IMT3150: Optimización Avanzada 23 de mayo de 2025

Tarea 2

Alumno: Vicente Opazo *Profesor: Cristóbal Guzmán*

Pregunta 1. *(a) Si $x \in \mathcal{X}$ entonces la proyección es si mismo trivialmente, así que supongamos que $x \notin \mathcal{X}$.*

Primero reduzcamos el problema al caso $x \geq 0$. Para esto primero consideremos el operador sign *definido como*

$$\text{sign}(x)_i = \begin{cases} 1 & \textit{si } x_i \geq 0 \\ -1 & \textit{si } x_i < 0 \end{cases}$$

Primero veamos que la proyección de x, llamémosla z, cumple que $\text{sign}(x) = \text{sign}(z)$. Supongamos por contradicción que existiese i tal que $\text{sign}(x)_i \neq \text{sign}(z)_i$, entonces reemplazando z_i por 0, reduce la norma 1 de z (así que sigue estando en $\mathcal{X}$) y reduce la norma 2 de $(x - z)$ (puesto que reduce el aporte de dicha componente a la suma, junto a que el operador raíz cuadrada es monótono), por lo cual z no puede ser la proyección ya que hay otro punto más cercano, absurdo. Así, se concluye que los signos coinciden componente a componente.

Ahora notemos que la bola $\overline{\mathcal{B}_{\|\cdot\|_1}}(0,1)$ es simétrica, puesto que

$$\|x\|_1 = \||x|\|_1$$

Así, si $x_i < 0$, entonces el z de la proyección cumple que $z_i < 0$. Consideremos ahora z', x' tales que coinciden con z, x respectivamente en todas las componentes menos en la i-ésima, en la cual consideramos $z'_i = -z_i, x'_i = -x_i$. Por ser $\mathcal{X}$ simétrico entonces $z' \in \mathcal{X}$. Además, en el aporte de la i-ésima componente a la norma 2 de $(x-z)$ se tiene que aporta $(x_i - z_i)^2 = (-x_i + z_i)^2 = (x'_i - z'_i)$ y entonces z' corresponde a la proyección de x', puesto que si no lo fuera entonces z no sería la proyección de x, absurdo. Así, podemos cambiar el signo de alguna componente de x y obtener un problema equivalente al que luego le tendremos que cambiar el signo respectivo a la proyección.

Repitiendo el procedimiento anterior para todos los signos negativos de x se concluye que podemos resolver el problema para $x' = |x|$, obtener la proyección z' de x' y luego la proyección original z de x es simplemente $z_i = \text{sign}(x)_i \cdot z'_i$. Y como $x' \geq 0$ por definición, se concluye finalmente que el problema puede reducirse al caso $x \geq 0$.

Ahora, formulemos el problema de proyección como un problema de optimización, tenemos que el problema corresponde a (sacando el valor absoluto puesto que $z \geq 0$)

$$\min_z \frac{1}{2}\|x - z\|_2^2 \quad \textit{s.a} \quad \sum_{i=1}^{d} z_i \leq 1, \quad z \geq 0$$

Y entonces, el Lagrangiano del problema corresponde a

$$\mathcal{L}(z, \lambda, \mu) = \frac{1}{2}\|x - z\|_2^2 + \lambda \left(\sum_{i=1}^{d} z_i - 1 \right) - \mu^T z$$

donde $\lambda \geq 0$ es el multiplicador de Lagrange para la restricción ℓ_1 y $\mu \geq 0$ para las restricciones de no negatividad.

Así, las condiciones KKT corresponden a

1. *Estacionariedad: Derivando con respecto a z resulta que*
$$z_i - x_i + \lambda - \mu_i = 0$$
2. *Factibilidad primal:* $\sum_{i=1}^{d} z_i \leq 1, \quad z \geq 0$
3. *Factibilidad dual:* $\lambda \geq 0$, $\mu \geq 0$
4. *Holgura complementaria:*
$$\lambda \left(\sum_{i=1}^{d} z_i - 1 \right) = 0, \quad \mu_i z_i = 0 \quad \forall i$$

Para encontrar las soluciones al sistema KKT primero notemos que podemos asumir que $\|z\|_1 = 1$ ya que la proyección debe estar en la frontera (puesto que inicialmente asumimos que $x \notin \mathcal{X}$). De esto, se tiene que la solución para $z_i > 0 \Rightarrow \mu_i = 0$ corresponde a
$$z_i = x_i - \lambda$$
Mientras que para $z_i = 0$, se tiene que
$$\mu_i = \lambda - x_i$$
Y como $\mu_i \geq 0$ entonces $\lambda \geq x_i$. De esta forma, juntando los dos casos la solución corresponde a
$$z_i = \max(x_i - \lambda, 0)$$
Así, nos basta determinar λ para poder encontrar la solución. La condición $\sum_{i=1}^{d} z_i = 1$ implica que
$$\sum_{i=1}^{d} \max(x_i - \lambda, 0) = 1$$
Entonces lo que podemos hacer es ordenar los valores de x de mayor a menor. De esta forma, sabemos que los z correspondientes que aportarán a la suma son un prefijo, pues al fijar λ solo los mayores que el aportan. Es claro además que la suma en función de λ es monótona, por lo cual podemos buscar el valor de λ desde valores mayores a menores, obteniendo desde sumas menores a 1 hasta mayores a 1 de forma monótona.

