Name: Kevin Zhang

Problem 1.

We are trying hire the best candidate. We have n candidates $\pi_1, \pi_2, ..., \pi_n$ but we can only hire one. Our interview process is as follows: We interview and reject the first $\pi_1, ...\pi_{n/e}$ candidates, and then hire the first candidate from the $\pi_{n/e+1}, ..., \pi_n$ order that outperforms all of the first n/e candidates.

(a) For an index $i > n/e$, let E_i be the event that π_i is the best candidate. Find $P(\pi_i \text{ is hired } |E_i)$.

In this scenario, π_i is hired if no one is hired from the candidates of $\pi_1, ...\pi_{i-1}$. Let π_k be the best candidate in this pool. Since we are conditioning on E_i, we can assume that $\pi_i > \pi_k$. The first question is where can π_k go? We can break up the pool as follows:

$$\underbrace{\pi_1, \pi_2, ..., \pi_{n/e}}_{\text{always rejected}}, \underbrace{\pi_{n/e+1}, ...\pi_{i-1}}_{\text{always hired}}$$

Since π_k is the best candidate in this pool, the only place it can go is in the first n/e spots. Otherwise, π_k will be hired instead of π_i. Since we have $i-1$ spots, total, we can express this as:

$$P(\pi_k \text{ not hired}) = \frac{n/e}{i-1}$$

The second question is how many choices of π_k we can have. The only condition we know is the $\pi_i > \pi_k$, but whether π_k is 2nd best, 3rd best, we don't really know. But, this is also easy to figure out, because we have a limited pool of candidates. The way to think of it is π_k is put into the first $i-1$ pool. But the candidates better than π_k that is not π_i must go after π_i, because otherwise they would come before, thereby replacing π_k. Thus we have the following:

$$\text{\# of choices for } \pi_k = n - (i-1)$$

Combine the two parts to get the probability of hiring π_i:

$$P(\pi_i \text{ is hired } |E_i) = \left(\frac{n/e}{i-1}\right)(n-(i-1))$$

(b) If π^* is the best candidate, find an approximation for $P(\pi^* \text{ is hired})$.

We can express this is a summation over all choices of i:

$P(\pi^* \text{ is hired}) = \sum_{i=1}^{n} P(\pi_i \text{ is hired}|E_i)P(E_i)$.

Since π_i won't be hired if it falls into the first n/e candidates, we can reduce the expression:

$P(\pi^* \text{ is hired}) = \sum_{i=n/e+1}^{n} P(\pi_i \text{ is hired}|E_i)P(E_i)$.

Expanding this, and then letting $j = i-1$ gets us:

$$
\begin{aligned}
P(\pi^* \text{ is hired}) &= \sum_{i=n/e}^{n} P(\pi_i \text{ is hired}|E_i)P(E_i) \\
&= \sum_{i=n/e+1}^{n} \left(\frac{(n/e)(n-(i-1))}{(i-1)}\right)\left(\frac{1}{n}\right) \\
&= \sum_{j=n/e}^{n} \left(\frac{(n/e)(n-j)}{j}\right)\left(\frac{1}{n}\right) \\
&= \sum_{j=n/e}^{n} \frac{(n-j)}{ej} \\
&= \sum_{j=n/e}^{n} \frac{n}{ej} - \frac{j}{ej} \\
&= \sum_{j=n/e}^{n} \frac{n}{ej} - \sum_{j=n/e}^{n} \frac{1}{e} \\
&= \sum_{j=n/e}^{n} \frac{n}{ej} - \left((n-n/e+1)\frac{1}{e}\right)
\end{aligned}
$$

We still have to deal with the summation term. Here, we can use the approximation $\int_a^{b+1} \frac{dx}{x} \leq \sum_{i=a}^{b} \frac{1}{i} \leq \int_{a-1}^{b} \frac{dx}{x}$. Then, we can express the summation as follows:

$$
\int_{n/e}^{n+1} \frac{dx}{x} \leq \sum_{j=n/e}^{n} \frac{n}{ej} \qquad \leq \int_{n/e-1}^{n} \frac{dx}{x} \tag{1}
$$

$$
\left((n-n/e+1)\frac{n}{e}\right) \leq \sum_{j=n/e}^{n} \frac{n}{ej} \leq \left((n-n/e+1)\frac{n}{e}\right) \tag{2}
$$

Thus, our final probability is:

$$
P(\pi^* \text{ is hired}) = \left((n-n/e+1)\frac{n-1}{e}\right)
$$

Problem 2.

