STAT 154 Spring 2019: Mid-term Exam Solutions

Instructor: Prof. Bin Yu
Mar 21, 8:10 AM–9:30 AM

Name: _____ Student ID: _____

Row Number: _____ Seat Number: _____

Maximum Points: 50 Time: 80 minutes

Instructions (please read carefully)

- Do not turn the page until you are told do so. Just count the number of sheets, **you should have** 11 **sheets** (22 pages counting both sides). If you have any issues contact the GSI/instructor immediately.
- WRITE your student ID CLEARLY on top of each page.
- This exam has **5 questions**.
- Your answer **will be graded only if** it is written in the space provided after the question. Use the blank pages at the end for rough work. There is a help-sheet at the end for your comfort.
- **When time is up**, please take your exam in your **left hand** and raise it. When asked pass it on to your left (facing the board). Please maintain the sequence of your seating. Instructor/GSI will collect the copies from the leftmost student in the row.

Pre-Exam Questions

1. What is (are) your favorite movie(s)?
   Ans. Inception, Avatar, Dark Knight, Dangal
2. What do you like the most about this class?
   Ans. The amazing students

1 True or False (no justification, 10 parts, 10 pts)

Please answer the following statements as True or False by darkening the corresponding bubble PROPERLY as shown below. *Don't get creative with bubbles.* For example, if your answer is True:

● True ○ False

(a) Data collection process usually has no-to-little influence on the outcome of a prediction problem.

○ True ● False

Ans. Refer to HW1 True/False

(b) Suppose $\mathbf{A}$ is a $d \times d$ matrix where $d \geq 3$ such that $\mathbf{A} = \mathbf{a}\mathbf{a}^\top + \mathbf{b}\mathbf{b}^\top$, with $\mathbf{a}$ and $\mathbf{b}$ as d-dimensional vectors. Then the matrix $\mathbf{A}$ can have rank d.

○ True ● False

Ans. Adding two rank 1 matrices can lead to a matrix of rank at most 2.

(c) If $\lambda_1, \ldots, \lambda_d$ denote the eigenvalues of the matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$, then the eigenvalues of the matrix $\mathbf{A} - \gamma\mathbf{I}$ are given by $1 - \gamma\lambda_1, \ldots, 1 - \gamma\lambda_d$.

○ True ● False

Ans. The eigenvalues are $\lambda_1 - \gamma, \ldots, \lambda_d - \gamma$.

(d) The objective value in K-means algorithm never increases as the iteration increases.

● True ○ False

Ans. Refer to HW 2 Problem 4.

(e) EM always finds the global maximizer of the log-likelihood, i.e., $\arg\max_\theta \sum_{i=1}^n \log p(X = x_i; \theta)$.

○ True ● False

Ans. The non-convexity of the problem may lead to local-maxima.

(f) Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $n > d$. If some columns ($\{\mathbf{X}_{\cdot 1}, \ldots, \mathbf{X}_{\cdot d}\}$) are correlated, at least one singular value of the feature matrix $\mathbf{X}$ is zero.

○ True ● False

Ans. Correlation is not same as collinearity and hence the matrix can be full rank. Usually high correlation among columns leads to ill-conditioning.

(g) For Gaussian linear regression model, OLS produces an unbiased estimate of the true parameter while the ridge and lasso produce biased estimates.

● True ○ False

Ans. Revise notes. The regularization term in ridge/lasso leads to biased estimates.

(h) Gradient descent on the ridge regression objective function is guaranteed to converge to the ridge estimate as long as the step size is positive.

○ True ● False

Ans. The step size has to be small enough.

(i) Subsequent principal component directions are always orthogonal to each other.

● True ○ False

Ans. That is how PCs are defined.

(j) When the features do not have similar units and are on different scales, simply centering them is sufficient to obtain meaningful results with PCA.

○ True ● False

Ans. The statement itself is suggesting that different units among the features may require scaling.

2 EM with Exponential mixture (5 parts, 7 pts)

We work with the following simple two mixture model:

$$\begin{aligned} Z &\sim \text{Bernoulli}(1-w)+1 \\ X|Z=1 &\sim \text{Exponential}(\lambda) \quad \text{and} \\ X|Z=2 &\sim \text{Exponential}(\mu) \end{aligned} \tag{1}$$

where Z denotes the label of the exponential distribution from which X is drawn. Suppose that we are given n i.i.d. observations $\{x_1, \ldots, x_n\}$ only for X (i.e., the labels are unobserved).

