Pregunta 1

(a) Tenemos que

$$\Lambda_{X^2}(\lambda) = \ln \mathbb{E}\left[e^{\lambda X^2}\right] = \ln \int_{-\infty}^{\infty} e^{\lambda x^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \ln \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}+\lambda x^2} dx$$

Notemos que si $\lambda \geq \frac{1}{2}$ entonces la integral no converge, por lo que es necesaria la suposición de que $\lambda < \frac{1}{2}$ Reescribiendo

$$\Lambda_{X^2}(\lambda) = \ln \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x^2-2\lambda x^2)}{2}} dx = \ln \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x\sqrt{1-2\lambda})^2}{2}} dx$$

Consideremos la sustitución $u = x\sqrt{1-2\lambda}$, entonces $du = \sqrt{1-2\lambda}dx$, convirtiendo la integral en

$$\Lambda_{X^2}(\lambda) = \ln \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \cdot \frac{1}{\sqrt{1-2\lambda}} du = \ln \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \cdot \frac{1}{\sqrt{1-2\lambda}} du$$

Podemos sacar de la integral el $\frac{1}{\sqrt{1-2\lambda}}$, y lo que queda es la integral de la función de densidad de una distribución normal estándar sobre todo su soporte, por lo que el resultado debe ser 1. De esto se concluye que

$$\Lambda_{X^2}(\lambda) = \ln \frac{1}{\sqrt{1-2\lambda}} = \ln(1-2\lambda)^{-\frac{1}{2}} = -\frac{1}{2}\ln(1-2\lambda)$$

Tal como queríamos demostrar. Notemos nuevamente que para $\lambda \geq \frac{1}{2}$ el logaritmo se nos indefine, por lo que λ debe ser menor que $\frac{1}{2}$

(b) Calculemos $\Lambda_Z(\lambda)$

$$\Lambda_Z(\lambda) = \ln \mathbb{E}\left[e^{\lambda Z}\right] = \ln \mathbb{E}\left[e^{\lambda \sum_{i=1}^k X_i^2}\right] = \ln \prod_{i=1}^k \mathbb{E}\left[e^{\lambda X_i^2}\right] = \sum_{i=1}^k \ln \mathbb{E}\left[e^{\lambda X_i^2}\right] = -\frac{k}{2}\ln(1-2\lambda)$$

Además tenemos que

$$\mu = \mathbb{E}[Z] = \sum_{i=1}^k \mathbb{E}[X_i^2] = \sum_{i=1}^k (Var(X_i) + \mathbb{E}[X_i]^2) = \sum_{i=1}^k (1+0^2) = k$$

Con esto podemos usar el teorema de los grandes desvíos. Pese a que esta técnica conocida como cotas de Chernoff se puede usar para cualquier distribución, pero para ser consistente con la formulación que vimos para el teorema consideremos $\bar{Z} = \frac{Z}{k}$, y por tanto tiene media 1 y $\Lambda_{\bar{Z}}(\lambda) = -\frac{1}{2}\ln(1-2\lambda)$

$$\frac{1}{k}\ln P(\bar{Z} \geq \mu + \epsilon) \leq \inf_{\lambda>0}\left[-\lambda(\mu+\epsilon) - \frac{1}{2}\ln(1-2\lambda)\right]$$

Reemplazemos $\epsilon = \frac{2\sqrt{tk}+2t}{k}$ y por el valor de μ obtenemos

$$\frac{1}{k}\ln P\left(\bar{Z} \geq 1 + \frac{2\sqrt{tk}+2t}{k}\right) \leq \inf_{\lambda>0}\left[-\lambda\left(1+\frac{2\sqrt{tk}+2t}{k}\right) - \frac{1}{2}\ln(1-2\lambda)\right]$$

Multiplicando por k en ambos lados de la ecuación y considerando que $\bar{Z} = \frac{Z}{k}$ y multiplicando por k en ambos lados dentro de la probabilidad, obtenemos que

$$\ln P\left(Z \geq k + 2\sqrt{tk} + 2tk\right) \leq \inf_{\lambda>0}\left[-\lambda\left(k+2\sqrt{tk}+2t\right) - \frac{k}{2}\ln(1-2\lambda)\right]$$

Consideremos $\lambda = \frac{t+\sqrt{tk}}{k+2\sqrt{tk}+2t}$ (basta derivar e igualar a cero para encontrar este candidato). Como $t > 0$ y $k > 0$ entonces se verifica que $0 < \lambda < \frac{1}{2}$ por lo que es válido para $\Lambda_Z(\lambda)$, reemplazando

$$\ln P\left(Z \geq k + 2\sqrt{tk} + 2tk\right) \leq -\frac{t+\sqrt{tk}}{k+2\sqrt{tk}+2t}\left(k+2\sqrt{tk}+2t\right) - \frac{k}{2}\ln\left(1 - 2\frac{t+\sqrt{tk}}{k+2\sqrt{tk}+2t}\right)$$

Operamos y simplificamos

$$\ln P\left(Z \geq k + 2\sqrt{tk} + 2tk\right) \leq -t - \sqrt{tk} - \frac{k}{2}\ln\left(\frac{k}{k+2\sqrt{tk}+2t}\right)$$

Consideremos $f(t) = \sqrt{tk} + \frac{k}{2}\ln\left(\frac{k}{k+2\sqrt{tk}+2t}\right)$, notemos que $f(0) = 0$, y a medida que t crece el término $\sqrt{tk}$ crece más rápidamente que lo que decrece $\frac{k}{2}\ln\left(\frac{k}{k+2\sqrt{tk}+2t}\right)$ (por estar en un logaritmo), por lo que podemos concluir que $f(t) \geq 0$ cuando $t > 0$, esto es, $-f(t) \leq 0$ cuando $t > 0$. Incluyendo esta cota en la ecuación anterior

$$\ln P\left(Z \geq k + 2\sqrt{tk} + 2tk\right) \leq -t - \sqrt{tk} - \frac{k}{2}\ln\left(\frac{k}{k+2\sqrt{tk}+2t}\right) \leq -t - 0 = -t$$