Consider the following linear program

$$
\begin{aligned}
&\min 2x + 5y - 3z \\
&\text{subject to} \\
&2x - y + 2z \geq 3 \\
&x - y - 2z \geq -1 \\
&-x + y + 5z \geq 1 \\
&x + y \geq 2 \\
&x, y, z \geq 0
\end{aligned}
$$

(a) Find the dual of the linear program.

The dual formulation can be expressed from the original linear program with the following transpose:

Linear	**Dual**
$\min c^T x$	$\max b^T x$
$Ax \geq b$	$A^T x \leq c$
$x \geq 0$	$x \geq 0$

As such, the values for A, b, c^T can be pulled from the original linear program:

$$
A = \begin{bmatrix} 2 & -1 & 2 \\ 1 & -1 & -2 \\ -1 & 1 & 5 \\ 2 & 1 & 0 \end{bmatrix} \quad b = \begin{bmatrix} 3 \\ -1 \\ 1 \\ 2 \end{bmatrix} \quad c^T = \begin{bmatrix} 2 & 5 & -3 \end{bmatrix}
$$

Applying the transpose, and formulated as a dual, we get this:

$$
\begin{aligned}
&\max 3w - x + y + 2z \\
&\text{subject to} \\
&2w + x - y + z \leq 2 \\
&-w - x + y + z \leq 5 \\
&2w - 2x + 5y \leq -3 \\
&w, x, y, z \geq 0
\end{aligned}
$$

(b) At $(x, y, z) = (2, 0, 3/2)$, the linear program has objective value $-1/2$. Show this is the optimal solution.

We can show this via linear combination. First, we can organize the program as matrix vectors:

$$\begin{aligned} v_1 &= 2x - y + 2z \geq 3 \\ v_2 &= x - y - 2z \geq -1 \\ v_3 &= -x + y + 5z \geq 1 \\ v_4 &= x + y \geq 2 \end{aligned}$$

Then, we need to find a linear combination B such that expressing $2x + 5y - 3z$ in this basis is solvable. (Note, this was reversed engineered from the objective value $-1/2$. But for the proof, we can *magically* find such a basis):

$$b_1 = v_2 + v_4 \qquad = 2x - 2z \geq 1 \tag{3}$$

$$b_2 = v_3 - v_1 \quad = -3x + 2y + 3z \geq -2 \tag{4}$$

$$b_3 = v_4 - 2v_1 = -3x + 3y - 4z \geq -4 \tag{5}$$

With a basis $\mathcal{B} = \text{span}\{\begin{bmatrix} 2 \\ 0 \\ -2 \end{bmatrix}, \begin{bmatrix} -3 \\ 2 \\ 3 \end{bmatrix} \begin{bmatrix} -3 \\ 3 \\ -4 \end{bmatrix}\}$ we would like to express $\begin{bmatrix} 2 \\ 5 \\ -3 \end{bmatrix}$ in terms of this basis. This can be thought of as the following:

$$\begin{bmatrix} 2 & -3 & -3 \\ 0 & 2 & 3 \\ -2 & 3 & -4 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 2 \\ 5 \\ -3 \end{bmatrix}$$

Solving for a, b, c nets us $a = \frac{65}{14}, b = \frac{16}{7}, c = \frac{1}{7}$. Plugging into our basis vectors, and we get the following:

$$\begin{aligned} ab_1 + bb_2 + cb_3 \\ \frac{65}{14}(2x - 2z) + \frac{16}{7}(-3x + 2y + 3z) + \frac{1}{7}(-3x + 3y - 4z) &\geq \frac{65}{14}(1) + \frac{16}{7}(-2) + \frac{1}{7}(-4) \\ \frac{1}{14}[65(2x - 2z) + 32(-3x + 2y + 3z) + 2(-3x + 3y - 4z)] &\geq \frac{1}{14}[65(1) + 32(-2) + 2(-4)] \\ \frac{1}{14}[130x - 130z - 96x + 64y + 96z + 6x + 6y - 8z] &\geq \frac{1}{14}[65(1) + 32(-2) + 2(-4)] \\ \frac{1}{14}[28x + 70y - 42z] &\geq \frac{1}{14}[-7] \\ 2x + 5y - 3z &\geq -1/2 \end{aligned}$$

Since the minimum value for $2x + 5y - 3z$ is $-1/2$, it must be the optimal solution.

Problem 3.