Note that for a random variable $Y \sim \text{Exponential}(\lambda)$, the probability density is given by

$$p(Y=y) = \lambda e^{-\lambda y} \quad \text{for } y \geq 0.$$

(a) (1 pt) **Derive the log-likelihood for the observed data , i.e., derive the explicit expression for $\sum_{i=1}^{n} \log p(X=x_i;\lambda,\mu,w)$ in terms of the parameters and the data $\{x_1,\ldots,x_n\}$.**

Ans. We have

$$\begin{aligned} \sum_{i=1}^{n} \log p(X=x_i;\lambda,\mu,w) &= \sum_{i=1}^{n} \log(\sum_{z=1}^{2} p(X=x_i, Z=z;\lambda,\mu,w)) \\ &= \sum_{i=1}^{n} \log(w\lambda e^{-\lambda x_i} + (1-w)\mu e^{-\mu x_i}). \end{aligned}$$

(b) (1 pt, no justification) EM algorithm makes use of auxiliary (new) variables to obtain a lower bound on the log-likelihood of the problem.

● True ○ False

Ans. The q variables are the auxiliary variables. Refer to HW3.

(c) (1 pt, no justification) For a positive valued random variable X, we have $\mathbb{E}[\log(X)] \leq \log(\mathbb{E}[X])$.

● True ○ False

Ans. This is the Jensen's inequality for the concave function $x \to \log x$.

(d) (2 pts) **Write the explicit expression for the posterior probability, i.e., $p(Z = 1|X = x_i; w_t, \lambda_t, \mu_t)$.** No derivation needed. (You can do it in rough work).
Ans. We have

$$\begin{aligned} p(Z = 1|X = x_i; w_t, \lambda_t, \mu_t) &= \frac{p(Z = 1, X = x_i; w_t, \lambda_t, \mu_t)}{p(X = x_i; w_t, \lambda_t, \mu_t)} \\ &= \frac{w\lambda e^{-\lambda x_i}}{w\lambda e^{-\lambda x_i} + (1 - w)\mu e^{-\mu x_i}}. \end{aligned}$$

(e) (2 pts) For the model given in the problem, we now consider the M-step in which the following lower bound on log-likelihood is maximized:

$$\mathcal{F}(\lambda, \mu, w, q^{t+1}) = \sum_{i=1}^{n} \left[q_i^{t+1} \log \left(w\lambda e^{-\lambda x_i} \right) + (1 - q_i^{t+1}) \log \left((1-w)\mu e^{-\mu x_i} \right) \right]$$

for some suitable choice of q_i^{t+1}. **Taking the above expression as given, show that the M-step update for the parameter λ is given by**

$$\lambda_{t+1} = \frac{\sum_{i=1}^{n} q_i^{t+1}}{\sum_{i=1}^{n} q_i^{t+1} x_i}.$$

Ans. Note that only two terms depend on λ, i.e.,

$$\mathcal{F}(\lambda, \mu, w, q^{t+1}) = \sum_{i=1}^{n} \left[q_i^{t+1} (\log \lambda - \lambda x_i) + C \right]$$

where C denotes the terms that do not depend on λ. Thus, we have

$$\frac{\partial \mathcal{F}}{\partial \lambda} = \sum_{i=1}^{n} q_i^{t+1} \left(\frac{1}{\lambda} - x_i \right)$$

Setting to zero, we get

$$\sum_{i=1}^{n} q_i^{t+1} \left(\frac{1}{\lambda} - x_i \right) = 0 \Longrightarrow \lambda_{t+1} = \frac{\sum_{i=1}^{n} q_i^{t+1}}{\sum_{i=1}^{n} q_i^{t+1} x_i}.$$

3 Newton's Method with Least Squares (4 parts, 9 pts)

A popular method in machine learning is Newton's method. In order to minimize a function $f : \mathbb{R}^d \to \mathbb{R}$, given a starting point θ_0, the Newton update at time k is given as follows:

$$\theta_{k+1} = \theta_k - [\nabla^2 f(\theta_k)]^{-1} \nabla f(\theta_k)$$

where $\nabla^2 f(\theta)$ is a $d \times d$ symmetric matrix corresponding to the Hessian of f, with ij-th entry given by $[\nabla^2 f(\theta)]_{i,j} = \dfrac{\partial^2 f}{\partial \theta_i \partial \theta_j}$. Under appropriate assumptions, the iterates of Newton's method converge to the minimizer of f in a few steps.