Exponenciando a ambos lados

$$P\left(Z \geq k + 2\sqrt{tk} + 2tk\right) \leq e^{-t}$$

Probando así la primera desigualdad pedida. Para la segunda podemos realizar un procedimiento análogo. Podemos usar el teorema de los grandes desvíos.

$$\frac{1}{k}\ln P(\bar{Z} \leq \mu - \epsilon) \leq \inf_{\lambda<0}\left[-\lambda(\mu-\epsilon) - \frac{1}{2}\ln(1-2\lambda)\right]$$

Reemplazemos $\epsilon = \frac{2\sqrt{tk}}{k}$ y por el valor de μ obtenemos

$$\frac{1}{k}\ln P\left(\bar{Z} \leq 1 - \frac{2\sqrt{tk}}{k}\right) \leq \inf_{\lambda<0}\left[-\lambda\left(1-\frac{2\sqrt{tk}}{k}\right) - \frac{1}{2}\ln(1-2\lambda)\right]$$

Multiplicando por k en ambos lados de la ecuación y considerando que $\bar{Z} = \frac{Z}{k}$ y multiplicando por k en ambos lados dentro de la probabilidad, obtenemos que

$$\ln P\left(Z \leq k - 2\sqrt{tk}\right) \leq \inf_{\lambda<0}\left[-\lambda\left(k-2\sqrt{tk}\right) - \frac{k}{2}\ln(1-2\lambda)\right]$$

Pero el mínimo se alcanza cuando $\lambda > 0$ y considerando $\lambda = 0$ aún así no es suficiente. Realmente lo llevo intentando demasiado tiempo así que ya me rendí, si pudiera probar que las cotas son simétricas estamos listos pero según yo no lo son.

(c) Como en el enunciado dice que podemos cambiar las constantes, yo voy a considerar $\epsilon = 4\sqrt{\frac{\ln(2/\delta)}{k}}$. Para los resultados de la parte 1b) consideremos $t = \ln\left(\frac{2}{\delta}\right) > 0$. Para la primera cota resulta que

$$P\left(Z \geq k + 2\sqrt{k\ln\left(\frac{2}{\delta}\right)} + 2\ln\left(\frac{2}{\delta}\right)\right) \leq e^{-\ln\left(\frac{2}{\delta}\right)} = e^{\ln\left(\frac{\delta}{2}\right)} = \frac{\delta}{2}$$

Ahora veamos que

$$(1+\epsilon)k = \left(1 + 4\sqrt{\frac{\ln(2/\delta)}{k}}\right)k = k + 4\sqrt{k\ln(2/\delta)} = k + 2\sqrt{k\ln(2/\delta)} + 2\sqrt{k\ln(2/\delta)}$$

Pero como $k > \ln(2/\delta)$, se concluye que

$$(1+\epsilon)k > k + 2\sqrt{k\ln\left(\frac{2}{\delta}\right)} + 2\ln\left(\frac{2}{\delta}\right)$$

Por lo tanto, podemos concluir que

$$P(Z \geq (1+\epsilon)k) \leq P\left(Z \geq k + 2\sqrt{k\ln\left(\frac{2}{\delta}\right)} + 2\ln\left(\frac{2}{\delta}\right)\right) \leq \frac{\delta}{2}$$

Para la segunda cota resulta

$$P\left(Z \leq k - 2\sqrt{k \ln\left(\frac{2}{\delta}\right)}\right) \leq e^{-\ln\left(\frac{2}{\delta}\right)} = e^{\ln\left(\frac{\delta}{2}\right)} = \frac{\delta}{2}$$

Ahora veamos que

$$(1-\epsilon)k = \left(1 - 4\sqrt{\frac{\ln(2/\delta)}{k}}\right)k = k - 4\sqrt{k\ln(2/\delta)} \leq k - 2\sqrt{k\ln(2/\delta)}$$

Por lo que

$$P(Z \leq (1-\epsilon)k) \leq P\left(Z \leq k - 2\sqrt{k\ln\left(\frac{2}{\delta}\right)}\right) \leq \frac{\delta}{2}$$

Pero entonces

$$P\left(Z \leq (1-\epsilon)k \text{ ó } Z \geq (1+\epsilon)k\right) \leq P(Z \leq (1-\epsilon)k) + P(Z \geq (1+\epsilon)k) \leq \frac{\delta}{2} + \frac{\delta}{2} = \delta$$

Y así podemos concluir que

$$P\left((1-\epsilon)k < Z < (1+\epsilon)k\right) = 1 - P\left(Z \leq (1-\epsilon)k \text{ ó } Z \geq (1+\epsilon)k\right) \geq 1 - \delta$$

Tal como queríamos demostrar.