We are given two $n \times d$ matrices A and B where $n \gg d$. Let $a_1, ..., a_n$ be the column vectors corresponding to the rows of A, and similarly for $b_1, ..., b_n$. We would like to compute $P = A^T B$ quickly. In this case, we will pick a $r \in 1, 2, ..., n$ with probability $p_r = \frac{||a_r|| \cdot ||b_r||}{\sum_{j=1}^{n} ||a_j|| \cdot ||b_j||}$. Our approximation is then $\hat{P} = \frac{1}{p_r} a_r b_r^T$.

(a) Our error can be expressed in terms of $||P - \hat{P}||_F$. Show that

$$\mathbb{E}\left[||P - \hat{P}||_F^2\right] \leq \left(\sum_{i=1}^{n} ||a_i|| \cdot ||b_i||\right)^2$$

The Frobenius norm is the sum of the squares of the entries. $P = A^T B$ is a $d \times d$ matrix:

$$\begin{aligned}
\mathbb{E}\left[||P - \hat{P}||_F^2\right] &= \mathbb{E}\left[\sum_{u=1}^{d} \sum_{v=1}^{d} (P_{uv} - \hat{P}_{uv})^2\right] \\
&\leq \mathbb{E}\left[\sum_{u=1}^{d} \sum_{v=1}^{d} (\hat{P}_{uv})^2\right] \\
&\leq \mathbb{E}\left[\sum_{u=1}^{d} \sum_{v=1}^{d} (\frac{1}{p_r}([a_r b_r^T]_{uv}))^2\right] \\
&\leq \mathbb{E}\left[\sum_{u=1}^{d} \sum_{v=1}^{d} \left(\frac{\sum_{i=1}^{n} ||a_i|| \cdot ||b_i||}{||a_r|| \cdot ||b_r||}\right)^2 ([a_r b_r^T]_{uv})^2\right] \\
&\leq \mathbb{E}\left[\left(\frac{\sum_{i=1}^{n} ||a_i|| \cdot ||b_i||}{||a_r|| \cdot ||b_r||}\right)^2 \sum_{u=1}^{d} \sum_{v=1}^{d} ([a_r b_r^T]_{uv})^2\right]
\end{aligned}$$

The expression $\sum_{u=1}^{d} \sum_{v=1}^{d} ([a_r b_r^T]_{uv})^2$ can be reduced by examining the matrix $a_r b_r^T$:

$$a_r b_r^T = \begin{bmatrix} a_r^{(1)} b_r^{(1)} & a_r^{(1)} b_r^{(2)} & ... & a_r^{(1)} b_r^{(d)} \\ a_r^{(2)} b_r^{(1)} & a_r^{(2)} b_r^{(2)} & ... & a_r^{(2)} b_r^{(d)} \\ \vdots & \vdots & \ddots & \vdots \\ a_r^{(d)} b_r^{(1)} & a_r^{(d)} b_r^{(2)} & ... & a_r^{(d)} b_r^{(d)} \end{bmatrix}$$

If we square all the entries, and then apply the summation, we get the following:

$$\sum_{u=1}^{d}\sum_{v=1}^{d}([a_r b_r^T]_{uv})^2 = \left(\begin{array}{ccccccccc} & (a_r^{(1)}b_r^{(1)})^2 & + & (a_r^{(1)}b_r^{(2)})^2 & + & \dots & + & (a_r^{(1)}b_r^{(d)})^2 \\ + & (a_r^{(2)}b_r^{(1)})^2 & + & (a_r^{(2)}b_r^{(2)})^2 & + & \dots & + & (a_r^{(2)}b_r^{(d)})^2 \\ + & \dots & & & & & & \\ + & (a_r^{(d)}b_r^{(1)})^2 & + & (a_r^{(1)}b_r^{(2)})^2 & + & \dots & + & (a_r^{(d)}b_r^{(d)})^2 \end{array}\right)$$

$$= \left(\begin{array}{ccc} & (a_r^{(1)})^2 & \left[\begin{array}{ccccccc} (b_r^{(1)})^2 & + & (b_r^{(2)})^2 & + & \dots & + & (b_r^{(d)})^2 \end{array}\right] \\ + & (a_r^{(2)})^2 & \left[\begin{array}{ccccccc} (b_r^{(1)})^2 & + & (b_r^{(2)})^2 & + & \dots & + & (b_r^{(d)})^2 \end{array}\right] \\ + & \dots & \\ + & (a_r^{(d)})^2 & \left[\begin{array}{ccccccc} (b_r^{(1)})^2 & + & (b_r^{(2)})^2 & + & \dots & + & (b_r^{(d)})^2 \end{array}\right] \end{array}\right)$$