Así, con x ya ordenado vamos a probar uno a uno considerar $\lambda = x_k$. Entonces la suma corresponde a
$$\sum_{i=1}^{d} \max(x_i - \lambda, 0) = \sum_{i=1}^{k} (x_i - x_k) = S_k - k x_k := R_k$$
Donde S_k es la suma acumulada de los x_i, y R_k corresponde a la suma acumulada de los z_i con el valor de λ como x_k. Así, buscamos el mayor k tal que se cumpla $R_k \leq 1$, y por la monotonía podemos concluir que $\lambda \in [x_k, x_{k+1})$ donde consideramos $x_{d+1} = \infty$.

Ahora que ya sabemos los elementos que aportaran a la suma, ya que encontramos en el intervalo en que vive λ, sabemos que solo me aportan los primeros k a la suma. Así, podemos calcular λ directamente de la ecuación.
$$\sum_{i=1}^{k} (x_i - \lambda) = S_k - k\lambda = 1 \Rightarrow \lambda = \frac{S_k - 1}{k}$$
Por lo que los pasos del algoritmo son

1. *Verificar si $x \in \mathcal{X}$ sumando sus componentes en $O(d)$*
2. *Cambiar (y recordar) los signos negativos de x en $O(d)$, ahora tenemos $x \geq 0$*
3. *Ordenar x de mayor a menor en $O(d \log d)$*
4. *Calcular S_k en $O(d)$*
5. *Calcular R_k en $O(d)$*
6. *Encontrar el mayor k tal que $R_k \leq 1$ en $O(d)$*
7. *Calcular $\lambda = \frac{S_k - 1}{k}$ en $O(1)$*
8. *Calcular $z_i = \max(x_i - \lambda, 0)$ en $O(d)$*
9. *Restaurar los signos al z según lo recordado en el paso 2 en $O(d)$*

Por lo cual encontramos un algoritmo $O(d \log d)$ que resuelve el problema, tal como se había solicitado.

(b) Intuitivamente los puntos extremos debieran ser las matrices que tienen un único autovalor igual a 1 y el resto igual a 0. Esto es, los puntos extremos son las matrices de la forma (notemos que además es de rango 1)

$$A = \pm uu^T, \quad \text{con } u \in \mathbb{R}^d : \|u\|_2 = 1$$

Ahora veámoslo con más rigor. Primero veamos que si A no es de esa forma entonces no puede ser punto extremo.

Si A tiene al menos dos autovalores distintos a cero entonces por la descomposición espectral A se puede escribir como $A = U\Lambda U^T$ con U ortogonal y $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d)$ donde por hipótesis $\lambda_1, \lambda_2 > 0$. Ahora consideremos las matrices

$$A^+ = U\Lambda^+ U^T, \quad A^- = U\Lambda^- U^T$$

con $\Lambda^+ = \text{diag}(\lambda_1 + \epsilon, \lambda_2 - \epsilon, \lambda_3, \ldots, \lambda_d)$ y $\Lambda^- = \text{diag}(\lambda_1 - \epsilon, \lambda_2 + \epsilon, \lambda_3, \ldots, \lambda_d)$, donde el ϵ es tal que $\lambda_1 \pm \epsilon \in [0, 1]$ y $\lambda_2 \pm \epsilon \in [0, 1]$, el cual existe ya que dichos dos autovalores son no nulos. Con lo anterior, A se puede escribir como

$$A = \frac{1}{2}A^+ + \frac{1}{2}A^-$$

Y como A^+, A^- son distintas, entonces A no puede ser un punto extremo, de lo que se concluye que si A es punto extremo entonces tiene un único autovalor (demostramos que debe tener 1 o 0, pero este segundo caso se descarta trivialmente ya que la matriz nula se puede escribir fácilmente como combinación de otras dos).

Si dicho único autovalor no es igual a 1, entonces debe ser menor a uno puesto que por definición se tiene que $\|A\|_{nuc} \leq 1$. Pero como es menor estricto a 1, entonces se cumple que

$$0 < \|A\|_{nuc} < 1$$

y entonces existe ϵ tal que

$$0 < \|(1 - \epsilon)A\|_{nuc} < \|(1 + \epsilon)A\|_{nuc} < 1$$

Y entonces podríamos escribir

$$A = \frac{1}{2}(1 - \epsilon)A + \frac{1}{2}(1 + \epsilon)A$$

Y entonces no es punto extremo. Sumado a lo anterior, ya demostramos la primera implicancia.

Ahora veamos que si A es de la forma, sin pérdida de generalidad, $A = uu^T$ con $\|u\|_2 = 1$, entonces debe ser punto extremo. Supongamos por contradicción pudieramos escribir A como

$$A = \lambda A_1 + (1-\lambda)A_2$$

con A_1, A_2 en $\mathcal{X}$ y $\lambda \in (0,1)$. Lo primero es notar que puesto que $\|A\|_{nuc} = 1$, entonces necesariamente $\|A_1\|_{nuc} = \|A_2\|_{nuc} = 1$, ya que en caso contrario no alcanzarían la norma de A al ser una combinación convexa.