Pregunta 2

(a) Primero veamos que es Ax.

$$Ax = \begin{bmatrix} a_{11} & \dots & a_{1d} \\ \vdots & \ddots & \vdots \\ a_{k1} & \dots & a_{kd} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{d} x_i \cdot a_{1i} \\ \vdots \\ \sum_{i=1}^{d} x_i \cdot a_{ki} \end{bmatrix} \tag{1}$$

Pero como cada $a_{ij} \sim \mathcal{N}(\mu = 0, \sigma^2 = 1)$, entonces para una entrada del vector x arbitraria se tiene que $x_k \cdot a_{ij} \sim \mathcal{N}(\mu = 0, \sigma^2 = x_k^2)$. Llamemos $X_k \sim \mathcal{N}(\mu = 0, \sigma^2 = x_k^2)$, entonces tenemos que

$$Ax = \begin{bmatrix} \sum_{i=1}^{d} X_i \sim \mathcal{N}(\mu = 0, \sigma^2 = x_i^2) \\ \vdots \\ \sum_{i=1}^{d} X_i \sim \mathcal{N}(\mu = 0, \sigma^2 = x_i^2) \end{bmatrix} = \begin{bmatrix} \bar{X} \sim \mathcal{N}(\mu = 0, \sigma^2 = \sum_{i=1}^{k} x_i^2) \\ \vdots \\ \bar{X} \sim \mathcal{N}(\mu = 0, \sigma^2 = \sum_{i=1}^{k} x_i^2) \end{bmatrix} \tag{2}$$

Pero recordemos que $\|x\|_2 = \sqrt{\sum_{i=1}^k x_i^2}$, por lo que dividiendo en este escalar obtenemos que

$$\frac{Ax}{\|x\|_2} = \begin{bmatrix} X \sim \mathcal{N}(\mu = 0, \sigma^2 = 1) \\ \vdots \\ \bar{X} \sim \mathcal{N}(\mu = 0, \sigma^2 = 1) \end{bmatrix} \tag{3}$$

Y entonces

$$\frac{\|Ax\|_2^2}{\|x\|_2^2} = \frac{\|Ax\|_2}{\|x\|_2} \cdot \frac{\|Ax\|_2}{\|x\|_2} = \left\| \frac{Ax}{\|x\|_2} \right\|_2^2 = \sum_{i=1}^k X_i^2 = Z \sim \chi_k^2$$

De lo que concluímos que $\frac{\|Ax\|_2^2}{\|x\|_2^2} \sim \chi_k^2$, por lo que podemos aplicar el resultado de la pregunta 1c, obteniendo que

$$P\left((1-\epsilon)k \leq \frac{\|Ax\|_2^2}{\|x\|_2^2} \leq (1+\epsilon)k\right) = P\left((1-\epsilon)k < \frac{\|Ax\|_2^2}{\|x\|_2^2} < (1+\epsilon)k\right) \geq 1 - \delta$$

Multiplicando por $\frac{\|x\|_2^2}{k} > 0$ en todos los lados de la desigualdad dentro de la probabilidad se obtiene

$$P\left((1-\epsilon)\|x\|_2^2 \leq \frac{1}{k}\|Ax\|_2^2 \leq (1+\epsilon)\|x\|_2^2\right) \geq 1 - \delta$$

Que es exactamente lo que queríamos demostrar

(b) Consideremos A como la matriz aleatoria descrita en el punto anterior. Sean $x, y \in Q$ arbitrarios. Por lo demostrado en el punto anterior y considerando $\delta = \frac{2}{e^{\frac{k\epsilon^2}{9}}}$ (si $\delta \geq 1$ también se cumple lo anterior trivialmente) tenemos que

$$P\left((1-\epsilon)\|x-y\|_2^2 \leq \frac{1}{k}\|A(x-y)\|_2^2 \leq (1+\epsilon)\|x-y\|_2^2\right) \geq 1 - \delta$$

Distribuyendo en el término central obtenemos que

$$P\left((1-\epsilon)\|x-y\|_2^2 \leq \frac{1}{k}\|Ax-Ay\|_2^2 \leq (1+\epsilon)\|x-y\|_2^2\right) \geq 1 - \delta$$

Notemos además que $\|x-y\|_2^2 = \|y-x\|_2^2$ y que $\|Ax-Ay\|_2^2 = \|Ay-Ax\|_2^2$. Ahora estudiemos la probabilidad de que para todo par $x, y \in Q$ cumpla con la desigualdad (notemos que hay $|Q|^2$ de estos pares), esto es

$$P\left((1-\epsilon)\|x-y\|_2^2 \leq \frac{1}{k}\|Ax-Ay\|_2^2 \leq (1+\epsilon)\|x-y\|_2^2 \ \ \forall x, y \in Q\right)$$

Por lo primero, la probabilidad de que un par dado falle la propiedad es menor o igual a δ (puesto que la probabilidad de cumplir la propiedad es mayor o igual a $1-\delta$). El evento de que exista un par que falle es una unión sobre los $|Q|^2$ eventos dados por pares posibles de fallar, por lo que la probabilidad de que exista un par que falle está acotada por la suma de la probabilidad de cada par de fallar, pero dicha probabilidad está acotada por δ, y como hay $|Q|^2$ pares posibles dicha suma resulta en $|Q|^2\delta$. La probabilidad que ningun par falle es el complemento, y por tanto es mayor o igual a $1-|Q|^2\delta$, esto es

$$P\left((1-\epsilon)\|x-y\|_2^2 \leq \frac{1}{k}\|Ax-Ay\|_2^2 \leq (1+\epsilon)\|x-y\|_2^2 \ \ \forall x,y\in Q\right) \geq 1-|Q|^2\delta$$

Reemplazando por el valor de δ obtenemos que

$$P\left((1-\epsilon)\|x-y\|_2^2 \leq \frac{1}{k}\|Ax-Ay\|_2^2 \leq (1+\epsilon)\|x-y\|_2^2 \ \ \forall x,y\in Q\right) \geq 1-\frac{2|Q|^2}{e^{\frac{k\epsilon^2}{9}}}$$

