CS 365 - Foundations of Data Science
Final
6-8pm, May 12th, 2023

(First, Last) Name: Jeong Yong Yang

BU ID: U95912941

Instructions

- Write your (official, no nicknames) first and last name, and your BU ID.
- This exam is CLOSED book, notes, and devices. No calculators are needed. For example, if you derive a fraction $\frac{5}{7}$ or cos(10), leave it in this form.
- The exam consists of 7 problems on pages 3-12. Page 2 contains helpful reminders.
- **Only pages 3-12 will be graded.** Write only your final answer and justification **in the space provided. Please answer all questions on this exam sheet.** The scratch paper is for your own use and it will not be graded.
- Only correct, and mathematically rigorous answers will receive full credit.
- Please read through all questions carefully, and make sure that you understand the problem before answering.
- Please ask if you have any questions but avoid disturbing your neighbors.
- Even an attempt to cheat will result in a grade of 0 in the final.
- **Great news:** There are 20 extra points.

GOOD LUCK!

1.) 19 (30 points)
2.) 9.5 (10 points)
3.) 15 (20 points)
4.) 10 (20 points)
5.) 15 (15 points)
6.) 10 (10 points)
7.) 11 (15 points)
Σ: 89.5 (100 points+20 extra)

Reminders

- **Markov's inequality:** If X is a non-negative random variable and $t > 0$, then

$$\Pr(X \geq t) \leq \frac{\mathbb{E}(X)}{t}.$$

- **Chebyshev's inequality:** Let X be a random variable with finite expected value μ and finite non-zero variance σ^2. Then for any real number $t > 0$

$$\Pr(|X - \mu| \geq t) \leq \frac{\text{Var}(X)}{t^2}.$$

- **Chernoff's inequality:** Let $X_i = 1$ with probability p_i, 0 with prob. $1-p_i$. Define $X = \sum_{i=1}^{n} X_i, \mu = \mathbb{E}(X) = \sum_{i=1}^{n} p_i$. Then, for any $0 < \epsilon < 1$,

$$\Pr(|X - \mu| \geq \epsilon\mu) \leq 2e^{-\frac{\epsilon^2}{3}\mu}.$$

- **Bayes' theorem:** Suppose that E and F are events from a sample space S such that $\Pr[E] \neq 0$ and $\Pr[F] \neq 0$. Then:

$$Pr[F|E] = \frac{\Pr[E|F]\Pr[F]}{\Pr[E|F]\Pr[F] + \Pr[E|\overline{F}]\Pr[\overline{F}]}.$$

- **Distributions**

  The moment generating function of a Gaussian variable $X \sim N(\mu, \sigma^2)$ is $M_X(t) = \mathbb{E}\left[e^{tX}\right] = e^{t\mu + \frac{1}{2}\sigma^2 t^2}$.
  The pdf is $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.

  The pdf of the uniform random variable $X \sim Unif(a, b)$ is defined as

$$f_X(x) = \begin{cases} \frac{1}{b-a} \text{ if } a < x < b, \\ 0 \text{ otherwise} \end{cases}$$

  A discrete random variable X that follows Poisson distribution has the pmf $\mathbb{P}[X = x] = \lambda^x e^{-\lambda}/x!, x = 0, 1, \ldots$.

  A discrete random variable K that follows a geometric distribution has the pmf $\mathbb{P}[K = k] = p(1-p)^{k-1}, k = 1, 2, \ldots$.

1 Multiple choice/Short answers [30 points]

1. (a) (3pts) A particle is performing random jumps on the real line, starting from 0. Every second, it jumps to the left with probability $\frac{1}{2}$, and to the right with probability $\frac{1}{2}$. Let X_t be its position after t seconds. What is the variance of X_t?

0/3

A. t B. $2t$ C. 0 D. 1 E. $\sqrt{t}$ F. t^2 G. None of the previous.

(b) (4pts) Upper bound the probability that after 100 steps the particle is positioned above 50 or below -50 using Chebyshev's inequality.

3/4

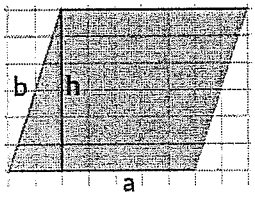
$$\Pr(|X_t - \mu| \geq 50) \leq \frac{Var(X_t)}{50^2} = \frac{1}{2500}$$

$$Var(X_t) = E(X_t^2) - E(X_t)^2 = E(X_t) - 0 = 1$$

we can use Chebyshev since $50 > 0$.