(a) (2 pts) We now consider Newton's method for least squares. In particular, consider the minimization problem $\min_\theta f(\theta)$ where

$$f(\theta) = \frac{1}{2} \|\mathbf{X}\theta - \mathbf{y}\|_2^2$$

where $\mathbf{X} \in \mathbb{R}^{n \times d}$ is the feature matrix and $\mathbf{y} \in \mathbb{R}^n$ is the response vector. **Compute the gradient $\nabla_\theta f(\theta)$ and the Hessian $\nabla_\theta^2 f(\theta)$.**
Hint: You may use the fact that $\nabla_\theta^2(\theta^\top \mathbf{A}\theta) = (\mathbf{A} + \mathbf{A}^\top)$.
Ans. We have

$$f(\theta) = \frac{1}{2}\left[\theta^\top(\mathbf{X}^\top\mathbf{X})\theta - 2\theta^\top\mathbf{X}^\top\mathbf{y} + \mathbf{y}^\top\mathbf{y}\right].$$

Note that

$$\nabla_\theta(\theta^\top \mathbf{v}) = \mathbf{v}$$
$$\nabla_\theta(\theta^\top \mathbf{A}\theta) = (\mathbf{A} + \mathbf{A}^\top)\theta$$
$$\nabla_\theta^2(\theta^\top \mathbf{A}\theta) = (\mathbf{A} + \mathbf{A}^\top)\theta.$$

Thus, we obtain

$$\nabla_\theta f(\theta) = \mathbf{X}^\top\mathbf{X}\theta - \mathbf{X}^\top\mathbf{y}$$
$$\nabla_\theta^2 f(\theta) = \mathbf{X}^\top\mathbf{X}.$$

(b) (2 pts) Suppose that $n \geq d$ and that $\mathbf{X}$ is full column rank. **Show that Newton's method converges to the OLS esimtate in one-step from any starting point θ_0.**

Ans. Using the expressions from previous part, we have

$$\begin{aligned}
\theta_1 &= \theta_0 - [\nabla^2 f(\theta_0)]^{-1} \nabla f(\theta_0) \\
&= \theta_0 - (\mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{X}^\top \mathbf{X}\theta_0 - \mathbf{X}^\top \mathbf{y}) \\
&= \theta_0 - (\mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{X}^\top \mathbf{X})\theta_0 + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \\
&= \theta_0 - \theta_0 + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \\
&= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \\
&= \widehat{\theta}^{\mathbf{OLS}}.
\end{aligned}$$

(c) (1 pt) **Will you use Newton's method when $\mathbf{X}$ is not full column rank? Why or why not?**

Ans. No. The matrix $\mathbf{X}$ is not full rank and hence $\mathbf{X}^\top \mathbf{X}$ will not be invertible, so a direct application of Newton's method is not possible.

(d) (4 pts) Now consider the problem of ridge regression:

$$\min_{\theta} \frac{1}{2}(\|\mathbf{X}\theta - \mathbf{y}\|_2^2 + \lambda \|\theta\|_2^2).$$

Show that Newton's method on ridge regression objective always converges in one step to the ridge regression estimate (independent of the rank of matrix X).

Ans. We have

$$\begin{aligned} f(\theta) &= \frac{1}{2}\left[\theta^\top(\mathbf{X}^\top\mathbf{X})\theta - 2\theta^\top\mathbf{X}^\top\mathbf{y} + \mathbf{y}^\top\mathbf{y} + \theta^\top\theta\right] \\ &= \frac{1}{2}\left[\theta^\top(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_d)\theta - 2\theta^\top\mathbf{X}^\top\mathbf{y} + \mathbf{y}^\top\mathbf{y}\right]. \end{aligned}$$

Doing algebra as in part (a), we have

$$\begin{aligned} \nabla_\theta f(\theta) &= (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_d)\theta - \mathbf{X}^\top\mathbf{y} \\ \nabla_\theta^2 f(\theta) &= \mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_d. \end{aligned}$$