Busquemos un valor de k tal que $1-\frac{2|Q|^2}{e^{\frac{k\epsilon^2}{9}}} > 0$

$$1-\frac{2|Q|^2}{e^{\frac{k\epsilon^2}{9}}} > 0 \to 1 > \frac{2|Q|^2}{e^{\frac{k\epsilon^2}{9}}} \to e^{\frac{k\epsilon^2}{9}} > 2|Q|^2 \to \frac{k\epsilon^2}{9} > \ln\left(2|Q|^2\right) \to k > \frac{9\ln\left(2|Q|^2\right)}{\epsilon^2}$$

Y considerando función cielo en vez de función suelo en el enunciado es claro que para la hipótesis del k se cumple. Por lo tanto tenemos que

$$P\left((1-\epsilon)\|x-y\|_2^2 \leq \frac{1}{k}\|Ax-Ay\|_2^2 \leq (1+\epsilon)\|x-y\|_2^2 \ \ \forall x,y\in Q\right) > 0$$

Como la probabilidad es mayor que 0, entonces debe existir al menos un resultado de las variables aleatorias de la matriz aleatoria tal que se cumple la propiedad, llamemos a dicho resultado de matriz $\bar{A}$. Por lo tanto, para dicha matriz $\bar{A}$ se cumple que

$$(1-\epsilon)\|x-y\|_2^2 \leq \frac{1}{k}\|\bar{A}x-\bar{A}y\|_2^2 \leq (1+\epsilon)\|x-y\|_2^2 \ \ \forall x,y\in Q$$

Metiendo el $\frac{1}{k}$ dentro de la norma obtenemos que

$$(1-\epsilon)\|x-y\|_2^2 \leq \left\|\frac{1}{\sqrt{k}}\bar{A}x-\frac{1}{\sqrt{k}}\bar{A}y\right\|_2^2 \leq (1+\epsilon)\|x-y\|_2^2 \ \ \forall x,y\in Q$$

Por lo tanto la matriz $\frac{1}{\sqrt{k}}\bar{A}$ cumple la propiedad descrita en el enunciado, demostrando la existencia de dicha matriz.

(c) Como A es una inmersión de baja distorsión sobre $\mathcal{Q}$ para un $\epsilon > 0$, entonces considerando la aplicación sobre x e y tenemos que

$$(1-\epsilon)\|x-y\|_2^2 \leq \|Ax-Ay\|_2^2 \leq (1+\epsilon)\|x-y\|_2^2$$

Y considerando la aplicación sobre x e $-y$ tenemos que

$$(1-\epsilon)\|x+y\|_2^2 \leq \|Ax+Ay\|_2^2 \leq (1+\epsilon)\|x+y\|_2^2$$

Restando ambas desigualdades obtenemos que

$$(1-\epsilon)\|x-y\|_2^2-(1-\epsilon)\|x+y\|_2^2 \leq \|Ax-Ay\|_2^2-\|Ax+Ay\|_2^2 \leq (1+\epsilon)\|x-y\|_2^2-(1+\epsilon)\|x+y\|_2^2$$

De la parte izquierda de la desigualdad tenemos que

$$(1-\epsilon)\|x-y\|_2^2-(1-\epsilon)\|x+y\|_2^2 = \|x-y\|_2^2-\|x+y\|_2^2-\epsilon\|x-y\|_2^2+\epsilon\|x+y\|_2^2$$

Pero como $x, y \in \mathcal{B}_2^d$, entonces $\|x-y\|_2^2 \leq 2$ y $\|x+y\|_2^2 \leq 2$, entonces

$$(1-\epsilon)\|x-y\|_2^2-(1-\epsilon)\|x+y\|_2^2 \geq \|x-y\|_2^2-\|x+y\|_2^2-2\epsilon$$

De la parte derecha de la desigualdad tenemos que

$$(1+\epsilon)\|x-y\|_2^2-(1+\epsilon)\|x+y\|_2^2 = \|x-y\|_2^2-\|x+y\|_2^2+\epsilon\|x-y\|_2^2-\epsilon\|x+y\|_2^2$$

Pero como $x, y \in \mathcal{B}_2^d$, entonces $\|x-y\|_2^2 \leq 2$ y $\|x+y\|_2^2 \leq 2$, entonces

$$(1+\epsilon)\|x-y\|_2^2-(1+\epsilon)\|x+y\|_2^2 \leq \|x-y\|_2^2-\|x+y\|_2^2+2\epsilon$$

De la desigualdad inicial y el desglose de las dos partes podemos concluir que

$$\|x-y\|_2^2-\|x+y\|_2^2-2\epsilon \leq \|Ax-Ay\|_2^2-\|Ax+Ay\|_2^2 \leq \|x-y\|_2^2-\|x+y\|_2^2+2\epsilon$$

Esto es

$$\|Ax-Ay\|_2^2-\|Ax+Ay\|_2^2 \in \left[\|x-y\|_2^2-\|x+y\|_2^2 \pm 2\epsilon\right]$$

De esto es claro que la distancia entre $\|Ax-Ay\|_2^2-\|Ax+Ay\|_2^2$ y $\|x-y\|_2^2-\|x+y\|_2^2$ es de a lo más 4ϵ, osea

$$\left|\|Ax+Ay\|_2^2-\|Ax-Ay\|_2^2-\|x+y\|_2^2+\|x-y\|_2^2\right| \leq 4\epsilon$$

Usando la identidad de polarización se tiene que

$$|\langle Ax, Ay\rangle - \langle x, y\rangle| = \frac{1}{4}\left|\|Ax+Ay\|_2^2-\|Ax-Ay\|_2^2-\|x+y\|_2^2+\|x-y\|_2^2\right| \leq \frac{1}{4}\cdot 4\epsilon = \epsilon$$

Tal como queríamos demostrar

Pregunta 3

(a) Realizaremos la demostración a través de dos implicancias.

