to *[5 points]* . For instructions, see `https://www.gradescope.com/get_started#student-submission`.
- Please recall that B problems, indicated in boxed text, are only graded for 546 students, and that they will be weighted at most 0.2 of your final GPA (see the course website for details). In Gradescope, there is a place to submit solutions to A and B problems separately. You are welcome to create a single PDF that contains answers to both and submit the same PDF twice, but associate the answers with the individual questions in Gradescope.
- If you collaborate on this homework with others, you must indicate who you worked with on your homework. Failure to do so may result in accusations of plagiarism.
- For every problem involving code, please include the code as part of your PDF for the PDF submission *in addition to* submitting your code to the separate assignment on Gradescope created for code. Not submitting all code files will lead to a deduction of *[1 point]* .
- Please indicate your final answer to each question by placing a box around the main result(s). To do this in LaTeX, one option is using the `boxed` command.

Not adhering to these reminders may result in point deductions.

Short Answer and "True or False" Conceptual questions

A1. The answers to these questions should be answerable without referring to external materials. Briefly justify your answers with a few words.

a. *[2 points]* Suppose that your estimated model for predicting house prices has a large positive weight on the feature `number of bathrooms`. If we remove this feature and refit the model, will the new model have a strictly higher error than before? Why?

b. *[2 points]* Compared to L2 norm penalty, explain why a L1 norm penalty is more likely to result in sparsity (a larger number of 0s) in the weight vector.

c. *[2 points]* In at most one sentence each, state one possible upside and one possible downside of using the following regularizer: $\left(\sum_i |w_i|^{0.5}\right)$.

d. *[1 point]* True or False: If the step-size for gradient descent is too large, it may not converge.

e. *[2 points]* In your own words, describe why stochastic gradient descent (SGD) works, even though only a small portion of the data is considered at each update.

f. *[2 points]* In at most one sentence each, state one possible advantage of SGD over GD (gradient descent), and one possible disadvantage of SGD relative to GD.

What to Submit:

- **Part d:** True or False.
- **Parts a-f:** Brief (2-3 sentence) explanation.

Answer:
Part a:
No it will not be strictly higher than before. If the attribute is linearly independent then it could be worse. However, it is likely the case that the feature "number of bathrooms" is really not an independent feature and in that case, the weights are distributed to other dependent features. In this new model, the collinearity is diminished and it will not be strictly worse

Part b:
For L1 norm, the penalty term is absolute while for L2 norm, the penalty term is squared. This implies that for L1 norm, the nature of its constraint results in coefficients that are exactly zero whereas for L2 norm, squaring small values already make it smaller, allow it to achieve the minimum square error without needing to make the values zero.

Part c:
Upside: It promotes sparsity better than L1.
Downside: It is non-convex

Part d:
True

Part e:
Stochastic gradient descents takes in feedback at each randomly selected point, or set of points, instead of computing the true gradient using all points as seen in gradient descent. This is less computationally expensive. Additionally, if the true function is relatively monotonic, analyzing one point at a time and following its slope allows it to reach a point very close to the actual minimum

Part f:
Advantage of SGD over GD: Less computationally expensive
Disadvantage of SGD over GD: Takes longer time and more steps to converge to the minimum

Convexity and Norms

A2. A *norm* $\|\cdot\|$ over $\mathbb{R}^n$ is defined by the properties: (i) non-negativity: $\|x\| \geq 0$ for all $x \in \mathbb{R}^n$ with equality if and only if $x = 0$, (ii) absolute scalability: $\|a\,x\| = |a|\,\|x\|$ for all $a \in \mathbb{R}$ and $x \in \mathbb{R}^n$, (iii) triangle inequality: $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in \mathbb{R}^n$.

a. *[3 points]* Show that $f(x) = (\sum_{i=1}^n |x_i|)$ is a norm. (Hint: for (iii), begin by showing that $|a+b| \leq |a|+|b|$ for all $a, b \in \mathbb{R}$.)

b. *[2 points]* Show that $g(x) = \left(\sum_{i=1}^n |x_i|^{1/2}\right)^2$ is not a norm. (Hint: it suffices to find two points in $n = 2$ dimensions such that the triangle inequality does not hold.)

Context: norms are often used in regularization to encourage specific behaviors of solutions. If we define $\|x\|_p := (\sum_{i=1}^n |x_i|^p)^{1/p}$ then one can show that $\|x\|_p$ is a norm for all $p \geq 1$. The important cases of $p = 2$ and $p = 1$ correspond to the penalty for ridge regression and the lasso, respectively.

What to Submit:

- **Parts a, b:** Proof.

Answers:
Part a:

$$|a+b|^2 = a^2 + b^2 + 2ab \leq a^2 + b^2 + 2|a||b| = (|a| + |b|)^2$$
$$\Rightarrow |a+b| \leq |a| + |b|$$
$$\text{since } |a+b| \geq 0, |a| + |b| \geq 0.\square$$

Let's check whether f(x) holds the properties of non-negativity, absolute scalability and triangle inequality to show that it is a norm:

Non-negativity:
$f(x) \geq 0$ since $\forall i : |x_i| \geq 0$. Now, $|x_i| = 0$ iff $x_i = 0$, so $f(x) = 0$ iff $x = 0$. $\square$

Absolute Scalability:
$\forall a \in \mathbb{R}, x \in \mathbb{R}^n : f(ax) = \sum_{i=1}^n |ax_i| = \sum_{i=1}^n |a||x_i| = |a|f(x).\square$

Triangle inequality:
$\forall x, y \in \mathbb{R}^n : f(x) + f(y) = \sum_{i=1}^n |x_i| + \sum_{i=1}^n |y_i| = \sum_{i=1}^n |x_i| + |y_i| \geq \sum_{i=1}^n |x_i + y_i| = f(x+y)$, where the second to last inequality follows from the triangle inequality for absolute values on $\mathbb{R}$ above. $\square$