2. (3pts) Write the optimization problem for finding the *parallelogram* of maximum area with a given perimeter ℓ. Use the same notation for the variables as shown in the following figure.

1/3



$$\frac{\ell}{2} - b = a$$

$$\ell = 2a + 2b$$

$$\max \frac{(a+b)h}{2} \quad \times$$

$b \geq h$

$$\max \frac{(\frac{\ell}{2})h}{2} = \max \frac{\ell h}{4}$$

3. (3pts) What is a saddle point of a function $f : \mathbb{R}^{n+m} \to \mathbb{R}$? Give the formal definition and an example of a function $f : \mathbb{R}^2 \to \mathbb{R}$ that has a saddle point.

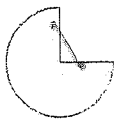
1/3

Saddle points are points that have a minimum on one axis but maximum on other axis, which can occur if the Hessian of the function has positive and negative eigenvalues.

X

4. (2pts) Is the following set convex or not? Explain your answer in few words.

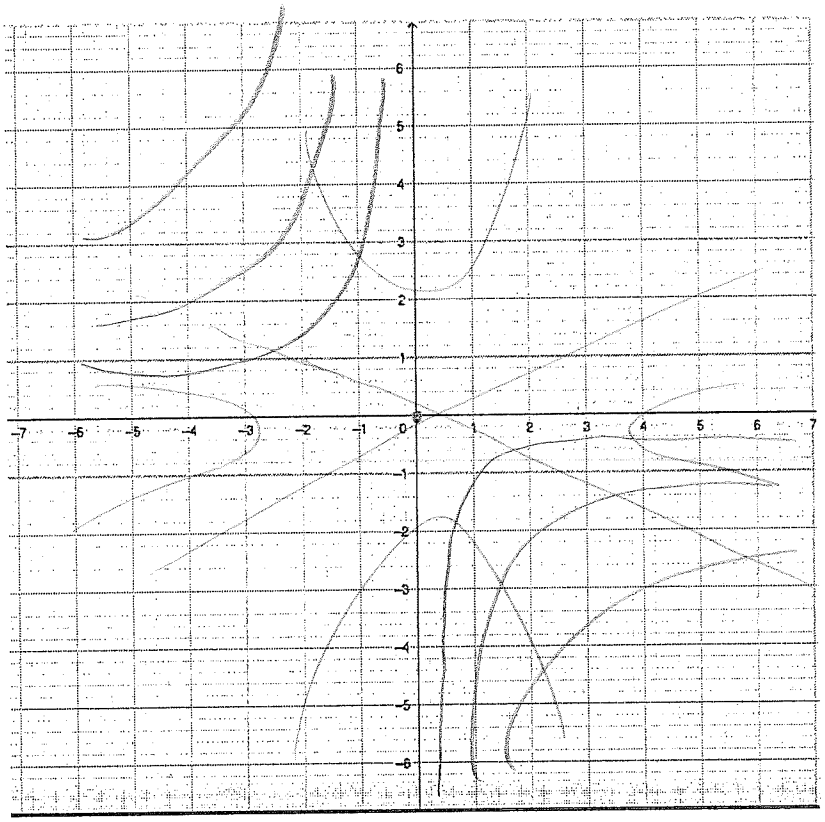
2/2.



The following is not a convex set because you can draw a line out of the set as shown.

5. (3pts) Draw the level curves of $f(x, y) = x^2 - 2y^2$. The x axis is the horizontal axis.

0/3



6. (2pts) Suppose you get n iid samples $\{x_1, \ldots, x_n\}$ from a Poisson distribution $\mathbb{P}[X = x] = \lambda^x e^{-\lambda}/x!$. What is MLE for parameter λ?

(A.) $\lambda_{MLE} = \frac{x_1+\ldots+x_n}{n}$ B. $\lambda_{MLE} = \frac{x_1^2+\ldots+x_n^2}{n}$ C. $\lambda_{MLE} = \frac{n}{x_1+\ldots+x_n}$ D. $\lambda_{MLE} = \frac{n}{x_1^2+\ldots+x_n^2}$

✓

$$\frac{\lambda^x e^{-\lambda}}{x!} \qquad \prod_i e^{-n\lambda} \qquad \frac{e^{-n\lambda}\lambda^{\sum x}}{\sum x}$$

$$-n + \frac{\sum x}{\lambda} = 0 \Rightarrow \lambda = \frac{\sum}{n}$$

7. (5 pts) Calculate the second-degree Taylor polynomial of $f(x, y) = e^{-x^2-y^2}$ at the point $(0, 0)$ (here $x, y \in \mathbb{R}$).

