# Final Exam, CPSC 8420, Fall 2024

Last Name, First Name

**Due 12/12/2024, Thursday, 5:59PM EST**

## Problem 1 [15 pts]

Consider the following problem:

$$\min_{\beta} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_1. \tag{1}$$

1. Prove that if $\lambda \geq \|\mathbf{X}^T\mathbf{y}\|_\infty$, then $\beta^* = 0$.

2. To validate the corretness of the conclusion above, let's find the optimal solution manually via experiment. As $\beta$ is a vector consisting of various elements $\beta[1], \beta[2], \ldots, \beta[-1]$, one of the most popular methods to find the optimal solution is so called 'coordinate descent' which minimizes a certain coordinate while fixing the rest. For example, we can first fix the rest while optimizing $\beta[1]$, then fix the rest to optimize $\beta[2]$, till $\beta[-1]$. By repeating the process until convergence, the optimal solution will be obtained. Please generate $\lambda \geq \|\mathbf{X}^T\mathbf{y}\|_\infty$ and make use of coordinate descent method described above to obtain the optimal $\beta$. It should be a zero vector (or very close to 0 due to machine precision issue).

## Problem 2 [10 pts]

- For any matrix with SVD decomposition $X = U\Sigma V^T$, define $\|X\|_2 = \Sigma(1,1), \|X\|_F = \sqrt{\sum_i \sum_j |x_{ij}|^2}$. Prove that $\|X\|_F \geq \|X\|_2$ and indicate when the equality holds.
- Use the fact that $vec(\mathbf{AXB}) = (\mathbf{B}^T \otimes \mathbf{A})vec(\mathbf{X})$ to find the best solution to $\min_{\mathbf{X}} \|\mathbf{AXB} - \mathbf{Y}\|_F^2$, where $\mathbf{A} \in \mathbb{R}^{m \times p}, \mathbf{X} \in \mathbb{R}^{p \times q}, \mathbf{B} \in \mathbb{R}^{q \times n}, \mathbf{Y} \in \mathbb{R}^{m \times n}$.

## Problem 3 [25 pts]

Please find *USArrests* dataset online and

- Implement your own program to reproduce the image on page 16/26 of 'PCA' slides on Canvas (if yours is flipped up and down, (or) left and right from the slide, it is totally Okay).
- For each state, out of 4 features, randomly mask one and assume it is missing (therefore you have your own $\Omega$ and $X$). Please write a program following what we discussed in class (you may refer to ProximalGradientDescent.pdf on Canvas) to optimize

$$\min_{Z} \frac{1}{2}\|P_{\Omega}(X - Z)\|_F^2 + \|Z\|_*, \tag{2}$$

  and plot the objective vs. iteration to demonstrate the algorithm will decrease the function.

## Problem 4 [15 pts]

Please reproduce Figure (14.29) in The Elements of Statistical Learning with your own codes. You are NOT allowed to call 'spectral clustering' function built-in python or matlab.

## Problem 5 [20 pts]

For Logistic Regression, assume each data $\mathbf{x}_i \in \mathbb{R}^{100}$. If the label is $\pm 1$, the objective is:

$$\min_{\mathbf{w}} \sum_{i=1}^{m} \log(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i)) \tag{3}$$

while if the label is $\{1, 0\}$ the objective is:

$$\min_{\mathbf{w}} \sum_{i=1}^{m} \log(1 + \exp(\mathbf{w}^T \mathbf{x}_i)) - y_i \mathbf{w}^T \mathbf{x}_i \tag{4}$$

- Write a program to show that the optimal solutions to the two cases are the same by making use of gradient descent method where $m = 100$ (please carefully choose the stepsize as we discussed in class). You can generate two class samples, one class's label is 1 and the other is -1 or 0 corresponding to the two formulations respectively. You can initialize $\mathbf{w}$ as $\mathbf{0}$.

- Consider the case where class label is $\{1, 0\}$ and $P(y = 1|\mathbf{x}, \mathbf{w}) = \frac{1}{1+\exp(-\mathbf{w}^T \mathbf{x})}$, the maximum likelihood function is $p^y(1-p)^{1-y}$. Please prove optimal $p^* = y$. If we use Mean Square Error instead of cross entropy: $\min_p \frac{1}{2}(y-p)^2$, and assume groundtruth $y = 1$ and our initial guess weight $\mathbf{w}$ result in $p$ very close to 0, if we optimize this objective by making use of gradient descent method, what will happen? Please explain why.

- For the second objective where the label is $\{1, 0\}$, implement Newton method (with unit stepsize) where $m = 100$. Compare with gradient descent method (constant stepsize) and plot objective versus **iteration** in one figure.

- Still consider the second formulation. Please write a stochastic gradient descent version (you may set the stepsize as $1/(t+1)$ where $t = 0, 1, 2, \dots$) and compare those two methods (gradient descent vs. stochastic gradient descent) for $m = [100000, 10000, 1000, 100]$ by plotting objective changes versus **time consumption** respectively.

## Problem 6 [15 pts]

In class, we discussed Kernel SVM, we said there are many options for the kernel, such as linear, polynoimal, Gaussian, etc.

- Show that if $K(i,j) = \frac{\langle \mathbf{x}_i, \mathbf{x}_j \rangle}{\|\mathbf{x}_i\|_2 \|\mathbf{x}_j\|_2}$, then $K$ defines a proper kernel.
- We define $K = K_1 + K_2$ where $K_1$ is Gaussian Kernel ($\gamma = 1$) and $K_2$ is linear Kernel. Assume we are to train SVM on iris dataset using the kernel defined above ($K$). Since there are 3 classes, we need train 3 hyperplanes (one vs. one). Please determine how many support vectors for each of the 3 SVMs. (You can use quadratic programming solvers in Matlab or Python at your convinience)

## Problem 7 [10 pts]

- Please tell me your favorite book, favorite travel destination and why.
- Please tell me the person who influences you most and why.
- Please tell me your favorite restaurant and dishes you order.
- Please tell me your favorite (or least favorite) part of this class.
- Please tell me your favorite machine learning algorithm(s) we discussed in class and why.