Part b:
Consider $x = [0,1]^T, y = [1,0]^T$. Then $g(x) + g(y) = 1^2 + 1^2 = 2, g(x+y) = (1+1)^2 = 4$, so for these x, y: $g(x+y) > g(x) + g(y)$ and the triangle inequality does not hold which means g is not a norm. $\square$

B1. A set $A \subseteq \mathbb{R}^n$ is *convex* if $\lambda x + (1-\lambda)y \in A$ for all $x, y \in A$ and $\lambda \in [0,1]$. Let $\|\cdot\|$ be a norm.

a. *[3 points]* Show that $f(x) = \|x\|$ is a convex function.

b. *[3 points]* Show that $\{x \in \mathbb{R}^n : \|x\| \leq 1\}$ is a convex set.

c. *[2 points]* Draw a picture of the set $\{(x_1, x_2) \ : \ g(x_1, x_2) \leq 4\}$ where $g(x_1, x_2) = \left(|x_1|^{1/2} + |x_2|^{1/2}\right)^2$. (This is the function considered in 1b above specialized to $n = 2$.) We know g is not a norm. Is the defined set convex? Why not?

Context: It is a fact that a function f defined over a set $A \subseteq \mathbb{R}^n$ is convex if and only if the set $\{(x, z) \in \mathbb{R}^{n+1} : z \geq f(x), x \in A\}$ is convex. Draw a picture of this for yourself to be sure you understand it.

What to Submit:

- **Parts a, b:** Proof.
- **Part c:** A picture of the set, and 1-2 sentence explanation.

Answers:
Part a:
For $\lambda \in [0,1]$, $x, y \in \mathbb{R}^n$, by using triangle inequality and absolute scalability we see that:

$$\|\lambda x + (1-\lambda)y\| \leq \|\lambda x\| + \|(1-\lambda)y\| \leq \lambda\|x\| + (1-\lambda)\|y\| \qquad \square$$

Part b:
Assume that $x, y \in [0,1]$ and $||x|| < ||y||$. By applying the definition we get:

$$||\lambda x + (1-\lambda)y|| \leq \lambda||x|| + (1-\lambda)||y|| < \lambda||x|| + (1-\lambda)||x||$$
$$||\lambda x + (1-\lambda)y|| < \lambda||x|| + ||x|| - \lambda||x|| = ||x|| \leq 1$$
$$||\lambda x + (1-\lambda)y|| < 1$$

Part c:
The set $L := \{(x, y) : g(x, y) \leq 4\}$, where $g(x, y) = (|x|^{1/2} + |y|^{1/2})^2$ is not convex. The extreme points x = (0, 4) and y = (4,0) does not follow $\lambda x + (1-\lambda)y \in L$

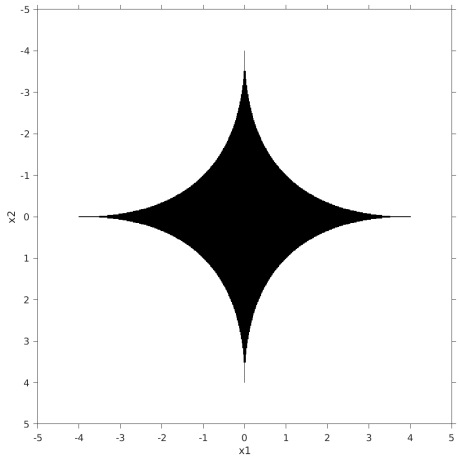

Figure 1: Plot of $\{(x, y) : g(x, y) \leq 4\}$, where $g(x, y) = (|x|^{1/2} + |y|^{1/2})^2$

B2. For $i = 1, \ldots, n$ let $\ell_i(w)$ be convex functions over $w \in \mathbb{R}^d$ (e.g., $\ell_i(w) = (y_i - w^\top x_i)^2$), $\|\cdot\|$ is any norm, and $\lambda > 0$.

a. *[3 points]* Show that

$$\sum_{i=1}^{n} \ell_i(w) + \lambda \|w\|$$

is convex over $w \in \mathbb{R}^d$ (Hint: Show that if f, g are convex functions, then $f(x) + g(x)$ is also convex.)

b. *[1 point]* Explain in one sentence why we prefer to use loss functions and regularized loss functions that are convex.

What to Submit

- **Part a:** Proof.
- **Part b:** 1-2 sentence explanation.

Answers:
Part a:
First, let's show that f(x) and g(x) are convex functions for any $x, y \in \mathbb{R}^n$:

$$\begin{aligned} f(\lambda x + (1-\lambda)y) + g(\lambda x + (1-\lambda)y) &\leq \lambda f(x) + (1-\lambda) f(y) + \lambda g(x) + (1-\lambda) g(y) \\ &= \lambda(f(x) + g(x)) + (1-\lambda)(f(y) + g(y)). \end{aligned}$$ □

$\implies$ that f(x)+g(x) is also convex.
By induction, $\sum_{i=1}^{n} f_i(x)$ is also convex because it can be broken down into the addition of two convex function sequentially. Then for any $\lambda > 0$, $0 \leq \alpha \leq 0$ and $\|w\|$ is convex, for $x, y \in \mathbb{R}^n$:

$$\lambda \|\alpha x + (1-\alpha) y\| \leq \lambda(\alpha \|x\| + (1-\alpha)\|y\|) = \alpha\lambda\|x\| + (1-\alpha)\lambda\|y\|$$

As shown above, $\lambda\|w\|$ is convex. $\implies$ $\sum_{i=1}^{n} l_i(w) + \lambda\|w\|$ is convex as a sum of two convex functions.