$f(x,y) = e^{-x^2-y^2}$

$f(0,0) = e^0 = 1$

$\nabla f(x,y) = \left(-2xe^{-x^2-y^2}, -2ye^{-x^2-y^2}\right)$

$\nabla f(0,0) = (0,0)$

$$Hf(x,y) = \begin{bmatrix} -2e^{-x^2-y^2} + 4x^2e^{-x^2-y^2} & +4xye^{-x^2-y^2} \\ 4xye^{-x^2-y^2} & -2e^{-x^2-y^2} + 4y^2e^{-x^2-y^2} \end{bmatrix}$$

$$Hf(0,0) = \begin{bmatrix} -2 & 0 \\ 0 & -2 \end{bmatrix}$$

$$T_2 = f(0,0) + \nabla f(0,0)\begin{bmatrix} x \\ y \end{bmatrix} + \frac{1}{2}[x\ y]\begin{bmatrix} -2 & 0 \\ 0 & -2 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix}$$

$$= 1 + [x\ y]\begin{bmatrix} -x \\ -y \end{bmatrix} = 1 + (-x^2 - y^2) = \boxed{-x^2 - y^2 + 1}$$

✓

8. (5 pts) Give the precise mathematical definition of a k-wise independent family of hash. What is the required space to store a hash function from such a family? Assume that the size of the universe is n.

The mathematical definition of k-wise independent family of hash is

$$\Pr_{h \in H}\left[h_1(x) = x_1, \ldots, h_k(x) = x_k\right] \leq \frac{1}{m^k}$$

The required size is $k \log n$.

✓

2 Data Streams [10 points]

1. (5 pts) Design an algorithm that samples $k \geq 1$ elements uniformly at random from an insert-only stream, whose length is unknown. Present the pseudocode and prove the correctness of the proposed algorithm.

4.5/5

```
i = 0
store = [ ]
while there stream[i] exists:
    if k > i:
        store[i] = stream[i]
    else:
        with probability k/i, randomly choose an element from
        the store and replace with stream[i]
    i += 1
```

In the base case, we have elements switching with probability $\frac{k}{i+1}$.

For the next element, we have $\left(\frac{k}{i}\right)\left(\frac{i}{i+1}\right) = \frac{k}{i+1}$.

How about after t steps?

2. (5 pts) Consider the F_0 estimation problem in a data stream of integers. Suppose you have access to an idealized hash function that maps each integer in $(0,1)$. Let $M = \max_{x \in \text{stream}} h(x)$. What is the expected value of M, and how can you use it to get an estimate of F_0?