Ahora usemos la desigualdad de Fan, de lo que resulta que

$$\langle A, A_1\rangle_F \leq \langle \lambda(A), \lambda(A_1)\rangle = 1 \cdot \lambda_1(A_1) + \sum_{i=2}^{d} 0 \cdot \lambda_i(A_1) = \lambda_1(A_1)$$

Pero como $\|A_1\|_{nuc} \leq 1$ entonces $|\lambda_1(A_1)| \leq 1$ y así

$$\langle A, A_1\rangle_F \leq 1$$

De forma análoga se concluye que

$$\langle A, A_2\rangle_F \leq 1$$

Ahora en la ecuación

$$A = \lambda A_1 + (1-\lambda)A_2$$

Tomemos producto interno con A, esto es

$$\langle A, A\rangle_F = \lambda\langle A_1, A\rangle_F + (1-\lambda)\langle A_2, A\rangle_F$$

Pero como $\langle A, A\rangle_F = \|A\|_F^2 = \|uu^T\|_F^2 = \|u\|_2^4 = 1$, y así

$$1 = \lambda\langle A_1, A\rangle_F + (1-\lambda)\langle A_2, A\rangle_F$$

Pero como $\langle A_1, A\rangle_F, \langle A_2, A\rangle_F \leq 1$ la única posibilidad es que $\langle A_1, A\rangle_F = \langle A_2, A\rangle_F = 1$. Así, como se alcanza la igualdad en la desigualdad de Fan, entonces A, A_1, A_2 son simultáneamente diagonalizables. Además de la desigualdad anterior se tiene que

$$1 = \langle A_i, A\rangle \leq \lambda_1(A_i) \leq 1 \quad \textit{para} \quad i = 1, 2$$

Concluyendo que $\lambda_1(A_1) = \lambda_1(A_2) = 1$ Entonces A_1, A_2 son matrices de rango 1 que comparten vectores y valores propios con A, así que son de la forma $\pm uu^T$.

Si ambos fueran uu^T entonces no contradice que A fuera punto extremo

No puede ser que ambos fueran $-uu^T$ puesto que tendríamos $uu^T = A = -uu^T$ y ya dijimos que $A \neq 0$.

Si $A_1 = uu^T$ y $A_2 = -uu^T$ entonces

$$A = \lambda uu^T + (1-\lambda)(-uu^T) = (2\lambda - 1)uu^T$$

Y entonces

$$\|A\|_{nuc} = |2\lambda - 1|$$

Y las únicas soluciones de $|2\lambda - 1| = 1$ *son* $\lambda = 1, \lambda = 0$*, contradiciendo que* $\lambda \in (0, 1)$.

Así, se concluye que si $A = uu^T$ *con* $\|u\|_2 = 1$*, entonces* A *es punto extremo. Como demostramos ambas implicancias, hemos demostrado rigurosamente la caracterización que dimos al inicio, por lo que hemos encontrado los puntos extremos de* $\mathcal{X}$*, tal como se pidió.*

(c) Primero veamos que la norma de Frobenius y la norma Nuclear son invariantes bajo transformaciones ortogonales, esto es, sea $A \in \mathbb{R}^{d\times d}$ *una matriz simétrica y* $Q \in \mathbb{R}^{d\times d}$ *una matriz ortogonal,* $Q^\top Q = I$*. Veamos que*

$$\|Q^T AQ\|_F = \|A\|_F \quad y \quad \|Q^T AQ\|_{nuc} = \|A\|_{nuc}$$

Para la norma de Frobenius se tiene que

$$\begin{aligned}
\|Q^T AQ\|_F^2 &= \mathrm{Tr}\left((Q^T AQ)^T (Q^T AQ)\right) \quad \textit{(caracterización)} \\
&= \mathrm{Tr}\left(Q^T AQQ^T AQ\right) \quad \textit{(Distribuímos el operador Transpusto)} \\
&= \mathrm{Tr}\left(Q^T A^2 Q\right) \quad \textit{(Usamos la ortogonalidad)} \\
&= \mathrm{Tr}\left(A^2 QQ^T\right) \quad \textit{(Ciclicidad de la traza)} \\
&= \mathrm{Tr}(A^2) \quad \textit{(Nuevamente ortogonalidad)} \\
&= \|A\|_F^2
\end{aligned}$$

Demostrando lo que queríamos, ahora veamoslo para la norma Nuclear. Para esto notemos que si Q *es ortogonal, la matriz* $Q^\top AQ$ *tiene los mismos autovalores que* A*. Esto es fácil de ver puesto que si consideramos la descomposición espectral de* A*, esto es* $A = U\Lambda U^T$*, entonces* $Q^T AQ = Q^T U\Lambda U^T Q$*, pero como* U, Q *son ortogonales, entonces* $Q^T U$ *también lo es, y por tanto ambos en la descomposición espectral tienen* Λ *como matriz diagonal. Esto es*

$$\lambda_i(Q^\top AQ) = \lambda_i(A) \quad \forall i$$

Y entonces

$$\|Q^\top AQ\|_{nuc} = \sum_{i=1}^{d} |\lambda_i(Q^\top AQ)| = \sum_{i=1}^{d} |\lambda_i(A)| = \|A\|_{nuc}$$

Demostrando lo que queríamos para la norma nuclear.