Part b:
Convexity in those functions would imply that every local minimum is a global minimum

Lasso on a Real Dataset

A Lasso Algorithm

Given $\lambda > 0$ and data $\left(x_i, y_i\right)_{i=1}^n$, the Lasso is the problem of solving

$$\arg \min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \sum_{i=1}^n (x_i^T w + b - y_i)^2 + \lambda \sum_{j=1}^d |w_j|$$

where λ is a regularization parameter. For the programming part of this homework, we have implemented the coordinate descent method shown in Algorithm 1 to solve the Lasso problem for you.

Algorithm 1: Coordinate Descent Algorithm for Lasso

while *not converged* **do**

$b \leftarrow \frac{1}{n} \sum_{i=1}^n \left(y_i - \sum_{j=1}^d w_j x_{i,j}\right)$

for $k \in \{1, 2, \cdots d\}$ **do**

$a_k \leftarrow 2 \sum_{i=1}^n x_{i,k}^2$

$c_k \leftarrow 2 \sum_{i=1}^n x_{i,k} \left(y_i - (b + \sum_{j \neq k} w_j x_{i,j})\right)$

$$w_k \leftarrow \begin{cases} (c_k + \lambda)/a_k & c_k < -\lambda \\ 0 & c_k \in [-\lambda, \lambda] \\ (c_k - \lambda)/a_k & c_k > \lambda \end{cases}$$

end

end

You will often apply Lasso on the same dataset for many values of λ. This is called a regularization path. One way to do this efficiently is to start at a large λ, and then for each consecutive solution, initialize the algorithm with the previous solution, decreasing λ by a constant ratio (e.g., by a factor of 2).
The smallest value of λ for which the solution $\widehat{w}$ is entirely zero is given by

$$\lambda_{max} = \max_{k=1,\ldots,d} 2 \left| \sum_{i=1}^n x_{i,k} \left(y_i - \left(\frac{1}{n} \sum_{j=1}^n y_j \right) \right) \right| \tag{1}$$

This is helpful for choosing the first λ in a regularization path.
A benefit of the Lasso is that if we believe many features are irrelevant for predicting y, the Lasso can be used to enforce a sparse solution, effectively differentiating between the relevant and irrelevant features.

Dataset

Download the training data set "crime-train.txt" and the test data set "crime-test.txt" from the course website. Store your data in your working directory, ensure you have `pandas` installed, and read in the files with the following Python code:

```
import pandas as pd
df_train = pd.read_table("crime-train.txt")
df_test = pd.read_table("crime-test.txt")
```

This stores the data as Pandas `DataFrame` objects. `DataFrame`s are similar to Numpy `array`s but more flexible; unlike `array`s, `DataFrame`s store row and column indices along with the values of the data. Each column of a `DataFrame` can also store data of a different type (here, all data are floats).
Here are a few commands that will get you working with Pandas for this assignment:

```
df.head()                    # Print the first few lines of DataFrame df.
df.index                     # Get the row indices for df.
df.columns                   # Get the column indices.
df[''foo'']                 # Return the column named ''foo''.
df.drop(''foo'', axis = 1)   # Return all columns except ''foo''.
df.values                    # Return the values as a Numpy array.
df[''foo''].values          # Grab column foo and convert to Numpy array.
df.iloc[:3,:3]               # Use numerical indices (like Numpy) to get 3 rows and cols.
```

The data consist of local crime statistics for 1,994 US communities. The response y is the rate of violent crimes reported per capita in a community. The name of the response variable is `ViolentCrimesPerPop`, and it is held in the first column of `df_train` and `df_test`. There are 95 features. These features include many variables. Some features are the consequence of complex political processes, such as the size of the police force and other systemic and historical factors. Others are demographic characteristics of the community, including self-reported statistics about race, age, education, and employment drawn from Census reports.

The dataset is split into a training and test set with 1,595 and 399 entries, respectively. The features have been standardized to have mean 0 and variance 1. We will use this training set to fit a model to predict the crime rate in new communities and evaluate model performance on the test set. As there are a considerable number of input variables and fairly few training observations, overfitting is a serious issue, and the coordinate descent Lasso algorithm may mitigate this problem during training.

The goals of this problem are threefold: (*i*) to encourage you to think about how data collection processes affect the resulting model trained from that data; (*ii*) to encourage you to think deeply about models you might train and how they might be misused; and (*iii*) to see how Lasso encourages sparsity of linear models in settings where d is large relative to n. **We emphasize that training a model on this dataset can suggest a degree of correlation between a community's demographics and the rate at which a community experiences and reports violent crime. We strongly encourage students to consider why these correlations may or may not hold more generally, whether correlations might result from a common cause, and what issues can result in misinterpreting what a model can explain.**

Applying Lasso

A3.

a. *[4 points]* Read the documentation for the original version of this dataset: `http://archive.ics.uci.edu/ml/datasets/communities+and+crime`. Report 3 features included in this dataset for which historical *policy* choices in the US would lead to variability in these features. As an example, the *number of police* in a community is often the consequence of decisions made by governing bodies, elections, and amount of tax revenue available to decision makers.

b. *[4 points]* Before you train a model, describe 3 features in the dataset which might, if found to have nonzero weight in model, be interpreted as *reasons* for higher levels of violent crime, but which might actually be a *result* rather than (or in addition to being) the cause of this violence.

Now, we will run the Lasso solver. Begin with $\lambda = \lambda_{\max}$ defined in Equation (1). Initialize all weights to 0. Then, reduce λ by a factor of 2 and run again, but this time initialize $\hat{w}$ from your $\lambda = \lambda_{\max}$ solution as your initial weights, as described above. Continue the process of reducing λ by a factor of 2 until $\lambda < 0.01$. For all plots use a log-scale for the λ dimension (Tip: use `plt.xscale('log')`).

c. *[4 points]* Plot the number of nonzero weights of each solution as a function of λ.

d. *[4 points]* Plot the regularization paths (in one plot) for the coefficients for input variables `agePct12t29`, `pctWSocSec`, `pctUrban`, `agePct65up`, and `householdsize`.

e. *[4 points]* On one plot, plot the squared error on the training and test data as a function of λ.

f. *[4 points]* Sometimes a larger value of λ performs nearly as well as a smaller value, but a larger value will select fewer variables and perhaps be more interpretable. Inspect the weights $\hat{w}$ for $\lambda = 30$. Which feature had the largest (most positive) Lasso coefficient? What about the most negative? Discuss briefly.

g. *[4 points]* Suppose there was a large negative weight on `agePct65up` and upon seeing this result, a politician suggests policies that encourage people over the age of 65 to move to high crime areas in an effort to reduce crime. What is the (statistical) flaw in this line of reasoning? (Hint: fire trucks are often seen around burning buildings, do fire trucks cause fire?)