Supongamos que los datos son estrictamente separables y veamos que entonces el margen es estrictamente positivo. Como los datos son estrictamente separables entonces se tiene que existe algún $\bar{w}$ tal que

$$\langle \bar{w}, X_i \rangle Y_i > 0 \quad \forall i \in [n]$$

Entonces si dividimos en ambos lados de la desigualdad por el escalar $\|\bar{w}\|_2$, el cual es positivo puesto que $\bar{w} \neq 0$ ya que si $\bar{w} = 0$ no cumpliría la propiedad descrita anteriormente

$$\frac{\langle \bar{w}, X_i \rangle Y_i}{\|\bar{w}\|_2} > 0 \quad \forall i \in [n]$$

Como dicha desigualdad es cierta para todo $i \in [n]$, podemos concluír que para el mínimo también lo es, esto es

$$\min_{i \in [n]} \frac{\langle \bar{w}, X_i \rangle Y_i}{\|\bar{w}\|_2} > 0$$

Pero $\bar{w} \in \mathbb{R}^d$, por lo que si consideramos todos los vectores en $\mathbb{R}^d$ dicho margen debe ser igual (considerando el propio $\bar{w}$ como candidato) o más grande, esto es

$$\sup_{w \in \mathbb{R}^d} \min_{i \in [n]} \frac{\langle w, X_i \rangle Y_i}{\|w\|_2} \geq \min_{i \in [n]} \frac{\langle \bar{w}, X_i \rangle Y_i}{\|\bar{w}\|_2} > 0$$

Tal como queríamos demostrar. Ahora supongamos que el margen es estrictamente positivo y veamos que entonces los datos son estrictamente separables. Como el margen es estrictamente positivo tenemos que

$$\rho := \sup_{w \in \mathbb{R}^d} \min_{i \in [n]} \frac{\langle w, X_i \rangle Y_i}{\|w\|_2} > 0$$

Entonces, existen $w_1, w_2, ...$ tales que $\lim_{n \to \infty} \min_{i \in [n]} \frac{\langle w_n, X_i \rangle Y_i}{\|w_n\|_2} = \rho$, pero como $\min_{i \in [n]} \frac{\langle w, X_i \rangle Y_i}{\|w\|_2}$ es una función continua, se tiene que $\lim_{n \to \infty} w_n \to \bar{w}$, y como $\mathbb{R}^d$ es un espacio completo, entonces dicho $\bar{w} \in \mathbb{R}^d$, por lo que podemos concluir que

$$\rho = \min_{i \in [n]} \frac{\langle \bar{w}, X_i \rangle Y_i}{\|\bar{w}\|_2} > 0$$

Por la linealidad del producto interno sobre el primer argumento tenemos que

$$\rho = \min_{i\in[n]} \left\langle \frac{\bar{w}}{\|\bar{w}\|_2}, X_i \right\rangle Y_i > 0$$

Considerando $\hat{w} = \frac{\bar{w}}{\|\bar{w}\|_2}$ se tiene que

$$\rho = \min_{i\in[n]} \langle \hat{w}, X_i \rangle Y_i > 0$$

Pero para cualquier i su valor debe ser mayor o igual al mínimo, por lo tanto

$$\langle \hat{w}, X_i \rangle Y_i > 0 \quad \forall i \in [n]$$

Por lo que los datos son estrictamente separables por el vector $\hat{w}$, el cual en particular cumple que

$$\|\hat{w}\|_2 = \left\| \frac{\bar{w}}{\|\bar{w}\|_2} \right\|_2 = \frac{\|\bar{w}\|_2}{\|\bar{w}\|_2} = 1$$

Notemos que por ser un cuerpo sobre los números reales y recuperando un resultado anterior tenemos que

$$\rho = \min_{i\in[n]} \langle \hat{w}, X_i Y_i \rangle > 0$$

Ahora consideremos el semiespacio $\mathcal{H} = \{z \in \mathbb{R}^d : \langle \hat{w}, z \rangle \leq 0\}$. Como para cada $i \in [n]$ se tiene que $\langle \hat{w}, X_i Y_i \rangle > 0$, entonces para cada $i \in [n]$ se tiene que $X_i Y_i \notin \mathcal{H}$, por lo que la distancia desde $X_i Y_i$ al semiespacio $\mathcal{H}$ viene dada por su distancia al hiperplano $\{z \in \mathbb{R}^d : \langle \hat{w}, z \rangle = 0\}$, por lo que podemos usar la fórmula de distancia de un vector a un hiperplano (y recordando que el hiperplano es lineal) encontrando que la distancia para cada $X_i Y_i$ viene dada por

$$\frac{|\langle X_i Y_i, \hat{w} \rangle|}{\|\hat{w}\|_2} = \frac{|\langle X_i Y_i, \hat{w} \rangle|}{1} = |\langle X_i Y_i, \hat{w} \rangle| \quad \forall i \in [n]$$

Pero la distancia entre el conjunto $\{X_i Y_i : i \in [n]$ al semiespacio $\mathcal{H}$ viene dada por el mínimo de la distancia entre pares de puntos, por lo que iterando sobre el primer conjunto tenemos que la distancia corresponde con

$$\min_{i\in[n]} \langle |\hat{w}, X_i Y_i \rangle |$$

Como sabemos anteriormente que $\langle |\hat{w}, X_i Y_i \rangle > 0 \ \forall i \in [n]$, entonces podemos sacar el valor absoluto, obteniendo la siguiente expresión para la distancia

$$\min_{i\in[n]} \langle \hat{w}, X_i Y_i \rangle$$

Pero recordemos que de lo demostrado antes dicha expresión es igual a ρ, y como $\|\hat{w}\|_2 = 1$ entonces se demuestra lo pedido. La intuición geométrica es que el margen de un conjunto de datos corresponde con la mínima distancia para la cual podemos encontrar un hiperplano de tal forma que todos los datos estén en su lado correspondiente a una distancia mayor o igual al margen.