Note that $\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_d$ is always invertible as long as $\lambda > 0$. Hence we have

$$\begin{aligned} \theta_1 &= \theta_0 - [\nabla^2 f(\theta_0)]^{-1}\nabla f(\theta_0) \\ &= \theta_0 - (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_d)^{-1}((\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_d)\theta_0 - \mathbf{X}^\top\mathbf{y}) \\ &= \theta_0 - (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_d)^{-1}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_d)\theta_0 + (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_d)^{-1}\mathbf{X}^\top\mathbf{y}) \\ &= \theta_0 - \theta_0 + (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_d)^{-1}\mathbf{X}^\top\mathbf{y}) \\ &= (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_d)^{-1}\mathbf{X}^\top\mathbf{y} \\ &= \widehat{\theta}^{\mathbf{ridge}}. \end{aligned}$$

4 Bias-variance of Gradient Descent (7 parts, 14 pts)

In this problem, we study the bias-variance trade off for gradient descent method with least squares in linear model:

We are given n samples $\{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$ from the following model:

$$y_i = \mathbf{x}_i^\top \theta^* + \varepsilon_i,$$

where $\mathbf{x}_i \in \mathbb{R}^d$ are fixed, $y_i \in \mathbb{R}$ and the noise is independent Gaussian, i.e., $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$. We run gradient descent for the problem of least squares:

$$\min_\theta f(\theta) \quad \text{where} \quad f(\theta) = \frac{1}{2}\|\mathbf{X}\theta - \mathbf{y}\|_2^2 \quad \text{with } \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_n \end{bmatrix} \text{ and } \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}. \tag{2}$$

In this following part of this problem, we consider the special case of fixed orthonormal design, i.e., $\mathbf{X}^\top\mathbf{X} = \mathbf{I}_d$ (identity matrix in d dimensions). For this case, the gradient descent updates with step size γ can be simplified as

$$\begin{aligned} \theta_{k+1} &= \theta_k - \gamma \nabla_\theta f(\theta_k) \\ &= (1-\gamma)\theta_k + \gamma\theta^* + \gamma\mathbf{X}^\top\varepsilon, \end{aligned} \tag{3}$$

where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ is the n-dimensional column vector of noise terms ε_i. Take the update (3) as given and now answer the following questions.

(a) (2 pts) **For what range of γ the gradient descent updates are guaranteed to converge when $k \to \infty$? Do the updates converge to θ^*?**
We expect a explicit range for γ (not a generic formula) that takes into account the fact that $\mathbf{X}^\top\mathbf{X} = \mathbf{I}_d$.
Ans. Step size for gradient descent with convex functions should lie between 0 and $2/L$ where L is the maximum eigenvalue of the Hessian $\nabla^2 f$.

For OLS the Hessian is $\mathbf{X}^\top\mathbf{X}$ and since here $\mathbf{X}^\top\mathbf{X} = \mathbf{I}_d$ the range is $(0, 2)$.

One can also see this directly as $0 < |1 - \gamma| < 1$ if $\gamma \in (0, 2)$.

As discussed in class, gradient descent converges to the OLS estimate and not θ^* which in this case is simply $\mathbf{X}^\top\mathbf{y}$.

(b) (2 pts) **For all the following parts**, we consider gradient descent with step size $\gamma = \frac{1}{2}$. Define the random vector $\mathbf{v} = \mathbf{X}^\top \varepsilon$ where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$. Given a deterministic starting point θ_0, **show that the error of the k-th iterate is given by**

$$\theta_k - \theta^* = \frac{1}{2^k}(\theta^0 - \theta^*) + \mathbf{v}\left(1 - \frac{1}{2^k}\right). \tag{4}$$

Hint: You may find this fact useful: $\sum_{j=1}^{k} a^j = \frac{a(1-a^k)}{1-a}$ *for* $a \neq 1$.