What to Submit:

- **Parts a, b:** 1-2 sentence explanation.
- **Part c:** Plot 1.
- **Part d:** Plot 2.
- **Part e:** Plot 3.
- **Parts f, g:** Answers and 1-2 sentence explanation.
- **Code** on Gradescope through coding submission.

Answers:
Part a:
OfficAssgnDrugUnits: number of officers assigned to a special drug units. Type: numeric(decimal)
NumKindsDugSeiz: number of different kinds of drugs seized. Type: numeric(decimal)
LemasPctOfficDrugUn: percent of officers assigned to drug units. Type: numeric(decimal)

Part b:
LemasTotReqPerPop: Total requests for police per 100k population. Type: numeric(decimal)
LemasTotalReg: Total requests for police. Type: numeric(decimal)
PolicReqPerOfficg: Total requests for police per police officer. Type: numeric(decimal)

Part c/d/e:

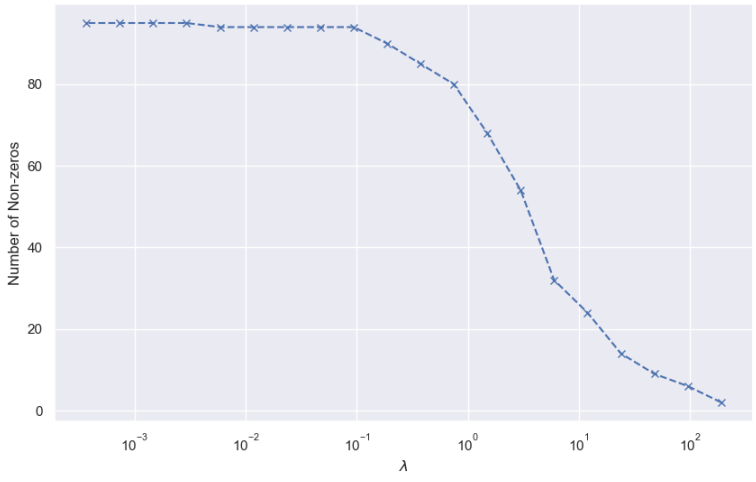

Figure 2: Number of non-zeros in $\widehat{w}_j$ with Lasso with different λ values ($\lambda_{max} = 191.98$

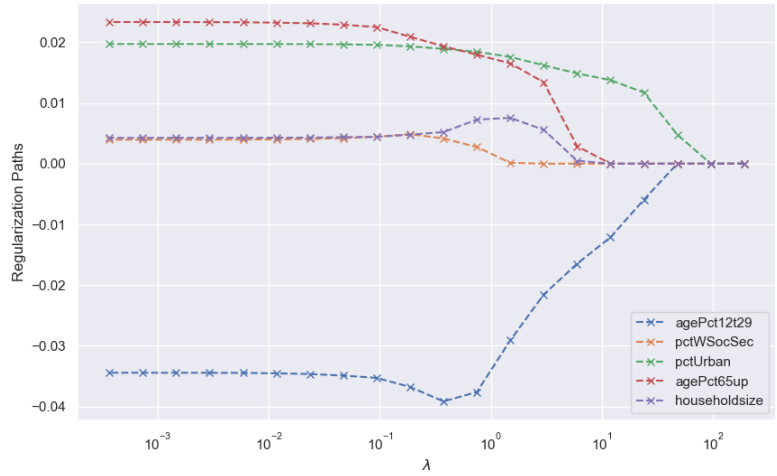

Figure 3: Regularization paths for various coefficients with Lasso using different *lambda* values

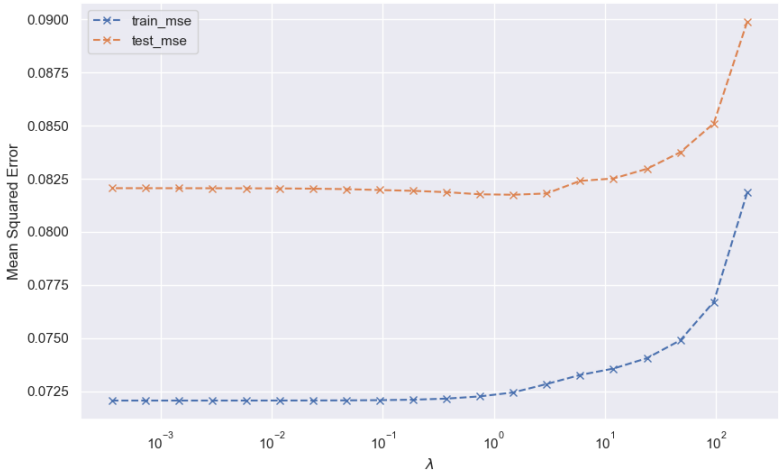

Figure 4: training and test MSE with Lasso using different λ values

Part f:
Largest Lasso Coefficient: PCTIlleg - $\lambda_{PctIlleg} = 0.068$
Smallest Lasso Coefficient: PCTKids2Par - $\lambda_{PetKids2Par} = -0.070$

Part g:
Correlation is not the same as casuality. Even though we might see that there are a low number of over 65 in high crime areas as seen in agePCT65up, it does not mean that people over 65 decreases the crime rate. It is more likely that people over the age of 65 would move out of areas with high crime rates and into areas with lower crime rates.