(b) Primero veamos que el conjunto $\{w \in \mathbb{R}^d : Y_i\langle w, X_i\rangle \geq 1 \quad \forall i \in [n]\}$ es cerrado y no vacío. Para ver que es cerrado basta notar que las restricciones son lineales y con $\geq$, por lo que el conjunto es cerrado. Para ver que es no vacío hay que notar que como los datos son estrictamente separables, entonces podemos considerar el $\hat{w}$ encontrado antes, el cual recordemos que cumple que $\|\hat{w}\|_2 = 1$ y que además se tiene que

$$\rho \leq |\langle X_i Y_i, \hat{w}\rangle| = \langle X_i Y_i, \hat{w}\rangle = Y_i\langle X_i, \hat{w}\rangle \quad \forall i \in [n]$$

Multiplicando ambos lados de la desigualdad por $\frac{1}{\rho}$ obtenemos que

$$1 = \frac{1}{\rho}\rho \leq Y_i\langle X_i, \hat{w}\rangle\frac{1}{\rho} = Y_i\left\langle X_i, \frac{\hat{w}}{\rho}\right\rangle \quad \forall i \in [n]$$

Por lo que tenemos que $\frac{\hat{w}}{\rho} \in \{w \in \mathbb{R}^d : Y_i\langle w, X_i\rangle \geq 1\}$ demostrando que el conjunto es no vacío.

Ahora veamos que la función $\frac{1}{2}\|w\|_2^2$ es coercitiva y estrictamente convexa. La función es coercitiva por definición puesto que si $\|w\|_2^2 \to \infty$ entonces $\frac{1}{2}\|w\|_2^2 \to \infty$ puesto que la constante no es relevante. Para ver que es estrictamente convexa podemos ver su matriz Hessiana. Llamemos f a la función, entonces tenemos que

$$f(w) = \frac{1}{2}\|w\|_2^2 = \frac{1}{2}\sum_{i=1}^{d} w_i^2$$

Por lo que estudiando la derivada parcial se tiene que

$$\frac{\partial f(w)}{\partial w_i} = \frac{1}{2}\cdot 2w_i = w_i$$

Por lo tanto $\frac{\partial f(w)}{\partial w_i \partial w_j} = 0$ si $i \neq j$ y $\frac{\partial f(w)}{\partial w_i \partial w_j} = 1$ si $i = j$. Por lo tanto la matriz Hessiana corresponde a la matriz identidad, la cual es definida positiva. Como la matriz Hessiana de $\frac{1}{2}\|w\|_2^2$ es definida positiva, se concluye que la función es estrictamente convexa.

Entonces, el problema de optimización (Hard-SVM) tiene un dominio cerrado no vacío con una función objetivo estrictamente convexa y coercitiva. De lo anterior por el teorema de Weierstrass extendido se concluye que la solución existe y es única.

Ahora notemos que si w es un candidato de solución óptima de (Hard-SVM) entonces debe existir $\hat{i}$ tal que $Y_{\hat{i}}\langle w, X_{\hat{i}}\rangle = 1$, ya que es inmediato que $w \neq 0$ y entonces por el estudio

anterior de la derivada parcial de la función objetivo existe alguna componente distinta a cero, entonces podemos movernos en dicha componente en la dirección que minimiza la función hasta alcanzar la igualdad en alguna restricción, lo cual mejora la solución.

Por lo que si w es candidato de solución óptima, entonces cumple que

$$\min_{i\in[n]} Y_{\hat{i}}\langle w, X_{\hat{i}}\rangle = 1$$

Por lo que un problema equivalente de (Hard-SVM) es

$$\min_{w\in\mathbb{R}^d}\left\{\frac{1}{2}\|w\|_2^2 : \min_{i\in[n]} Y_{\hat{i}}\langle w, X_{\hat{i}}\rangle = 1\right\}$$

Pero también podemos sacar la constante de la función objetivo, y como para los valores factibles la función objetivo es positiva, tenemos que otro problema equivalente de (Hard-SVM) es

$$\max_{w\in\mathbb{R}^d}\left\{\frac{1}{\|w\|_2} : \min_{i\in[n]} Y_{\hat{i}}\langle w, X_{\hat{i}}\rangle = 1\right\} =$$

Entonces si w^* es solución óptima de (Hard-SVM), también es solución óptima de este problema anterior, y como máximiza sobre $\frac{1}{\|w\|_2}$ y mantiene constante $\min_{i\in[n]} Y_{\hat{i}}\langle w, X_{\hat{i}}\rangle$, tenemos que

$$\frac{1}{\|w^*\|_2} = \sup_{w\in\mathbb{R}^d}\min_{i\in[n]}\frac{1}{\|w\|_2^2}\cdot Y_i\langle w, X_i\rangle = \rho$$

Por lo tanto w^* se corresponde con el $\bar{w}$ que encontramos en la pregunta anterior, el cual dijiimos que cumplía $\hat{w} = \frac{\bar{w}}{\|\bar{w}\|_2}$, por lo que podemos concluír que

$$\hat{w} = \frac{w^*}{\|w^*\|_2}$$

Demostrando las relaciones pedidas.