Ahora comencemos a ver el algoritmo. Dado que X *es simétrica, posee una descomposición espectral ortogonal*

$$X = U\Lambda U^T$$

donde $U \in \mathbb{R}^{d\times d}$ *es ortogonal y* $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ *es una matriz diagonal con los autovalores de* X*. Formalmente el problema de proyección corresponde a*

$$\min_{Y\in S^d} \|X - Y\|_F^2 \quad s.a \quad \|Y\|_{nuc} \leq 1$$

Reemplazando por la descomposición espectral de X

$$\min_{Y\in S^d} \|U\Lambda U^T - Y\|_F^2 \quad s.a \quad \|Y\|_{nuc} \leq 1$$

Multiplicando ambas normas por U^T *por la izquierda y* U *por la derecha, debido a la invarianza de la norma al multiplicar por una matriz ortogonal resulta que el problema es equivalente a*

$$\min_{Y\in S^d} \|U^T(U\Lambda U^T - Y)U\|_F^2 \quad s.a \quad \|U^T YU\|_{nuc} \leq 1$$

Y entonces

$$\min_{Y \in S^d} \|\Lambda - U^T Y U\|_F^2 \quad \textit{s.a} \quad \|U^T Y U\|_{nuc} \leq 1$$

Llamemos $\tilde{Y} = U^T Y U$*, entonces el problema consiste en*

$$\min_{\tilde{Y} \in S^d} \|\Lambda - \tilde{Y}\|_F^2 \quad \textit{s.a} \quad \|\tilde{Y}\|_{nuc} \leq 1$$

Y de la caracterización de la norma de Frobenius

$$\min_{\tilde{Y} \in S^d} \sum_{i=1}^{d} \sum_{j=1}^{d} (\Lambda_{ij} - \tilde{Y}_{ij})^2 \quad \textit{s.a} \quad \|\tilde{Y}\|_{nuc} \leq 1$$

Pero como Λ *es una matriz diagonal, entonces* $\Lambda_{ij} = 0$ *si* $i \neq j$*, por lo que no conviene dejar la componente* $\tilde{Y}_{ij} \neq 0$ *puesto que aumentaría la función objetivo y la norma nuclear. Así,* $\tilde{Y}$ *también debe ser diagonal. Así, resulta que el problema es equivalente a*

$$\min_{\tilde{Y} \in S^d, \tilde{Y} \textit{ diagonal}} \sum_{i=1}^{d} (\Lambda_{ii} - \tilde{Y}_{ii})^2 \quad \textit{s.a} \quad \sum_{i=1}^{d} \tilde{Y}_{ii} \leq 1$$

Considerando $\mathrm{vec}(X)_i = X_{ii}$ *el vector que toma la diagonal de la matriz, entonces el problema es equivalente a*

$$\min_{\mathrm{vec}(\tilde{Y}) \in \mathbb{R}^d} \|\mathrm{vec}(\Lambda) - \mathrm{vec}(\tilde{Y})\|_2^2 \quad \textit{s.a} \quad \|\mathrm{vec}(\tilde{Y})\|_1 \leq 1$$

Y este problema ya lo sabemos resolver según lo encontrado en (a). Entonces podemos encontrar el $\tilde{Y}^*$ *del problema anterior. Y entonces el óptimo del problema original corresponde a* $U\tilde{Y}^*U^T$*, encontrando así un algoritmo para encontrar el óptimo, junto con la correctitud del algoritmo. Entonces, en resumen el algoritmo es el siguiente*

- *Encontrar la descomposición espectral de* X*. La descomposición espectral se puede realizar en tiempo* $\mathcal{O}(d^3)$*, usando por ejemplo* `np.linalg.eigh` *de NumPy (ver documentación).*
- *Proyectar sobre la bola* ℓ_1 *el vector de valores propios para encontrar* $\tilde{Y}^*$*. Esto toma tiempo* $\mathcal{O}(d \log d)$ *según el algoritmo encontrado en (a)*
- *Reconstruir* $Y^* = U\tilde{Y}^*U^T$*. Toma tiempo* $\mathcal{O}(d^3)$ *ya que hay que hacer dos multiplicaciones matriciales de tamaño* $d \times d$*. La multiplicación de matrices se puede realizar en tiempo* $\mathcal{O}(d^3)$*, usando por ejemplo* `np.matmul` *de NumPy (ver documentación).*

Por tanto, el algoritmo completo tiene complejidad total $\mathcal{O}(d^3)$*. Con esto, encontramos lo solicitado.*

Pregunta 2. *(a) Intuitivamente los puntos extremos debieran ser las matrices que tienen un único valor singular igual a* 1 *y el resto igual a* 0*. Esto es, los puntos extremos son las matrices de la forma*

$$A = \pm uv^T, \quad \textit{con } u \in \mathbb{R}^n, v \in \mathbb{R}^d : \|u\|_2 = \|v\|_2 = 1$$

Ahora veámoslo con más rigor. Primero veamos que si A *no es de esa forma entonces no puede ser punto extremo.*

Si A tiene al menos dos valores singulares distintos a cero entonces por la descomposición espectral A se puede escribir como $A = U\Sigma V^T$ con $U \in \mathbb{R}^{n\times n}, V \in \mathbb{R}^{d\times d}$ ortogonales y $\Sigma \in \mathbb{R}^{n\times d} = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_d)$, y el resto ceros, donde por hipótesis $\sigma_1, \sigma_2 > 0$. Ahora consideremos las matrices

$$A^+ = U\Sigma^+ U^T, \quad A^- = U\Sigma^- U^T$$

con $\Sigma^+ = \mathrm{diag}(\sigma_1 + \epsilon, \sigma_2 - \epsilon, \sigma_3, \ldots, \sigma_d)$ y el resto ceros, y $\Sigma^- = \mathrm{diag}(\sigma_1 - \epsilon, \sigma_2 + \epsilon, \sigma_3, \ldots, \sigma_d)$, y el resto ceros, donde el ϵ es tal que $\sigma_1 \pm \epsilon \in [0, 1]$ y $\sigma_2 \pm \epsilon \in [0, 1]$, el cual existe ya que dichos dos valores singulares son no nulos. Con lo anterior, A se puede escribir como

$$A = \frac{1}{2}A^+ + \frac{1}{2}A^-$$

Y como A^+, A^- son distintas, entonces A no puede ser un punto extremo, de lo que se concluye que si A es punto extremo entonces tiene un único valor singular (demostramos que debe tener 1 o 0, pero este segundo caso se descarta trivialmente ya que la matriz nula se puede escribir fácilmente como combinación de otras dos).