Logistic Regression

Binary Logistic Regression

A4. Here we consider the MNIST dataset, but for binary classification. Specifically, the task is to determine whether a digit is a 2 or 7. Here, let $Y = 1$ for all the "7" digits in the dataset, and use $Y = -1$ for "2". We will use regularized logistic regression. Given a binary classification dataset $\{(x_i, y_i)\}_{i=1}^n$ for $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$ we showed in class that the regularized negative log likelihood objective function can be written as

$$J(w, b) = \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-y_i(b + x_i^T w))) + \lambda ||w||_2^2$$

Note that the offset term b is not regularized. For all experiments, use $\lambda = 10^{-1}$. Let $\mu_i(w, b) = \frac{1}{1+\exp(-y_i(b+x_i^T w))}$.

a. *[8 points]* Derive the gradients $\nabla_w J(w, b)$, $\nabla_b J(w, b)$ and give your answers in terms of $\mu_i(w, b)$ (your answers should not contain exponentials).

b. *[8 points]* Implement gradient descent with an initial iterate of all zeros. Try several values of step sizes to find one that appears to make convergence on the training set as fast as possible. Run until you feel you are near to convergence.

   (a) For both the training set and the test, plot $J(w, b)$ as a function of the iteration number (and show both curves on the same plot).

   (b) For both the training set and the test, classify the points according to the rule $\text{sign}(b + x_i^T w)$ and plot the misclassification error as a function of the iteration number (and show both curves on the same plot).

   Reminder: Make sure you are only using the test set for evaluation (not for training).

c. *[7 points]* Repeat (b) using stochastic gradient descent with a batch size of 1. Note, the expected gradient with respect to the random selection should be equal to the gradient found in part (a). Show both plots described in (b) when using batch size 1. Take careful note of how to scale the regularizer.

d. *[7 points]* Repeat (b) using stochastic gradient descent with batch size of 100. That is, instead of approximating the gradient with a single example, use 100. Note, the expected gradient with respect to the random selection should be equal to the gradient found in part (a).

What to Submit

- **Part a:** Proof
- **Part b:** Separate plots for b(i) and b(ii).
- **Part c:** Separate plots for c which reproduce those from b(i) and b(ii) for this case.
- **Part d:** Separate plots for c which reproduce those from b(i) and b(ii) for this case.
- **Code** on Gradescope through coding submission.

Answers:
Part a:

Note that $\exp(-y_i(b + x_i^T w)) = \frac{1}{\mu_i(w,b)} - 1$. Then,

$$\begin{aligned}
\nabla_w J(w, b) &= \frac{1}{n} \sum_i \nabla_w \log(1 + \exp(-y_i(b + x_i^T w))) + \nabla_w \lambda \|w\|^2 \\
&= \frac{1}{n} \sum_i \mu_i(w, b)(\frac{1}{\mu_i(w, b)} - 1)(-y_i)x_i + 2\lambda w
\end{aligned}$$

So,

$$\boxed{\nabla_w J(w,b) = \frac{1}{n}\sum_i (\mu_i(w,b) - 1)(y_i)x_i + 2\lambda w}$$

Now,

$$\begin{aligned}\nabla_b J(w,b) &= \frac{1}{n}\sum_i \nabla_b \log(1 + \exp(-y_i(b + x_i^T w))) + \nabla_b \lambda \|w\|^2 \\ &= \frac{1}{n}\sum_i \mu_i(w,b)(\frac{1}{\mu_i(w,b)} - 1)(-y_i)\end{aligned}$$

So,

$$\boxed{\nabla_b J(w,b) = \frac{1}{n}\sum_i (\mu_i(w,b) - 1)y_i} \quad \square$$

Part b:

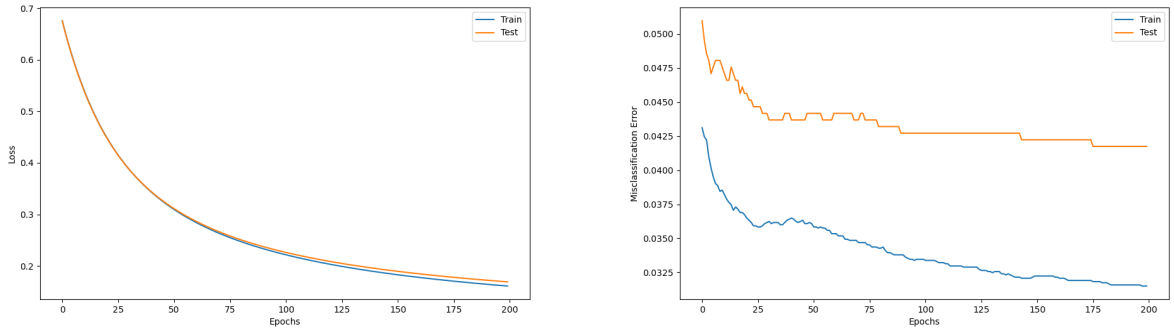

Figure 5: Problem A4.b Left: A4.bi, Right: A4.bii (Plots for Gradient Descent)

Part c:

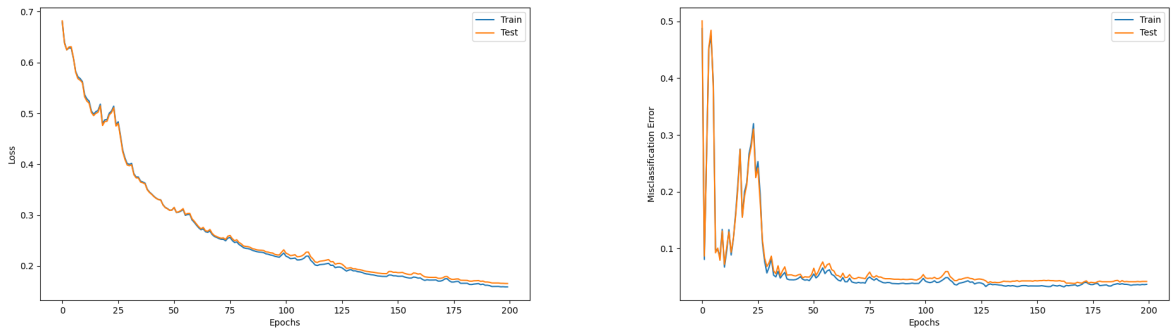

Figure 6: Problem A4.c Left: A4.ci, Right: A4.cii (Plots for Stochastic Gradient Descent with batch 1.)