(c) Primero veamos que la función es convexa, por lo que solo necesitamos que se cumplan las condiciones de Slatter para poder usar el teorema de dualidad fuerte. Para ver que cumple las condiciones de Slatter podemos considerar una construcción análoga a la realizada para ver que el conjunto es no vacío (simplemente multiplicamos por $\frac{2}{\rho}$ en vez de por $\frac{1}{\rho}$) para mostrar que si los datos son estrictamente separables entonces existe un vector $\bar{w}$ tal que

$$Y_i\langle\bar{w}, X_i\rangle \geq 2 \;\; \forall i\in[n] \to Y_i\langle\bar{w}, X_i\rangle > 1 \;\; \forall i\in[n]$$

Por lo tanto, existe un punto tal que cumple todas las restricciones con holgura mayor a 0, así que se cumplen las condiciones de Slatter en este problema en particular.

Construyamos el langrangeano, para esto vamos a considerar $\alpha \in \mathbb{R}^n$ tal que $\alpha \geq 0$ (y ponemos un signo menos antes por el signo $\geq$ de la desigualdad), obteniendo que el problema primal es equivalente a

$$\max_{\alpha \geq 0} \min_{w \in \mathbb{R}^d} \mathcal{L}(w, \alpha) = \frac{1}{2}\|w\|_2^2 - \sum_{i=1}^{n} \alpha_i (Y_i \langle w, X_i \rangle - 1)$$

Lo cual lo podemos reescribir como

$$\max_{\alpha \geq 0} \min_{w \in \mathbb{R}^d} \mathcal{L}(w, \alpha) = \frac{1}{2}\sum_{i=1}^{d} w_i^2 - \sum_{i=1}^{n} \alpha_i (Y_i \langle w, X_i \rangle - 1)$$

Por lo que podemos cambiar el orden del mínimo y el máximo, obteniendo que el problema dual es equivalente a

$$\min_{w \in \mathbb{R}^d} \max_{\alpha \geq 0} \mathcal{L}(w, \alpha) = \frac{1}{2}\sum_{i=1}^{d} w_i^2 - \sum_{i=1}^{n} \alpha_i (Y_i \langle w, X_i \rangle - 1)$$

Para estudiar la relación entre los w y α en el óptimo podemos derivar con respecto a w e igualar a 0 (puesto que queremos minimizar sobre w), de lo que obtenemos que

$$\nabla_w \mathcal{L}(w, \alpha) = w - \sum_{i=1}^{n} \alpha_i Y_i X_i$$

Si lo igualamos a 0 obtenemos que

$$w - \sum_{i=1}^{n} \alpha_i Y_i X_i = 0 \to w = \sum_{i=1}^{n} \alpha_i Y_i X_i$$

Reemplazando por dicha relación de w en el langrangeano obtenemos el problema dual, el cual resulta que

$$\min_{w \in \mathbb{R}^d} \max_{\alpha \geq 0} \mathcal{L}(w, \alpha) = \frac{1}{2}\left\|\sum_{i=1}^{n} \alpha_i Y_i X_i\right\|_2^2 - \sum_{i=1}^{n} \alpha_i \left(Y_i \left\langle \sum_{j=1}^{n} \alpha_j Y_j X_j, X_i \right\rangle - 1\right)$$

Pero ya no hay dependencia sobre el w, por lo que podemos quitar el $\min_{w \in \mathbb{R}^d}$ y podemos dejar de hacer a $\mathcal{L}(w, \alpha)$ depender de w, y ponemos la función directamente para que quepa en la hoja

$$\max_{\alpha \geq 0} \frac{1}{2}\left\|\sum_{i=1}^{n} \alpha_i Y_i X_i\right\|_2^2 - \sum_{i=1}^{n} \alpha_i \left(Y_i \left\langle \sum_{j=1}^{n} \alpha_j Y_j X_j, X_i \right\rangle - 1\right)$$

Desarrollamos, expresamos el cuadrado de la norma de otra forma y separamos el paréntesis

$$\max_{\alpha\geq 0} \frac{1}{2}\left(\sum_{i=1}^{n}\alpha_i Y_i X_i\right)^T\left(\sum_{i=1}^{n}\alpha_i Y_i X_i\right) - \sum_{i=1}^{n}\alpha_i Y_i\left\langle\sum_{j=1}^{n}\alpha_j Y_j X_j, X_i\right\rangle + \sum_{i=1}^{n}\alpha_i$$

Distribuimos en el primer término y en el segundo usamos la linealidad de la esperanza

$$\max_{\alpha\geq 0} \frac{1}{2}\left(\sum_{i=1}^{n}\alpha_i Y_i X_i\right)^T\left(\sum_{i=1}^{n}\alpha_i Y_i X_i\right) - \sum_{i=1}^{n}\alpha_i Y_i\sum_{j=1}^{n}\alpha_j Y_j\left\langle X_j, X_i\right\rangle + \sum_{i=1}^{n}\alpha_i$$

Resolviendo la convolución en el primer término y metiendo el $\alpha_i Y_i$ en el segundo término obtenemos que

$$\max_{\alpha\geq 0} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i Y_i\alpha_j Y_j\left\langle X_j, X_i\right\rangle - \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i Y_i\alpha_j Y_j\left\langle X_j, X_i\right\rangle + \sum_{i=1}^{n}\alpha_i$$

Simplificando los terminos comunes obtenemos que

$$\max_{\alpha\geq 0} \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i Y_i\alpha_j Y_j\left\langle X_j, X_i\right\rangle$$

El cual es el problema dual de (Hard-SVM), tal como queríamos demostrar.