Si dicho único valor singular no es igual a 1, entonces debe ser menor a uno puesto que por definición se tiene que $\|A\|_{nuc} \leq 1$. Pero como es menor estricto a 1, entonces se cumple que

$$0 < \|A\|_{nuc} < 1$$

y entonces existe ϵ tal que

$$0 < \|(1-\epsilon)A\|_{nuc} < \|(1+\epsilon)A\|_{nuc} < 1$$

Y entonces podríamos escribir

$$A = \frac{1}{2}(1-\epsilon)A + \frac{1}{2}(1+\epsilon)A$$

Y entonces no es punto extremo. Sumado a lo anterior, ya demostramos la primera implicancia.

Ahora veamos que si A es de la forma, sin pérdida de generalidad, $A = uv^T$ con $\|u\|_2 = \|v\|_2 = 1$, entonces debe ser punto extremo. Supongamos por contradicción pudieramos escribir A como

$$A = \lambda A_1 + (1-\lambda)A_2$$

con A_1, A_2 en $\mathcal{X}$ y $\lambda \in (0, 1)$. Lo primero es notar que puesto que $\|A\|_{nuc} = 1$, entonces necesariamente $\|A_1\|_{nuc} = \|A_2\|_{nuc} = 1$, ya que en caso contrario no alcanzarían la norma de A al ser una combinación convexa.

Ahora usemos la desigualdad de Fan generalizada, de lo que resulta que

$$\langle A, A_1\rangle_F \leq \sum_{i=1}^{d} \sigma_i(A)\cdot\sigma_i(A_1) = 1\cdot\sigma_1(A_1) + \sum_{i=2}^{d} 0\cdot\sigma_i(A_1) = \sigma_1(A_1)$$

Pero como $\|A_1\|_{nuc} \leq 1$ entonces $|\sigma_1(A_1)| \leq 1$ y así

$$\langle A, A_1\rangle_F \leq 1$$

De forma análoga se concluye que

$$\langle A, A_2\rangle_F \leq 1$$

Ahora en la ecuación

$$A = \lambda A_1 + (1-\lambda)A_2$$

Tomemos producto interno con A, esto es

$$\langle A, A\rangle_F = \lambda\langle A_1, A\rangle_F + (1-\lambda)\langle A_2, A\rangle_F$$

Pero como

$$\langle A, A\rangle_F = \|A\|_F^2 = \sum_{i,j}(A_{ij})^2 = \sum_{i,j}(u_i v_j)^2 = \sum_{i=1}^{n} u_i \sum_{j=1}^{d} v_j = \|u\|_2^2 \cdot \|v\|_2^2 = 1 \cdot 1 = 1$$

Y así

$$1 = \lambda\langle A_1, A\rangle_F + (1-\lambda)\langle A_2, A\rangle_F$$

Pero como $\langle A_1, A\rangle_F, \langle A_2, A\rangle_F \leq 1$ la única posibilidad es que $\langle A_1, A\rangle_F = \langle A_2, A\rangle_F = 1$. Así, como se alcanza la igualdad en la desigualdad de Fan, entonces A, A_1, A_2 son simultáneamente diagonalizables. Además de la desigualdad anterior se tiene que

$$1 = \langle A_i, A\rangle \leq \sigma_1(A_i) \leq 1 \quad \textit{para} \quad i = 1, 2$$

Concluyendo que $\sigma_1(A_1) = \sigma_1(A_2) = 1$ Entonces A_1, A_2 son matrices de rango 1 que comparten la diagonalización y los valores singulares con A, así que son de la forma $\pm uv^T$.

Si ambos fueran uv^T entonces no contradice que A fuera punto extremo

No puede ser que ambos fueran $-uv^T$ puesto que tendríamos $uv^T = A = -uv^T$ y ya dijimos que $A \neq 0$.

Si $A_1 = uv^T$ y $A_2 = -uv^T$ entonces

$$A = \lambda uv^T + (1-\lambda)(-uv^T) = (2\lambda - 1)uv^T$$

Y entonces

$$\|A\|_{nuc} = |2\lambda - 1|$$

Y las únicas soluciones de $|2\lambda - 1| = 1$ son $\lambda = 1, \lambda = 0$, contradiciendo que $\lambda \in (0, 1)$.

Así, se concluye que si $A = uv^T$ con $\|u\|_2 = \|v\|_2 = 1$, entonces A es punto extremo. Como demostramos ambas implicancias, hemos demostrado rigurosamente la caracterización que dimos al inicio, por lo que hemos encontrado los puntos extremos, tal como se pidió.