$$= ||a_r||^2 \cdot ||b_r||^2$$

$$= (||a_r|| \cdot ||b_r||)^2$$

Plug this back into our original problem and we get the following:

$$\begin{aligned} \mathbb{E}\left[||P - \hat{P}||_F^2\right] &\leq \mathbb{E}\left[\left(\frac{\sum_{i=1}^{n} ||a_i|| \cdot ||b_i||}{||a_r|| \cdot ||b_r||}\right)^2 \sum_{u=1}^{d}\sum_{v=1}^{d}([a_r b_r^T]_{uv})^2\right] \\ &= \mathbb{E}\left[\left(\frac{\sum_{i=1}^{n} ||a_i|| \cdot ||b_i||}{||a_r|| \cdot ||b_r||}\right)^2 (||a_r|| \cdot ||b_r||)^2\right] \\ &= \left(\sum_{i=1}^{n} ||a_i|| \cdot ||b_i||\right)^2 \quad \checkmark \end{aligned}$$

(b) Instead of using 1 sample r, we use m i.i.d. samples $r_1, ..., r_m$ and then take the average result. The new estimate is $\hat{P} = \frac{1}{m}\sum_{i=1}^{m} \frac{1}{p_{r_i}} a_{r_i} b_{r_i}^T$. What happens to the bound of $\mathbb{E}\left[\|P - \hat{P}\|_F^2\right]$?

We can use the same logic as above to manipulate terms:

$$
\begin{aligned}
\mathbb{E}\left[\|P - \hat{P}\|_F^2\right] &= \mathbb{E}\left[\sum_{u=1}^{d}\sum_{v=1}^{d}(P_{uv} - \hat{P}_{uv})^2\right] \\
&\leq \mathbb{E}\left[\sum_{u=1}^{d}\sum_{v=1}^{d}(\hat{P}_{uv})^2\right] \\
&\leq \mathbb{E}\left[\sum_{u=1}^{d}\sum_{v=1}^{d}\left(\frac{1}{m}\sum_{i=1}^{m}\frac{1}{p_{r_i}}([a_{r_i} b_{r_i}^T]_{uv})\right)^2\right] \\
&\leq \mathbb{E}\left[(\frac{1}{m})^2\sum_{i=1}^{m}(\frac{1}{p_{r_i}})^2\sum_{u=1}^{d}\sum_{v=1}^{d}\left([a_{r_i} b_{r_i}^T]_{uv})\right)^2\right] \\
&\leq \mathbb{E}\left[(\frac{1}{m})^2\sum_{i=1}^{m}\left(\sum_{i=1}^{n}\|a_i\| \cdot \|b_i\|\right)^2\right] \\
&\leq \frac{1}{m}\left(\sum_{i=1}^{n}\|a_i\| \cdot \|b_i\|\right)^2
\end{aligned}
$$

The error gets reduced by a factor of $\frac{1}{m}$

Problem 4.

Consider a function $f(x) : \mathbb{R}^2 \to \mathbb{R}$ with

$$f(x) = \frac{7}{2}x_1^2 + 4x_1x_2 + \frac{13}{2}x_2^2 = \frac{1}{2}x^T \begin{bmatrix} 7 & 4 \\ 4 & 13 \end{bmatrix} x$$

(a) Show $f(x)$ is β-smooth for as small β as you can.

β-smoothness is defined by:

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\|$$

We can find $\nabla f(x)$ using the Jacobian matrix:

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 7x_1 + 4x_2 \\ 4x_1 + 13x_2 \end{bmatrix} = \begin{bmatrix} 7 & 4 \\ 4 & 13 \end{bmatrix} x$$

We can apply this to the definition of β-smoothness:

$$\begin{aligned} \|\nabla f(x) - \nabla f(y)\| &\leq \beta \|x - y\| \\ \left\| \begin{bmatrix} 7 & 4 \\ 4 & 13 \end{bmatrix} x - \begin{bmatrix} 7 & 4 \\ 4 & 13 \end{bmatrix} y \right\| &\leq \beta \|x - y\| \\ \left\| \begin{bmatrix} 7 & 4 \\ 4 & 13 \end{bmatrix} \right\| \cdot \|x - y\| &\leq \beta \|x - y\| \\ \left\| \begin{bmatrix} 7 & 4 \\ 4 & 13 \end{bmatrix} \right\| &\leq \beta \\ \left\| \begin{bmatrix} 7 & 4 \\ 4 & 13 \end{bmatrix} \right\|_2 &\leq \beta \\ \sigma_1 &\leq \beta \end{aligned}$$

Computing the SVD of $\begin{bmatrix} 7 & 4 \\ 4 & 13 \end{bmatrix}$, we have that $\sigma_1 = 15$. Thus, $\beta = 15$.