Ans. Using equation (3), we have

$$\begin{aligned}\theta_k - \theta^* &= \left[\frac{1}{2}\theta_{k-1} + \frac{1}{2}\theta^* + \frac{1}{2}\mathbf{X}^\top\varepsilon\right] - \theta^* \\ &= \frac{1}{2}(\theta_{k-1} - \theta^*) + \frac{1}{2}\mathbf{v}.\end{aligned}$$

Now, noting that we have the same recursion between $k-1$ and $k-2$, we obtain

$$\begin{aligned}\theta_k - \theta^* &= \frac{1}{2}(\theta_{k-1} - \theta^*) + \frac{1}{2}\mathbf{v} \\ &= \frac{1}{2}\left(\frac{1}{2}(\theta_{k-2} - \theta^*) + \frac{1}{2}\mathbf{v}\right) + \frac{1}{2}\mathbf{v} \\ &= \frac{1}{2^2}(\theta_{k-2} - \theta^*) + \left(\frac{1}{2} + \frac{1}{2^2}\right)\mathbf{v} \\ &\vdots \\ &= \frac{1}{2^k}(\theta_0 - \theta^*) + \left(\frac{1}{2} + \frac{1}{2^2} + \ldots + \frac{1}{2^k}\right)\mathbf{v} \\ &= \frac{1}{2^k}(\theta_0 - \theta^*) + \left(1 - \frac{1}{2^k}\right)\mathbf{v}\end{aligned}$$

where in the last step we have used the hint with $a = \frac{1}{2}$. The proof is now complete.

(c) (2 pts) **What is the expectation of θ_k? What is the distribution of θ_k?** You may directly use equation (4).

Ans. We can compute the expectation directly from equation (4). Note that

$$\theta^k = \underbrace{\theta^* + \frac{1}{2^k}(\theta^0 - \theta^*)}_{\text{deterministic part}} + \underbrace{\mathbf{v}\left(1 - \frac{1}{2^k}\right)}_{\text{random part}}$$

Hence, we have

$$\begin{aligned}
\mathbb{E}[\theta^k] &= \theta^* + \frac{1}{2^k}(\theta^0 - \theta^*) + \left(1 - \frac{1}{2^k}\right)\mathbb{E}[\mathbf{v}] \\
&= \theta^* + \frac{1}{2^k}(\theta^0 - \theta^*) + \left(1 - \frac{1}{2^k}\right)\mathbf{X}^\top \underbrace{\mathbb{E}[\varepsilon]}_{=0} \\
&= \theta^* + \frac{1}{2^k}(\theta^0 - \theta^*).
\end{aligned}$$

More generally, you should remember that a linear combination of jointly Gaussian random variables is also Gaussian. In particular, remember the following: If we have a random Gaussian vector z given by

$$z \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

then given a fixed vector $\mathbf{z}^*$ and another matrix $\mathbf{A}$, the random vector $z_{\text{new}} = \mathbf{z}^* + \mathbf{Az}$ has a Gaussian distribution that satisfies

$$z_{\text{new}} = \mathbf{z}^* + \mathbf{Az} \sim \mathcal{N}(\mathbf{z}^* + \mathbf{A}\boldsymbol{\mu}+, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top).$$

In our case, we have

$$\theta^k = \theta^* + \frac{1}{2^k}(\theta^0 - \theta^*) + z$$

such that $z \sim \mathcal{N}(0, (1 - \frac{1}{2^k})^2\sigma^2\mathbf{X}^\top\mathbf{X})$ and since $\mathbf{X}^\top\mathbf{X} = \mathbf{I}_d$, we have

$$\theta^k \sim \mathcal{N}(\theta^* + \frac{1}{2^k}(\theta^0 - \theta^*), \left(1 - \frac{1}{2^k}\right)^2 \sigma^2\mathbf{I}_d).$$

(d) (2 pts) **Compute the squared-bias of θ_k, i.e., $\|\mathbb{E}[\theta_k] - \theta^*\|_2^2$ in terms of k and the initial squared error $e_0^2 = \|\theta_0 - \theta^*\|_2^2$. How does the bias change as k increases to ∞?**

Ans. One can compute the squared bias directly from equation (4):

$$\mathbb{E}[\theta^k] - \theta^* = \frac{1}{2^k}(\theta^0 - \theta^*) \implies \left\|\mathbb{E}[\theta^k] - \theta^*\right\|_2^2 = \frac{1}{2^{2k}} \underbrace{\left\|\theta^0 - \theta^*\right\|_2^2}_{e_0^2}.$$

Clearly, the bias decreases to 0 as $k \to \infty$. One could also directly conclude that bias decreases to 0 as $k \to \infty$ noting that θ^k converges to the unbiased (since we have a linear model) OLS estimator as $k \to \infty$.