5/5

$E(M) = 1 -$ min's expected value

$= 1 - \frac{1}{F_0 + 1}$

$= \frac{F_0}{F_0 + 1}$

$\frac{F_0}{F_0+1} = m$

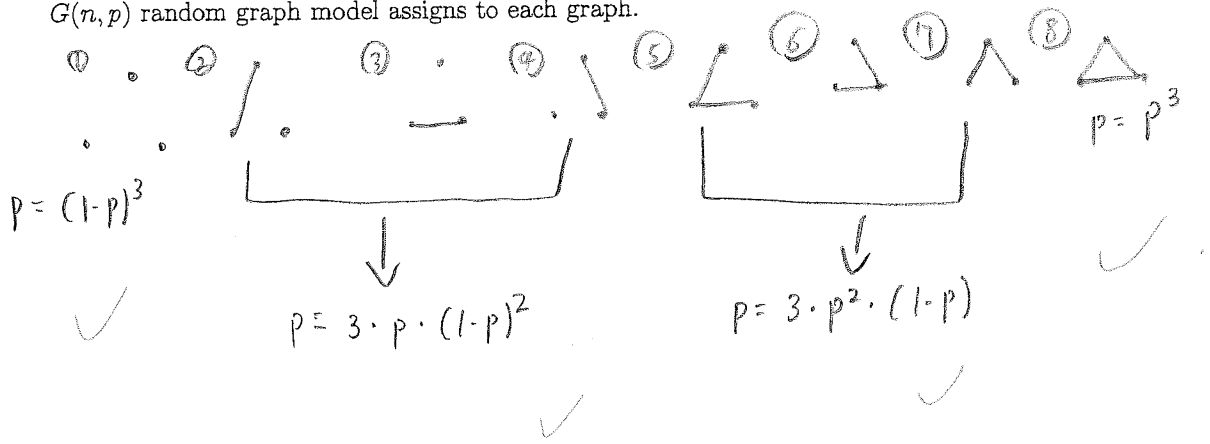
$F_0 = m \cdot F_0 + m$

$F_0 - mF_0 = m$

$F_0 = \frac{m}{1-m}$

3 Graphs [20 points]

1. (6 pts) Draw all possible labeled graphs on nodes $\{1, 2, 3\}$ and write clearly the probability that the $G(n, p)$ random graph model assigns to each graph.



2. (8 pts) Consider a graph G sampled from the $G(n, p)$ model where $p = o(1/n)$. Prove that the graph has no triangles with high probability.

Assume number of triangles of $G(n,p)$ model is X.
Therefore, we are looking for $Pr(X > 0) = Pr(X \geq 1)$ tends to 0.
Since X is a non-negative random variable, (number of triangles can't be negative), we can apply Markov's inequality.

$Pr(X \geq 1) \leq \frac{E(X)}{1} = E(X) = \frac{1}{g^3(n)}$, which goes to 0 with high probability

$E(X) = n(3 \cdot p^3 \leq n^3 p^3$

Now, if $p = o(1/n)$, it means $p = \frac{1}{n g(n)}$ such that $g(n) \to \infty$ when $n \to \infty$

Plugging in, we get

$E(X) \leq n^3 \left(\frac{1}{n^3 g(n)^3}\right) = \frac{1}{g(n^3)}$

3. (6 pts) Consider a graph chosen uniformly at random among all graphs with exactly m edges on n nodes where $0 \leq m \leq \binom{n}{2}$. What is the expected number of triangles as a function of n, m?

$E(X) = 2^{n \choose 2} \left(3\ p^3 (1-p)^{2^{n \choose 2} - 3}\right)$

4 MLE and MoM [20 points]

Let $X_1, \ldots, X_n \sim Unif[0, \theta]$ be an iid sample of size n. The parameter θ of the uniform distribution is unknown. You need to estimate θ from the sample using the method of moments and maximum likelihood principles. Specifically answer the following questions.

(a) (4 pts) Find $\hat{\theta}_{MoM}$, i.e., the method of moments estimator of the unknown parameter θ.

First moment $\rightarrow \frac{\sum_{i=1}^{n} X_i}{n} = \frac{0+\theta}{2}$

$= \frac{\theta}{2}$ $\quad \hat{\theta}_{MoM} = \frac{2\sum_{i=1}^{n} X_i}{n}$ ✓

(b) (8 pts) Write down the likelihood function of the sample.

4/8

$L(X_1, \ldots, X_n; \theta) = \prod_{i=1}^{n} \frac{1}{\theta - 0} = \frac{1}{\theta^n}$

$= \sum_{i=1}^{n} \theta^{-1}$

$L(X_1, \ldots, X_n; \theta) = \prod_{i=1}^{n} \left(\frac{1}{\theta - 0}\right) = \frac{1}{\theta^n}$

$= \frac{1}{\sum_{i=1}^{n} \theta}$

{ when $\theta < \max(X_i)$, $L = 0$

$L(X_1, \ldots, X_n; \theta) = \prod_{i=1}^{n} \left(\frac{1}{\theta - 0}\right) = \frac{1}{\theta^n} = \frac{1}{\sum_{i=1}^{n} \theta}$ ✓

(c) (8 pts) Find $\hat{\theta}_{MLE}$, i.e., the maximum likelihood estimator of the unknown parameter θ.

2/8

Log-Likelihood

↳ $\sum_{i=1}^{n} \theta^{-1} = \log\left(\sum_{i=1}^{n} \theta^{-1}\right) - \log\left(\sum_{i=1}^{n} \theta\right)$

$\frac{d}{d\theta}\left(-\ln \sum_{i=1}^{n} \theta\right) = \frac{-1}{\sum_{i=1}^{n} \theta}$

Log-likelihood

↳ $\ln\left(\frac{1}{\sum_{i=1}^{n} \theta}\right) = -\ln \sum_{i=1}^{n} \theta$

$\frac{d}{d\theta}\left(-\ln \sum_{i=1}^{\theta} \theta\right) = \frac{-1}{\sum_{i=1}^{\theta} \theta} = 0$ can't be 0. ✓