(b) Primero veamos que la norma de Frobenius y la norma Nuclear son invariantes bajo transformaciones ortogonales, esto es, sea $A \in \mathbb{R}^{n\times d}$ y $U \in \mathbb{R}^{n\times n}, V \in \mathbb{R}^{d\times d}$ matrices ortogonales. Veamos que

$$\|U^T AV\|_F = \|A\|_F \quad y \quad \|U^T AV\|_{nuc} = \|A\|_{nuc}$$

Para la norma de Frobenius se tiene que

$$\begin{aligned}
\|U^T AV\|_F^2 &= \operatorname{Tr}\left((U^T AV)^T (U^T AV)\right) \quad \textit{(caracterización)} \\
&= \operatorname{Tr}\left(V^T AUU^T AV\right) \quad \textit{(Distribuímos el operador Transpusto)} \\
&= \operatorname{Tr}\left(V^T A^2 V\right) \quad \textit{(Usamos la ortogonalidad)} \\
&= \operatorname{Tr}\left(A^2 VV^T\right) \quad \textit{(Ciclicidad de la traza)} \\
&= \operatorname{Tr}(A^2) \quad \textit{(Nuevamente ortogonalidad)} \\
&= \|A\|_F^2
\end{aligned}$$

Demostrando lo que queríamos, ahora veamoslo para la norma Nuclear. Para esto notemos que si U, V son ortogonales, la matriz $U^T AV$ tiene los mismos valores singulares que A. Esto es fácil de ver puesto que si consideramos la descomposición espectral de A, esto es $A = U'\Lambda V'^T$, entonces $U^T AV = U^T U'\Lambda V'^T V$, pero como U, V, U', V' son ortogonales, entonces $U^T U', V'^T V$ también lo son, y por tanto ambos en la descomposición espectral tienen Σ como matriz diagonal. Esto es

$$\sigma_i(U^T AV) = \sigma_i(A) \quad \forall i$$

Y entonces

$$\|U^T AV\|_{nuc} = \sum_{i=1}^{d} |\sigma_i(U^T AV)| = \sum_{i=1}^{d} |\sigma_i(A)| = \|A\|_{nuc}$$

Demostrando lo que queríamos para la norma nuclear.

Ahora comencemos a ver el algoritmo. Dado que X es simétrica, posee una descomposición espectral de la forma

$$X = U\Sigma V^T$$

donde $U \in \mathbb{R}^{n\times n}, V \in \mathbb{R}^{d\times d}$ son ortogonales y $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_d)$ es una matriz diagonal con los valores singulares de X (y el resto 0). Formalmente el problema de proyección corresponde a

$$\min_{Y\in\mathbb{R}^{n\times d}} \|X - Y\|_F^2 \quad \textit{s.a} \quad \|Y\|_{nuc} \leq 1$$

Reemplazando por la descomposición espectral de X

$$\min_{Y\in\mathbb{R}^{n\times d}} \|U\Sigma V^T - Y\|_F^2 \quad \textit{s.a} \quad \|Y\|_{nuc} \leq 1$$

Multiplicando ambas normas por U^T por la izquierda y V por la derecha, debido a la invarianza de la norma al multiplicar por una matriz ortogonal resulta que el problema es equivalente a

$$\min_{Y\in\mathbb{R}^{n\times d}} \|U^T(U\Sigma V^T - Y)V\|_F^2 \quad \textit{s.a} \quad \|U^T YV\|_{nuc} \leq 1$$

Y entonces

$$\min_{Y\in\mathbb{R}^{n\times d}} \|\Sigma - U^T YV\|_F^2 \quad \textit{s.a} \quad \|U^T YV\|_{nuc} \leq 1$$

Llamemos $\tilde{Y} = U^T YV$, entonces el problema consiste en

$$\min_{\tilde{Y}\in\mathbb{R}^{n\times d}} \|\Sigma - \tilde{Y}\|_F^2 \quad \textit{s.a} \quad \|\tilde{Y}\|_{nuc} \leq 1$$

Y de la caracterización de la norma de Frobenius

$$\min_{\tilde{Y}\in\mathbb{R}^{n\times d}} \sum_{i=1}^{n}\sum_{j=1}^{d} (\Sigma_{ij} - \tilde{Y}_{ij})^2 \quad \textit{s.a} \quad \|\tilde{Y}\|_{nuc} \leq 1$$

Pero como Σ es una matriz diagonal, entonces $\Sigma_{ij} = 0$ si $i \neq j$, por lo que no conviene dejar la componente $\tilde{Y}_{ij} \neq 0$ puesto que aumentaría la función objetivo y la norma nuclear. Así, $\tilde{Y}$ también debe ser diagonal (entendiéndose que usa la diagonal principal y el resto son nulos). Así, resulta que el problema es equivalente a

$$\min_{\tilde{Y}\in\mathbb{R}^{n\times d}, \tilde{Y}\ \textit{diagonal}} \sum_{i=1}^{d} (\Sigma_{ii} - \tilde{Y}_{ii})^2 \quad \textit{s.a} \quad \sum_{i=1}^{d} \tilde{Y}_{ii} \leq 1$$

Considerando $\text{vec}(X)_i = X_{ii}$ *el vector que toma la diagonal de la matriz, entonces el problema es equivalente a*

$$\min_{\text{vec}(\tilde{Y})\in\mathbb{R}^d} \|\text{vec}(\Sigma) - \text{vec}(\tilde{Y})\|_2^2 \quad \textit{s.a} \quad \|\text{vec}(\tilde{Y})\|_1 \leq 1$$

Y este problema ya lo sabemos resolver según lo encontrado en (a). Entonces podemos encontrar el $\tilde{Y}^*$ *del problema anterior. Y entonces el óptimo del problema original corresponde a* $U\tilde{Y}^*V^T$*, encontrando así un algoritmo para encontrar el óptimo, junto con la correctitud del algoritmo. Entonces, en resumen el algoritmo es el siguiente.*