(e) (2 pts) **Compute the variance of θ_k, i.e., compute $\mathbb{E}[\|\theta_k - \mathbb{E}[\theta_k]\|_2]^2$ in terms of k, d and σ^2. How does the variance change as k increases to ∞?**
Ans. From the previous parts, we have

$$\textbf{Covariance}(\theta^k) = \mathbb{E}[(\theta^k - \mathbb{E}[\theta^k])(\theta^k - \mathbb{E}[\theta^k])^\top] = \left(1 - \frac{1}{2^k}\right)^2 \sigma^2 \mathbf{I}_d,$$

and noting that

$$\begin{aligned}\textbf{variance}(\theta^k) &= \mathbb{E}[\left\|\theta^k - \mathbb{E}[\theta^k]\right\|_2^2] \\ &\overset{(i)}{=} \textbf{trace}[\textbf{Covariance}(\theta^k)] \\ &= d\sigma^2 \left(1 - \frac{1}{2^k}\right)^2.\end{aligned}$$

Note the useful trick (i), which is based on swapping the trace and expectations. We now provide a proof of this trick.

Proof of (i): Note the following useful facts:

(a) For any vector a, we have $\mathbf{a}^\top \mathbf{a} = \textbf{trace}[\mathbf{a}\mathbf{a}^\top]$.

(b) For a random matrix A, we have $\mathbb{E}[\textbf{trace}(\mathbf{A})] = \textbf{trace}(\mathbb{E}[\mathbf{A}])$.

Thus, we have

$$\begin{aligned}\mathbb{E}[\left\|\theta^k - \mathbb{E}[\theta^k]\right\|_2^2] &= \mathbb{E}\left[\left(\theta^k - \mathbb{E}[\theta^k]\right)^\top \left(\theta^k - \mathbb{E}[\theta^k]\right)\right] \\ &= \mathbb{E}\left[\textbf{trace}(\theta^k - \mathbb{E}[\theta^k])(\theta^k - \mathbb{E}[\theta^k])^\top\right] \\ &= \textbf{trace}\left[\mathbb{E}(\theta^k - \mathbb{E}[\theta^k])(\theta^k - \mathbb{E}[\theta^k])^\top\right] \\ &= \textbf{trace}\left[\textbf{Covariance}(\theta^k)\right].\end{aligned}$$

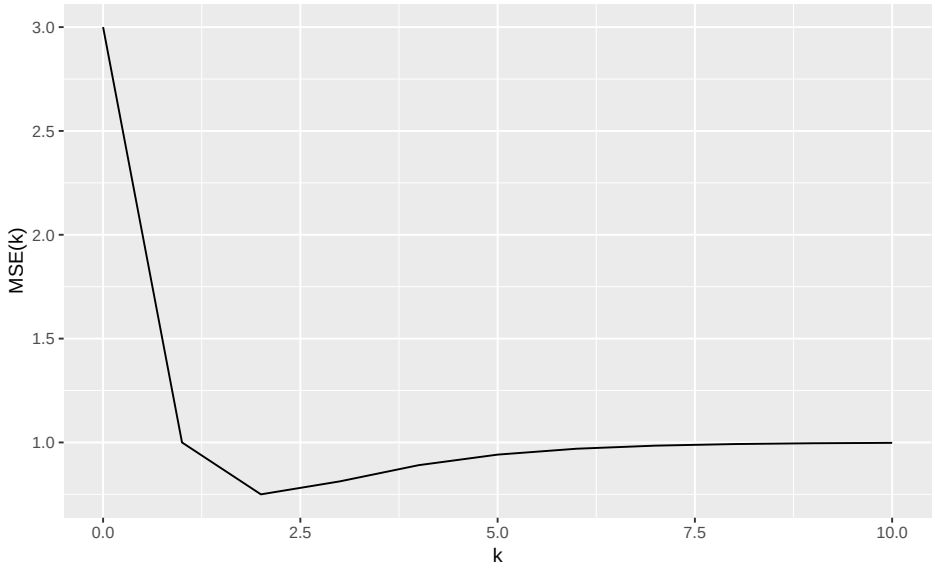
(f) (2 pts) **What is the mean square error MSE(k) error, i.e., $\mathbb{E}[\|\theta_k - \theta^*\|_2^2]$ of the k-th iterate? Manually draw an (approximate) curve of the MSE(k) error on y-axis versus k on the x-axis when $e_0^2 = 3$, $d = 1$ and $\sigma^2 = 1$.**
Hint: No additional calculations are necessary to derive MSE. You might want to use what we have shown in previous questions.
Ans. We saw quite a few times that for any estimator $\widehat{\theta}$, we can decomposes its mean-squared error as the sum of its squared bias and variance. Here the estimator is θ^k, and hence we have