Part d:

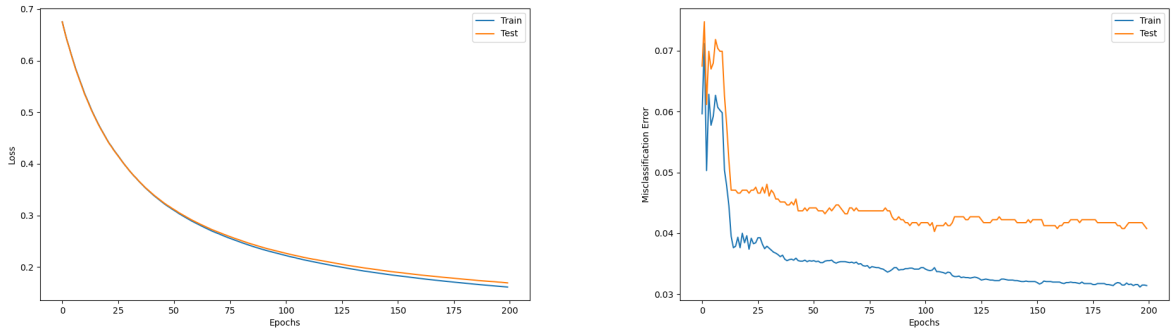

Figure 7: Problem A4.d Left: A4.di, Right: A4.dii (Plots for Stochastic Gradient Descent with batch 100.)

Ridge Regression on MNIST

These problems were moved from HW1 and are reproduced identically here. If you already started these, you may wish to reuse your work from HW1.

A5. In this problem we will implement a regularized least squares classifier for the MNIST data set. The task is to classify handwritten images of numbers between 0 to 9.

You are **NOT** allowed to use any of the pre-built classifiers in `sklearn`. Feel free to use any method from `numpy` or `scipy`. **Remember:** if you are inverting a matrix in your code, you are probably doing something wrong (Hint: look at `scipy.linalg.solve`).

Each example has features $x_i \in \mathbb{R}^d$ (with $d = 28 * 28 = 784$) and label $z_j \in \{0, \ldots, 9\}$. You can visualize a single example x_i with `imshow` after reshaping it to its original 28×28 image shape (and noting that the label z_j is accurate). We wish to learn a predictor $\widehat{f}$ that takes as input a vector in $\mathbb{R}^d$ and outputs an index in $\{0, \ldots, 9\}$. We define our training and testing classification error on a predictor f as

$$\widehat{\epsilon}_{\text{train}}(f) = \frac{1}{N_{\text{train}}} \sum_{(x,z) \in \text{Training Set}} \mathbf{1}\{f(x) \neq z\}$$

$$\widehat{\epsilon}_{\text{test}}(f) = \frac{1}{N_{\text{test}}} \sum_{(x,z) \in \text{Test Set}} \mathbf{1}\{f(x) \neq z\}$$

We will use one-hot encoding of the labels: for each observation (x, z), the original label $z \in \{0, \ldots, 9\}$ is mapped to the standard basis vector e_{z+1} where e_i is a vector of size k containing all zeros except for a 1 in the i^{th} position (positions in these vectors are indexed starting at one, hence the $z + 1$ offset for the digit labels). We adopt the notation where we have n data points in our training objective with features $x_i \in \mathbb{R}^d$ and label one-hot encoded as $y_i \in \{0, 1\}^k$. Here, $k = 10$ since there are 10 digits.

a. *[10 points]* In this problem we will choose a linear classifier to minimize the regularized least squares objective:

$$\widehat{W} = \text{argmin}_{W \in \mathbb{R}^{d \times k}} \sum_{i=1}^{n} \|W^T x_i - y_i\|_2^2 + \lambda \|W\|_F^2$$

Note that $\|W\|_F$ corresponds to the Frobenius norm of W, i.e. $\|W\|_F^2 = \sum_{i=1}^{d} \sum_{j=1}^{k} W_{i,j}^2$. To classify a point x_i we will use the rule $\arg\max_{j=0,\ldots,9} e_{j+1}^T \widehat{W}^T x_i$. Note that if $W = \begin{bmatrix} w_1 & \ldots & w_k \end{bmatrix}$ then

$$\begin{aligned} \sum_{i=1}^{n} \|W^T x_i - y_i\|_2^2 + \lambda \|W\|_F^2 &= \sum_{j=1}^{k} \left[\sum_{i=1}^{n} (e_j^T W^T x_i - e_j^T y_i)^2 + \lambda \|W e_j\|^2 \right] \\ &= \sum_{j=1}^{k} \left[\sum_{i=1}^{n} (w_j^T x_i - e_j^T y_i)^2 + \lambda \|w_j\|^2 \right] \\ &= \sum_{j=1}^{k} \left[\|X w_j - Y e_j\|^2 + \lambda \|w_j\|^2 \right] \end{aligned}$$

where $X = \begin{bmatrix} x_1 & \ldots & x_n \end{bmatrix}^\top \in \mathbb{R}^{n \times d}$ and $Y = \begin{bmatrix} y_1 & \ldots & y_n \end{bmatrix}^\top \in \mathbb{R}^{n \times k}$. Show that

$$\widehat{W} = (X^T X + \lambda I)^{-1} X^T Y$$

b. *[10 points]*

- Implement a function `train` that takes as input $X \in \mathbb{R}^{n \times d}$, $Y \in \{0, 1\}^{n \times k}$, $\lambda > 0$ and returns $\widehat{W} \in \mathbb{R}^{d \times k}$.