Primero recordemos que podemos encontrar descomposición espectral en $\mathcal{O}(d^3)$ *y multiplicar matrices de* $n \times d$ *con* $d \leq n$ *en* $O(d^2n)$ *según los recursos proporcionados en la pregunta 1c. Por lo tanto el algoritmo consiste en*

- *Encontrar la descomposición en valores singulares de* X*. Esta descomposición se puede calcular en tiempo* $\mathcal{O}(nd^2)$*, utilizando por ejemplo la función* `np.linalg.svd` *de NumPy con el parámetro* `full_matrices=False` *(ver documentación).*
- *Proyectar sobre la bola* ℓ_1 *el vector de valores singulares para encontrar* $\tilde{Y}^*$*. Esto toma tiempo* $\mathcal{O}(d \log d)$ *según el algoritmo encontrado en (a)*
- *Reconstruir* $Y^* = U\tilde{Y}^*V^T$ *con dos multiplicaciones matriciales de tamaño* $n \times d$ *y* $d \times n$*. La multiplicación de matrices se puede realizar en tiempo* $\mathcal{O}(n^2d)$ *y* $\mathcal{O}(d^2n)$*, usando por ejemplo* `np.matmul` *de NumPy (ver documentación).*

Por tanto, el algoritmo completo tiene complejidad total $\mathcal{O}(n^2d)$*. Con esto, encontramos lo solicitado.*

Pregunta 3. *(a) Reescribamos la función* f *como el producto punto usual (eso me dijeron en el foro) para más claridad.*

$$f(x) = \frac{1}{2}[x^TA^TAx - b^Tx]$$

Y entonces su gradiente viene dado por

$$\nabla f(x) = A^TAx - \frac{1}{2}b$$

Vamos a querer usar la definición de C_f*, por lo que calculemos lo necesario. Tenemos que* $y = (1-\gamma)x + \gamma z = x + \gamma(z-x)$*. Así, resulta que*

$$f(y) = \frac{1}{2}(x+\gamma(z-x))^TA^TA(x+\gamma(z-x)) - \frac{1}{2}b^T(x+\gamma(z-x))$$

Expandiendo

$$f(y) = \frac{1}{2}x^TA^TAx + \gamma x^TA^TA(z-x) + \frac{\gamma^2}{2}(z-x)^TA^TA(z-x) - \frac{1}{2}b^Tx - \frac{\gamma}{2}b^T(z-x)$$

Además

$$f(x) = \frac{1}{2}x^TA^TAx - \frac{1}{2}b^Tx$$

Y podemos calcular entonces que

$$f(y) - f(x) = \gamma x^TA^TA(z-x) + \frac{\gamma^2}{2}(z-x)^TA^TA(z-x) - \frac{\gamma}{2}b^T(z-x)$$

Ahora notemos que $y - x = x + \gamma(z - x) - x = \gamma(z - x)$, y entonces $\langle \nabla f(x), y - x \rangle = \gamma \langle \nabla f(x), z - x \rangle$. Calculemos esto último, se tiene directamente que

$$\gamma \langle \nabla f(x), z - x \rangle = \gamma (x^T A^T A(z - x) - \frac{1}{2} b^T (z - x))$$

Y entonces

$$f(y) - f(x) - \gamma \langle \nabla f(x), z - x \rangle = \frac{\gamma^2}{2} (z - x)^T A^T A(z - x)$$

Por definición la constante de curvatura resulta en

$$C_f = \sup_{x,z \in X, \gamma \in [0,1]} \frac{\gamma^2}{2} (z - x)^T A^T A(z - x) = \frac{1}{2} \sup_{x,z \in X, \gamma \in [0,1]} \gamma^2 \|A(z - x)\|_2^2$$

Pero el máximo se alcanza cuando $\gamma = 1$, puesto que el interior de la norma no depende de γ, y entonces se tiene que

$$C_f = \frac{1}{2} \sup_{x,z \in X} \|A(z - x)\|_2^2$$

Calculando lo pedido.

(b) Como f es convexa y L_1-suave, entonces de las caracterizaciones vistas en clases se tiene que

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_1}{2} \|y - x\|^2$$

Sea $y = (1 - \gamma)x + \gamma z = x + \gamma(z - x)$ con $x, y \in \mathcal{X}$ arbitrarios y $\gamma \in [0, 1]$. Entonces resulta que

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leq \frac{L_1}{2} \|y - x\|^2 = \frac{L_1}{2} \gamma^2 \|z - x\|^2$$

Multiplicando ambos lados por $\frac{2}{\gamma^2}$ se obtiene

$$\frac{2}{\gamma^2} \left[f(y) - f(x) - \langle \nabla f(x), y - x \rangle \right] \leq L_1 \|z - x\|^2$$

Tomando supremo sobre $x, z \in X$ y $\gamma \in (0, 1]$, concluimos que

$$C_f \leq \sup_{x,z \in \mathcal{X}} L_1 \|z - x\|^2 = L_1 \cdot \left(\sup_{x,z \in \mathcal{X}} \|x - z\| \right)^2 = L_1 \cdot \operatorname{diam}_{\|\cdot\|}(X)^2$$

Tal como queríamos demostrar

(c) De la definición de C_f, al ser un supremo, es directo que

$$\frac{2}{\gamma^2} [f(y) - f(x) - \langle \nabla f(x), y - x \rangle] \leq C_f \quad \forall x, z \in \mathcal{X}, \gamma \in [0, 1], y = (1 - \gamma)x + \gamma z$$

Apliquemos el resultado anterior al caso particular $x := x_k$, $z := v_k$ y $\gamma := \eta_k$. Notemos entonces por definición que $y := (1 - \eta_k)x_k + \eta_k v_k = x_{k+1}$. Y entonces, reemplazando

$$\frac{2}{\eta_k^2} [f(x_{k+1}) - f(x_k) - \langle \nabla f(x_k), x_{k+1} - x_k \rangle] \leq C_f$$

Y reorganizando

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{\eta_k^2}{2} C_f$$

Finalmente, notando que $x_{k+1} - x_k = (1 - \eta_k)x_k + \eta_k v_k - x_k = \eta_k(v_k - x_k)$ *resulta que*

$$f(x_{k+1}) \leq f(x_k) + \eta_k \langle \nabla f(x_k), v_k - x_k \rangle + \frac{\eta_k^2 C_f}{2}$$

Demostrando así la desigualdad pedida.