$$\begin{aligned}
\textbf{MSE}(k) &= \mathbb{E}[\|\theta_k - \theta^*\|_2^2] \\
&= \underbrace{\|\mathbb{E}[\theta_k] - \theta^*\|_2^2}_{\textbf{Squared-bias}} + \underbrace{\mathbb{E}[\|\theta_k - \mathbb{E}[\theta_k]\|_2^2]}_{\textbf{Variance}} \\
&= \frac{1}{2^{2k}}\left\|\theta^0 - \theta^*\right\|_2^2 + d\sigma^2\left(1 - \frac{1}{2^k}\right)^2 \\
&= \frac{1}{2^{2k}}e_0^2 + d\sigma^2\left(1 - \frac{1}{2^k}\right)^2.
\end{aligned}$$

For the particular choice of values, we have

$$\textbf{MSE}(k) = \frac{3}{2^{2k}} + \left(1 - \frac{1}{2^k}\right)^2.$$

and one can obtain the following plot.



MSE vs iteration k for gradient descent

(g) (2 pts) **What is an optimal k for which the MSE(k) is minimized?**
Your answer should be in terms of σ^2, d and initial squared error $e_0 = \|\theta_0 - \theta^*\|_2^2$.
Hint: $\frac{d}{dk}a^k = a^k \ln a$ for $a > 0$.
Ans. We have

$$\textbf{MSE}(k) = \frac{1}{2^{2k}} e_0^2 + d\sigma^2 \left(1 - \frac{1}{2^k}\right)^2.$$

Using the given hint, we have

$$\begin{aligned}
\frac{d\textbf{MSE}(k)}{dk} = = \frac{d}{dk}\left(\frac{1}{2^{2k}} e_0^2 + d\sigma^2 \left(1 - \frac{1}{2^k}\right)^2\right) \\
= \frac{d}{dk}\left(\frac{1}{4^k} e_0^2 + d\sigma^2 \left(1 - \frac{1}{2^k}\right)^2\right) \\
= e_0^2 \frac{1}{4^k} \ln 4 - d\sigma^2 2\left(1 - \frac{1}{2^k}\right) \cdot \frac{1}{2^k} \ln 2 \\
= 2 \ln 2 \cdot \frac{1}{2^k}\left[e_0^2 \frac{1}{2^k} - d\sigma^2 \left(1 - \frac{1}{2^k}\right)\right].
\end{aligned}$$

Setting the derivative to zero, we obtain

$$2^k d\sigma^2 = e_0^2 + d\sigma^2 \implies k^* = \log_2\left(1 + \frac{e_0^2}{d\sigma^2}\right).$$

Thus, we see that only a few steps of gradient descent can obtain a better estimate than the OLS estimator. The catch is we don't know the different quantities to begin with!

Note that there are two missing points in the argument above: We should also verify that the second derivative at $k = k^*$ should be positive for it to be a minima. Also, k takes discrete values so one has to be careful in making such arguments, but for the purpose of the exam; we did not focus on these aspects.

5 EDA with MPG DATA (4 parts, 10 pts)

In this problem we consider some EDA steps with the Auto MPG dataset. This dataset contains a subset of the fuel economy data that the EPA (Environmental Protection Agency) makes available.

(a) (1 pt) In the two figures below, you are given two scatter-plots of 398 data points using two features "weight" and "acceleration" of the Auto MPG dataset. We have two data frames: Data 1–without any preprocessing. Data 2–after some standard preprocessing. Figure 1 corresponds to the scatter plot of Data 1 and Figure 2 corresponds to the scatter plot of Data 2. **What kind of standardization do you think was done?**