- Implement a function `one_hot` that takes as input $Y \in \{0, ..., k-1\}^n$, and returns $Y \in \{0,1\}^{n \times k}$.
- Implement a function `predict` that takes as input $W \in \mathbb{R}^{d \times k}$, $X' \in \mathbb{R}^{m \times d}$ and returns an m-length vector with the ith entry equal to $\arg\max_{j=0,...,9} e_j^T W^T x_i'$ where $x_i' \in \mathbb{R}^d$ is a column vector representing the ith example from X'.
- Using the functions you coded above, train a model to estimate $\widehat{W}$ on the MNIST training data with $\lambda = 10^{-4}$, and make label predictions on the test data. This behavior is implemented in `main` function provided in zip file. **What is the training and testing error?** Note that they should both be about 15%.

What to Submit:

- **Part A:** Derivation of expression for $\widehat{W}$
- **Part B:** Values of training and testing errors
- **Code** on Gradescope through coding submission

Answers:
Part A:

$$\begin{aligned}\sum_{i=1}^{n} \|W^T x_i - y_i\|_2^2 + \lambda \|W\|_F^2 &= \sum_{j=0}^{k} \left[\sum_{i=1}^{n} (e_j^T W^T x_i - e_j^T y_i)^2 + \lambda \|W e_j\|_F^2 \right] \\ &= \sum_{j=0}^{k} \left[\sum_{i=1}^{n} (w_j^T x_i - e_j^T y_i)^2 + \lambda \|w_j\|^2 \right] \\ &= \sum_{j=0}^{k} \left[\|X w_j - Y e_j\|^2 + \lambda \|w_j\|^2 \right]\end{aligned}$$

Take the derivative wrt to W and set it to zero in order to find the minimum:

$$\begin{aligned}&\frac{\partial}{\partial w_j} \sum_{j=0}^{k} \left[\|X w_j - Y e_j\|^2 + \lambda \|w_j\|^2 \right] \\ &= \frac{\partial}{\partial w_j} \left[\|X w_j - Y_{:,j}\|^2 + \lambda \|w_j\|^2 \right] \\ &= X^T (X w_j - Y_{:,j}) + \lambda w_j \\ &= 0 \\ &\Leftrightarrow \widehat{w}_j = (X^T X + \lambda I)^{-1} Y_{:,j}\end{aligned}$$

where $Y_{:,j}$ stands for j-th column of Y

$$\boxed{\therefore \widehat{W} = (X^T X + \lambda I)^{-1} X^T Y \qquad \square}$$

Part B:
Training error = 0.14805
Test error = 0.1466

B3.

a. *[5 points]* Instead of reporting just the test error, which is an unbiased estimate of the *true* error, we would like to report a *confidence interval* around the test error that contains the true error.

Lemma 1. *(Hoeffding's inequality) Fix $\delta \in (0,1)$. If for all $i = 1, \ldots, m$ we have that X_i are i.i.d. random variables with $X_i \in [a,b]$ and $\mathbb{E}[X_i] = \mu$ then*

$$\mathbb{P}\left(\left|\left(\frac{1}{m}\sum_{i=1}^{m} X_i\right) - \mu\right| \geq \sqrt{\frac{(b-a)^2 \log(2/\delta)}{2m}}\right) \leq \delta$$

We will use the above equation to construct a confidence interval around the true classification error $\epsilon(\widehat{f}) = \mathbb{E}_{\text{test}}[\widehat{\epsilon}_{\text{test}}(\widehat{f})]$ since the test error $\widehat{\epsilon}_{\text{test}}(\widehat{f})$ is just the average of indicator variables taking values in $\{0,1\}$ corresponding to the ith test example being classified correctly or not, respectively, where an error happens with probability $\mu = \epsilon(\widehat{f}) = \mathbb{E}_{\text{test}}[\widehat{\epsilon}_{\text{test}}(\widehat{f})]$, the *true* classification error.

Let $\widehat{p}$ be the value of p that approximately minimizes the validation error on the plot you just made and use $\widehat{f}(x) = \arg\max_j x^T \widehat{W^p} e_j$ to compute the classification test error $\widehat{\epsilon}_{\text{test}}(\widehat{f})$. Use Hoeffding's inequality, of above, to compute a confidence interval that contains $\mathbb{E}_{\text{test}}[\widehat{\epsilon}_{\text{test}}(\widehat{f})]$ (i.e., the *true* error) with probability at least 0.95 (i.e., $\delta = 0.05$). Report $\widehat{\epsilon}_{\text{test}}(\widehat{f})$ and the confidence interval.

What to Submit:

- **Part a:** Testing error along with confidence interval around it.

Answers:

$$\epsilon(\widehat{f}) \in \widehat{\epsilon}_{test}(\widehat{f}) \pm \sqrt{\frac{log(2/0.05)}{2m}}$$

Confidence Interval of Least Squares Estimation

Bounding the Estimate

B4. Let us consider the setting, where we have n inputs, $X_1, ..., X_n \in \mathbb{R}^d$, and n observations $Y_i = \langle X_i, \beta^* \rangle + \epsilon_i$, for $i = 1, ..., n$. Here, β^* is a ground truth vector in $\mathbb{R}^d$ that we are trying to estimate, the noise $\epsilon_i \sim \mathcal{N}(0, 1)$, and the n examples piled up — $X \in R^{n \times d}$. To estimate, we use the least squares estimator $\widehat{\beta} = \min_\beta \|X\beta - Y\|_2^2$. Moreover, we will use $n = 20000$ and $d = 10000$ in this problem.