(b) Suppose we run gradient descent from starting point $x^{(0)} = (-1, 3)^T$ with constant step size $\eta = \frac{1}{\beta}$. Derive a closed form expression for $x^{(t)}$ and $f(x^{(t)})$.

We can derive a closed form expression for $x^{(t)}$ as follows:

$$
\begin{aligned}
x^{(t)} &= x^{(t-1)} - \eta \nabla f(x^{(t-1)}) \\
&= x^{(t-1)} - \eta \begin{bmatrix} 7 & 4 \\ 4 & 13 \end{bmatrix} x^{(t-1)} \\
&= \left(I - \frac{1}{\beta} \begin{bmatrix} 7 & 4 \\ 4 & 13 \end{bmatrix} \right) x^{(t-1)} \\
&= \left(I - \frac{1}{15} \begin{bmatrix} 7 & 4 \\ 4 & 13 \end{bmatrix} \right)^t x^{(0)} \\
&= \left(\frac{1}{15} \right)^t \begin{bmatrix} 8 & 4 \\ 4 & 2 \end{bmatrix}^t x^{(0)} \\
&= \left(\frac{1}{15} \right)^t \left(\underbrace{\begin{bmatrix} \frac{2\sqrt{5}}{5} & \frac{-\sqrt{5}}{5} \\ \frac{\sqrt{5}}{5} & \frac{2\sqrt{5}}{5} \end{bmatrix} \begin{bmatrix} 10 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{2\sqrt{5}}{5} & \frac{\sqrt{5}}{5} \\ \frac{-\sqrt{5}}{5} & \frac{2\sqrt{5}}{5} \end{bmatrix}}_{\text{SVD of inner}} \right)^t x^{(0)} \\
&= \left(\frac{1}{15} \right)^t \begin{bmatrix} \frac{2\sqrt{5}}{5} & \frac{-\sqrt{5}}{5} \\ \frac{\sqrt{5}}{5} & \frac{2\sqrt{5}}{5} \end{bmatrix} \begin{bmatrix} 10 & 0 \\ 0 & 0 \end{bmatrix}^t \begin{bmatrix} \frac{2\sqrt{5}}{5} & \frac{\sqrt{5}}{5} \\ \frac{-\sqrt{5}}{5} & \frac{2\sqrt{5}}{5} \end{bmatrix} x^{(0)} \\
&= \left(\frac{1}{15} \right)^t \begin{bmatrix} (\frac{4}{5})(10^t) & (\frac{2}{5})(10^t) \\ (\frac{2}{5})(10^t) & (\frac{1}{5})(10^t) \end{bmatrix} \begin{bmatrix} -1 \\ 3 \end{bmatrix} \\
&= \begin{bmatrix} \frac{2}{5}(\frac{2}{3})^t \\ \frac{1}{5}(\frac{2}{3})^t \end{bmatrix}
\end{aligned}
$$

As $x^{(t)}$ must be a vector, the final answer is $x^{(t)} = \left(\frac{2}{5}(\frac{2}{3})^t, \frac{1}{5}(\frac{2}{3})^t \right)^T$.

For $f(x^{(t)})$, the process is similar, but also different. We can rely on the fact that $x^{(t)} = \frac{2}{3} x^{(t-1)}$:

$$
\begin{aligned}
f(x^{(t)}) &= \frac{1}{2} x^{(t)^T} \begin{bmatrix} 7 & 4 \\ 4 & 13 \end{bmatrix} x^{(t)} \\
&= (\frac{2}{3})^2 \cdot \frac{1}{2} x^{(t-1)^T} \begin{bmatrix} 7 & 4 \\ 4 & 13 \end{bmatrix} x^{(t-1)} \\
&= (\frac{2}{3})^2 \cdot f(x^{(t-1)}) \\
&= (\frac{2}{3})^{2t} \cdot f(x^{(0)})
\end{aligned}
$$

Solving for $f(x^0) = 50$ results in $f(x^{(t)}) = (\frac{2}{3})^{2t} \cdot 50$.