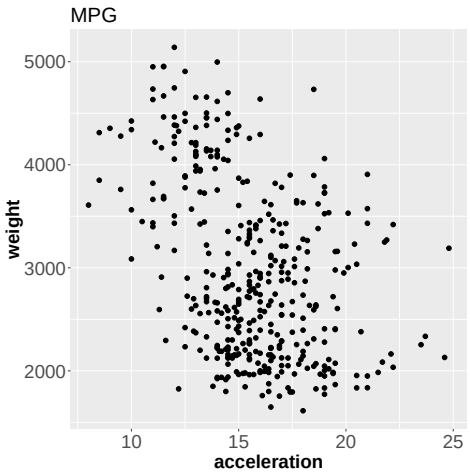


Fig 1. Scatter plot of Data 1

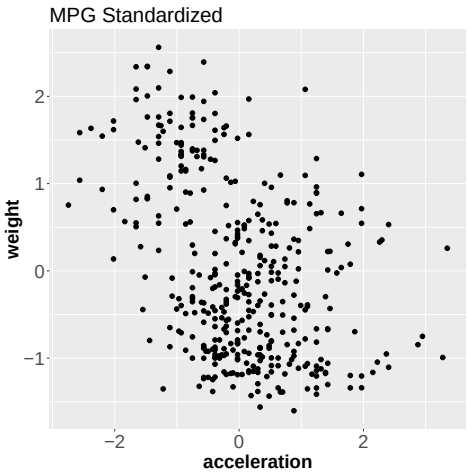


Fig 2. Scatter plot of Data 2

Ans. It seems that both the axes were centered (removing their own mean) and scaled (by their own standard deviation).

(b) (3 pts) **Which data frame (Data 1 or Data 2) would you prefer to use to compute the principal vectors? Why? How many principal components does the data have? Plot as many principal components as you can on the figure corresponding to your answer.**
Ans. Data 2, because it is standardized and the scales of the two axes are different. We will have 2 PCs since the data is two dimensional.

(c) (2 pts) Suppose we use K-means or to estimate the hidden clusters for Data 2. **What choice of K would you use in K-means? What happens if we decide to fit a Gaussian mixture model on Data 2, i.e., how many mixtures would you try to fit? Give a brief reasoning.**
Ans. From the scatter-plot, a suitable choice would be $K = 2$ or 3 is possible. Similar number of mixtures with EM is a valid answer too.

(d) (4 pts) We fit both K-means and Gaussian mixture model with EM algorithm on Data 1 (Figure 1) and Data 2 (Figure 2), i.e., two procedures on both the data frames and in total four results. But these four results can be *qualitatively summarized* in the two figures below (meaning that some results were similar). The x and y axis ticks have been omitted deliberately. **Can you guess which Figure (Figure 3 or Figure 4) corresponds to the results for the different procedures on each dataset?**

(i) K-means on raw data: Figure 3

(ii) K-means standardized data: Figure 4

(iii) EM on raw data: Figure 4

(iv) EM on standardized data: Figure 4

Simply enter the figure number. No justification required.
Ans. Since covariance is also estimated in EM, it is usually unaffected by scaling and hence recovers approximately the correct estimates for both datasets. For K-means the scaling matters (unless we change the objective function), and as a result it finds reasonable estimates when the data is scaled and has (almost) nicely separated clusters.

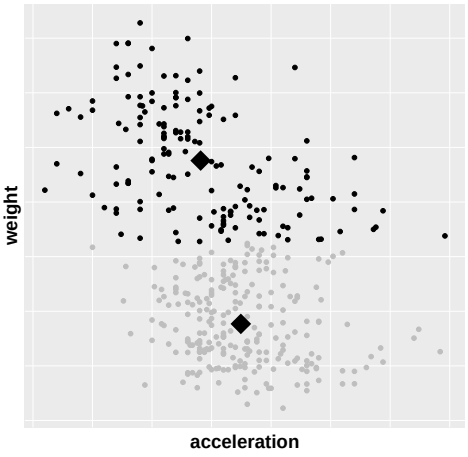


Figure 3

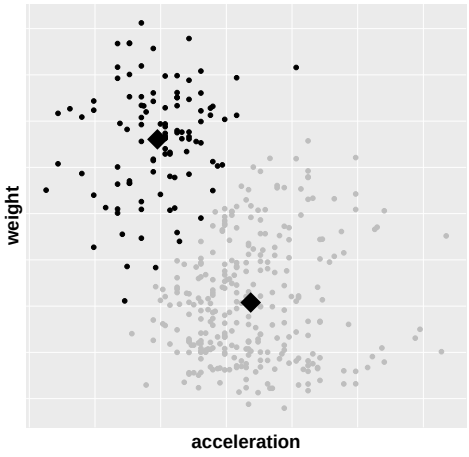


Figure 4