a. *[3 points]* Show that $\widehat{\beta}_j \sim \mathcal{N}(\beta_j^*, (X^T X)^{-1}_{j,j})$ for each $j = 1, ..., d$. *(Hint: see notes on confidence intervals from lecture.)*

b. *[4 points]* Fix $\delta \in (0, 1)$ suppose $\beta^* = 0$. Applying the proposition from the notes, conclude that for each $j \in [d]$, with probability at least $1 - \delta$, $|\widehat{\beta}_j| \leq \sqrt{2(X^T X)^{-1}_{j,j} \log(2/\delta)}$. Can we conclude that with probability at least $1 - \delta$, $|\widehat{\beta}_j| \leq \sqrt{2(X^T X)^{-1}_{j,j} \log(2/\delta)}$ for all $j \in [d]$ simultaneously? Why or why not?

c. *[5 points]* Let's explore this question empirically. Assume data is generated as $x_i = \sqrt{(i \mod d) + 1} \cdot e_{(i \mod d)+1}$ where e_i is the ith canonical vector and $i \mod d$ is the remainder of i when divided by d. Generate each y_i according to the model above. Compute $\widehat{\beta}$ and plot each $\widehat{\beta}_j$ as a scatter plot with the x-axis as $j \in \{1, \ldots, d\}$. Plot $\pm\sqrt{2(X^T X)^{-1}_{j,j} \log(2/\delta)}$ as the upper and lower confidence intervals with $1 - \delta = 0.95$. How many $\widehat{\beta}_j$'s are outside the confidence interval? *Hint: Due to the special structure of how we generated x_i, we can compute $(X^T X)^{-1}$ analytically without computing an inverse explicitly.*

What to Submit:

- **Parts a, b:** Proof.
- **Part b:** Answer.
- **Part c:** Plots of $\hat{\beta}$ and its confidence interval **on the same plot**.

Answers:
Part a:
Using least square estimator, we get:

$$\widehat{\beta} = \min_\beta \|X\beta - Y\|_2^2$$

Taking the derivative and setting it to zero, we get:

$$\begin{aligned} 2X(X\widehat{\beta} - Y) &= 0 \\ X\widehat{\beta} &= Y \\ X^T X\widehat{\beta} &= X^T Y \\ \widehat{\beta} &= (X^T X)^{-1} X^T Y \end{aligned}$$

Finally,

$$\begin{aligned} Y &\sim \mathcal{N}(X\beta, 1) \\ X^T &\sim \mathcal{N}(X^T X\beta, X^T X)) \\ (X^T X)^{-1} X^T Y = Y &\sim \mathcal{N}(\beta, (X^T X)^{-}1) \\ \widehat{\beta}_j &\sim \mathcal{N}(B_j^*, (X^T X)^{-1}_{j,j} \end{aligned}$$

□

Part b:
No we cannot draw that conclusion. The probability in the sub-problem considers only a point-wise confidence band which is independent for each β_j. If we want to consider all j at the same time, then we will need to consider the simultaneous interval of β as a whole.

Part c:
There are 3 $\widehat{\beta}_j$ outside the confidence interval of $1 - \sigma = 0.95$

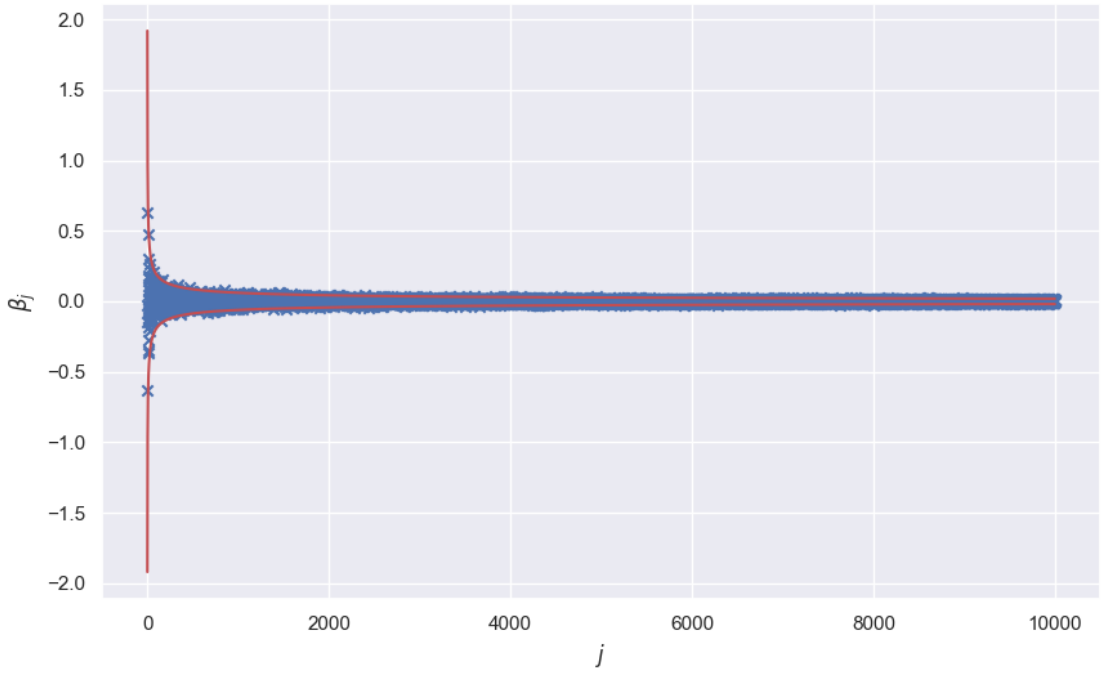

Figure 8: Plot of $\widehat{\beta}_j$ with the confidence interval of $1 - \sigma = 0.95$

Administrative

A6.

a. *[2 points]* About how many hours did you spend on this homework? There is no right or wrong answer :)

I spent about 30-40 hours spread out over a couple of days