5 Vector calculus [15 points]

Compute the gradient $\frac{df}{dx}$ for the function $f : \mathbb{R}^n \to \mathbb{R}$ (i.e., $x, \mu \in \mathbb{R}^n$, $\Sigma \in \mathbb{R}^{n \times n}$) using the chain rule, that is defined as follows

$$f(x) = e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}.$$

$f: \mathbb{R}^n \to \mathbb{R}$

$$\frac{df}{da} = e^a$$

$$a = -\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)$$

$$= -\frac{1}{2}(x^T - \mu^T)\Sigma^{-1}(x-\mu)$$

$$\sum_{i=1}^{n} -\frac{1}{2}(x_i^T - \mu_i^T)\Theta_i^{-1}(x_i - \mu_i)$$

$$= -\frac{1}{2}(x_i^T - \mu_i^T)(\Theta_i^{-1}x_i - \Theta_i^{-1}\mu_i) = -\frac{1}{2}\Theta_i^{-1}(x_i^T - \mu_i^T)(x_i - \mu_i)$$

$$= \sum_{i=1}^{n} -\frac{1}{2}\Theta_i^{-1}(x_i^T x_i - x_i^T\mu_i - x_i\mu_i^T - \mu_i^T\mu_i)$$

$$a = -\frac{1}{2}\Sigma^{-1}(x^T x - x^T\mu - \mu^T x - \mu^T\mu)$$

$$\frac{df}{dx} = \Sigma^{-1}(\mu^T - x^T)e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

$$\frac{da}{dx} = -\frac{1}{2}\Sigma^{-1}(2x^T - \mu^T - \mu^T)$$

$$= -\frac{1}{2}\Sigma^{-1}(2x^T - 2\mu^T)$$

$$= \Sigma^{-1}(\mu^T - x^T)$$

6 SVD [10 points]

Compute the SVD of the following matrices. Write explicitly U, Σ, V such that $A = U\Sigma V^T$, and explain your reasoning briefly. Notice that due to the simplicity of the matrices, one can eyeball the "SVD". However, if a full computation is carried out and done correctly, it will receive full credit.

1. (5 pts) $A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \\ 3 & 6 \end{bmatrix}$.

2 8 18 4 16 36

$A^TA = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \end{bmatrix}\begin{bmatrix} 1 & 2 \\ 2 & 4 \\ 3 & 6 \end{bmatrix} = \begin{bmatrix} 14 & 28 \\ 28 & 56 \end{bmatrix}$ $\det(A^TA - \lambda I) = \lambda^2 - 70\lambda + 784 - 784$

$0 = \lambda^2 - 70\lambda$

$\lambda = 0, 70$ $\Sigma = \begin{bmatrix} \sqrt{70} & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$

$V \rightarrow$ row space, nullspace

$v = \begin{bmatrix} 1 & -2 \\ 2 & 1 \end{bmatrix}$ $v^T = \begin{bmatrix} 1 & 2 \\ -2 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1/\sqrt{5} & 2/\sqrt{5} \\ -2/\sqrt{5} & 1/\sqrt{5} \end{bmatrix}$

$\vec{u} \rightarrow$ column space

$u = \begin{bmatrix} 1 & -2 & -3 \\ 2 & 1 & 0 \\ 3 & 0 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1/\sqrt{14} & -2/\sqrt{5} & -3/\sqrt{10} \\ 2/\sqrt{14} & 1/\sqrt{5} & 0 \\ 3/\sqrt{14} & 0 & 1/\sqrt{10} \end{bmatrix}$

$A = U\Sigma V^T = \begin{bmatrix} 1/\sqrt{14} & -2/\sqrt{5} & -3/\sqrt{10} \\ 2/\sqrt{14} & 1/\sqrt{5} & 0 \\ 3/\sqrt{14} & 0 & 1/\sqrt{10} \end{bmatrix}\begin{bmatrix} \sqrt{70} & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} 1/\sqrt{5} & 2/\sqrt{5} \\ -2/\sqrt{5} & 1/\sqrt{5} \end{bmatrix}$ ✓

2. (5 pts) $A = \begin{bmatrix} 4 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$.