Como el problema cumple con las condiciones de Slatter y la función es convexa, entonces se tiene por el teorema de dualidad fuerte que si el primal es factible entonces el dual es factible y coinciden en el valor del óptimo, y por la relación que estudiamos anteriormente se tiene que si w^* y α^* son soluciones óptimas del primal y dual respectivamente, entonces debe cumplirse que

$$w^* = \sum_{i=1}^{n}\alpha_i^* Y_i X_i$$

Tal como queríamos demostrar. La interpretación geométrica de los vectores de soporte se la podemos dar usando el teorema de holguras complementarias. Como $\alpha_i^* > 0$ se tiene que en la i-ésima restricción del primal debe cumplirse con igualdad, esto es, $Y_i\langle w^*, X_i\rangle = 1$, por lo que el i-ésimo dato toca al hiperplano correspondiente, lo que podemos pensarlo como que lo "soporta".

Pregunta 4

Consideremos el conjunto $\bar{S} = S \cup -S$, y notemos que $|\bar{S}| = |S \cup -S| = |S| + |-S| + |S \cap -S| \leq |S| + |-S| = 2|S|$. Por lo que por lo demostrado en el punto 2b) considerando $\epsilon = \frac{2}{3}$, se tiene que existe una matriz $A \in \mathbb{R}^{k \times d}$ tal que es una matriz de baja distorsión en S para el ϵ dado, con k que cumple lo del resultado, esto es

$$k = \left\lfloor \frac{9 \ln(4|Q|^2)}{\epsilon^2} \right\rfloor$$

Como en este caso particular se tiene que $\epsilon = \frac{2}{3}$, $|Q| = 2|S|$, entonces

$$k = \left\lfloor \frac{9 \ln(4(2|S|)^2)}{\left(\frac{2}{3}\right)^2} \right\rfloor = \left\lfloor \frac{81 \ln(16|S|^2)}{4} \right\rfloor$$

El cual es parecido al k dado en el enunciado. La ventaja de este k obtenido es que como no depende de ρ, podemos encontrar el separador de la dimensión más baja sin conocer exactamente el separador de la dimensión más alta (ya que necesitamos conocer o al menos estimar ρ con el otro valor de k). Ahora, podemos aplicar el resultado de la pregunta 1c), de lo que concluímos que

$$|\langle Ax, Ay \rangle - \langle x, y \rangle| \leq \frac{2}{3} \quad \forall x, y \in S$$

Por el resultado 3b) y dado que los datos S son estrictamente separables con un margen ρ sabemos que existe un w tal que $\frac{1}{\|w\|_2} = \rho$ y $Y_i \langle w, X_i \rangle \geq 1$ para todo $i \in [n]$, podemos meter el Y_i dentro obteniendo que $\langle w, Y_i X_i \rangle \geq 1$ para todo $i \in [n]$. Por lo mostrado antes (y agregando el vector w a S ya que es lo mismo obviando un $+1$ constante), tenemos que existe una matriz A tal que

$$|Y_i \langle Aw, AX_i \rangle - Y_i \langle w, X_i \rangle| \leq \frac{2}{3} \quad \forall i \in [n]$$

De lo que se puede concluir que

$$Y_i \langle Aw, AX_i \rangle \geq 1 - \frac{2}{3} = \frac{1}{3} \quad \forall i \in [n]$$

Considerando $w' = 3w$, se tiene que

$$Y_i \langle Aw', AX_i \rangle = Y_i \langle A3w, AX_i \rangle = 3Y_i \langle Aw, AX_i \rangle \geq 3\frac{1}{3} = 1 \quad \forall i \in [n]$$

Por lo que considerando el problema (Hard-SVM) para S se tiene que w' es un punto factible. Considerando además que $\|w'\|_2^2 = \|3w\|_2^2 = 9\|w\|_2^2$, por lo que el punto óptimo x^* debe

cumplir que

$$\|w^*\|_2^2 \leq 9\|w\|_2^2$$

Puesto que en caso contrario w' sería mejor que el punto óptimo, absurdo. Si a dicha expresión le aplicamos la raíz cuadrada resulta que

$$\|w^*\|_2 \leq 3\|w\|_2$$

Y si la elevamos a -1, tenemos que

$$\frac{1}{\|w^*\|_2} \geq \frac{1}{3}\frac{1}{\|w\|_2} = \frac{1}{3}\rho$$

Así, se demuestra que el plano separador definido por el problema (Hard-SVM) para el conjunto de datos con la dimensionalidad reducida por la matriz A mencionada anteriormente, se cumple que el margen es de al menos $\frac{1}{3}\rho$, tal como queríamos demostrar.

Esto nos permite un enfoque aleatorizado para encontrar un hiperplano separador de los datos. Lo que podemos hacer es análogo a la construcción de A en la demostración del Lema de Johnson-Lindenstrauss dado en la pregunta 1b), construir una matriz A con dimensiones $k \times d$ en donde cada entrada la obtenemos a partir de una distribución normal. Sabemos que la probabilidad de tener éxito en la construcción es alta, y encima podemos calcular dicha cota y podemos considerar una dimensión más alta si creemos que es muy pequeña (de todas formas es una cota muy burda, así que debería ser decentemente probable). Además algo relevante es que le quité la dependencia del ρ para poder estimar de mejor manera la probabilidad de que A sea exitosa. Con dicha matriz A multiplicamos nuestros datos X por dicha matriz para reducir sus d dimensiones a k dimensiones donde k es pequeño (ya que tiene un logaritmo), por lo que los separadores sobre este nuevo espacio viven en $\mathbb{R}^k$. Buscamos un separador lineal en dicho conjunto de datos con menos características resolviendo el problema (Hard-SVM) o algo similar. En caso de encontrar dicho separador, entonces podemos clasificar nuevos datos usando la matriz A y el separador obtenido, logrando lo pedido. En caso de no encontrar un separador lo podemos intentar nuevamente con otra matriz A al azar.

La ventaja de este algoritmo es que al reducir la dimensión de los datos, entonces los algoritmos pueden funcionar mucho más rápido, ya que la complejidad computacional depende del número de características que tengan los datos.