(d) Por definición de v_k *se tiene que*

$$\langle \nabla f(x_k), v_k \rangle \leq \langle \nabla f(x_k), v \rangle \quad \forall v \in \mathcal{X}$$

Así, en particular se tiene que

$$\langle \nabla f(x_k), v_k \rangle \leq \langle \nabla f(x_k), x^* \rangle$$

Y de la linealidad del producto interno, restando $\langle \nabla f(x_k), x_k \rangle$ *a ambos lados resulta que*

$$\langle \nabla f(x_k), v_k - x_k \rangle \leq \langle \nabla f(x_k), x^* - x_k \rangle$$

Por ser f *convexa se tiene además que*

$$f(x^*) \geq f(x_k) + \langle \nabla f(x_k), x^* - x_k \rangle \Rightarrow \langle \nabla f(x_k), x^* - x_k \rangle \leq f(x^*) - f(x_k) = -\gamma_k$$

Mezclando ambas desigualdades resulta que

$$\langle \nabla f(x_k), v_k - x_k \rangle \leq \langle \nabla f(x_k), x^* - x_k \rangle \leq -\gamma_k$$

Y usando esta desigualdad en lo obtenido en (c) se tiene que

$$f(x_{k+1}) \leq f(x_k) - \eta_k \gamma_k + \frac{\eta_k^2 C_f}{2}$$

Restando $f(x^*)$ *a ambos lados*

$$f(x_{k+1}) - f(x^*) \leq f(x_k) - f(x^*) - \eta_k \gamma_k + \frac{\eta_k^2 C_f}{2}$$

Esto es por definición

$$\gamma_{k+1} \leq \gamma_k - \eta_k \gamma_k + \frac{\eta_k^2 C_f}{2}$$

Factorizando

$$\gamma_{k+1} \leq (1 - \eta_k)\gamma_k + \frac{\eta_k^2 C_f}{2}$$

Y reemplazando por el valor d η_k *dado en el algoritmo resulta que*

$$\gamma_{k+1} \leq \left(1 - \frac{2}{k+2}\right)\gamma_k + \frac{2C_f}{(k+2)^2}$$

Tal como queríamos demostrar.

(e) Demostrémoslo por inducción. Comencemos por probar el caso base. De la definición de C_f, tomando $x := x^*$, $z := x_0$, $\gamma := 1$, y entonces $y := x^*$, se obtiene que

$$2[f(x_0) - f(x^*) - \langle \nabla f(x^*), x_0 - x^* \rangle] \leq C_f$$

Reordenando y sumando a que $C_f \geq 0$ (lo cual resulta de tomar $x = z$ en la definición) se obtiene

$$f(x_0) - f(x^*) - \langle \nabla f(x^*), x_0 - x^* \rangle \leq \frac{1}{2} C_f \leq C_f$$

Pero por la condición de Fermat $\nabla f(x^*) = 0$, y entonces

$$\gamma_0 \leq f(x_0) - f(x^*) \leq C_f = \frac{2C_f}{0+2}$$

Lo que demuestra el caso base. Como hipótesis de inducción supongamos que la cota se cumple para k, esto es, que

$$\gamma_k \leq \frac{2C_f}{k+2}$$

Y ahora queremos demostrarlo para $k+1$, comencemos desde lo encontrado en (d), esto es

$$\gamma_{k+1} \leq \left(1 - \frac{2}{k+2}\right) \gamma_k + \frac{2C_f}{(k+2)^2}$$

Y usando la hipótesis inductiva resulta que

$$\gamma_{k+1} \leq \left(1 - \frac{2}{k+2}\right) \cdot \frac{2C_f}{k+2} + \frac{2C_f}{(k+2)^2} = \frac{k}{k+2} \cdot \frac{2C_f}{k+2} + \frac{2C_f}{(k+2)^2}$$

Y factorizando resulta que

$$\gamma_{k+1} \leq \frac{2C_f \cdot (k+1)}{(k+2)^2}$$

Y para ver que esto cumple la cota pedida comencemos desde que

$$k^2 + 4k + 3 \leq k^2 + 4k + 4$$

Factorizando

$$(k+1)(k+3) \leq (k+2)^2$$

Y entonces

$$\frac{k+1}{(k+2)^2} \leq \frac{1}{(k+3)}$$

Y usando esta desigualdad en el resultado anterior, se tiene que

$$\gamma_{k+1} \leq \frac{2C_f}{k+3}$$

Demostrando así la tesis de inducción, y entonces concluyendo la cota anterior para todo k, y en particular se tiene que

$$\gamma_K = f(x_K) - \min_{x \in \mathcal{X}} f(x) \leq \frac{2C_f}{K+2}$$

Tal como queríamos demostrar