$A^TA = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} 4 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 16 & 0 \\ 0 & 0 \end{bmatrix}$ $\lambda = \sqrt{16}, 0$ $\Sigma = \begin{bmatrix} 4 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$

$v = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ $v^T = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

$u = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

$A = U\Sigma V^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 4 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ ✓

7 Least Squares and Optimization [15 points]

1. Express the following problems as a least squares minimization problem of the form

$$\min_{\theta \in \mathbb{R}^n} ||\Phi\theta - b||_2^2.$$

Specify $\Phi \in \mathbb{R}^{m \times n}, \theta \in \mathbb{R}^n, b \in \mathbb{R}^m$.

Note that **you do not need to solve the optimization problems (a), (b), (c).** Use the following parametric equations: for a line $y = Ax + B$, for a parabola $y = Ax^2 + Bx + C$ and for an ellipse $A^2x^2 + B^2y^2 = C^2$.

(a) (2 pts) Suppose that we have measured three data points in 2 dimensions, $(0, 6), (1, 0), (2, 0)$. The goal is to find the least squares line that best approximates the data points.

$\min_{\theta \in \mathbb{R}^n} ||\Phi\theta - b||_2^2$

$$\Phi = \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 2 & 1 \end{bmatrix} \quad \theta = \begin{bmatrix} A \\ B \end{bmatrix} \quad b = \begin{bmatrix} 6 \\ 0 \\ 0 \end{bmatrix}$$ ✓

(b) (2 pts) Find the parabola that best approximates the data points $(0, 2), (-3, 1), (-1, 1), (2, 1), (1, -1)$.

$\min_{\theta \in \mathbb{R}^n} ||\Phi\theta - b||_2^2$

$$\Phi = \begin{bmatrix} 0 & 0 & 1 \\ 9 & -3 & 1 \\ 1 & -1 & 1 \\ 4 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix} \quad \theta = \begin{bmatrix} A \\ B \\ C \end{bmatrix} \quad b = \begin{bmatrix} 2 \\ 1 \\ 1 \\ 1 \\ -1 \end{bmatrix}$$ ✓

(c) (2 pts) Find the ellipse that best approximates the data points $(0, 2), (-3, 1), (-1, 1), (2, 1), (1, -1)$.

1/2

$\min_{\theta \in \mathbb{R}^n} ||\Phi\theta - b||_2^2$

$$\Phi = \begin{bmatrix} 0 & 0 & -1 \\ 9 & 9 & -1 \\ 1 & 1 & -1 \\ 4 & 4 & -1 \\ 1 & 1 & -1 \end{bmatrix} \quad \theta = \begin{bmatrix} A^2 \\ B^2 \\ C^2 \end{bmatrix} \quad b = \begin{bmatrix} 4 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$ all 0.

2. (a) (5 pts) Derive the first order optimality condition of the stationary point x^* for the following regularized least squares problem that is known as ridge regression in the literature:

$$\min_{\theta \in \mathbb{R}^n} ||\Phi\theta - b||_2^2 + \lambda||\theta||_2^2.$$

Here, $\Phi \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$ and $\lambda > 0$ are given.

4/5

Stationary point means that the first derivative must equal 0.

As solved in b, its derivative is

$$2(\Phi^T\theta^T - b^T)\Phi - 2\lambda\theta^T = 0$$

So far correct.

$$\lambda\theta^T = (\Phi^T\theta^T - b^T)\Phi$$

$$\lambda\theta^T\Phi^{-1} = \Phi^T\theta^T - b^T$$

$$(\Phi^{-1})^T\theta\lambda^T = \Phi\theta - b$$

$$\Phi^{-1}\left((\Phi^{-1})^T\theta\lambda^T + b\right) = 0 \quad \times$$

(b) (4 pts) How would gradient descent solve the problem in 2(a)? Derive the precise equation that we should apply in each iteration for this problem, assuming a step $\eta > 0$.

2 3/4

$$\min_{\theta \in \mathbb{R}^n} ||\Phi\theta - b||_2^2$$

Let $e = \Phi\theta - b$. Then, $\min_{\theta \in \mathbb{R}^n} ||e||_2^2 = e^Te$

$$\frac{\partial f}{\partial e} = 2e^T \qquad \frac{\partial e}{\partial \theta} = \Phi \qquad \frac{\partial f}{\partial \theta} = \Phi \cdot 2e^T$$

$$\min_{\theta \in \mathbb{R}^n} ||\Phi\theta - b||_2^2 = 2 \cdot (\Phi\theta - b)^T \cdot \Phi$$

$$= 2(\Phi^T\theta^T - b^T) \cdot \Phi$$

What about gradient descent with step size η?

Scratch paper (will not be graded)