

Pregunta 1

- (a) Como W es convexo, se tiene que $(1-t)\Pi_W^H(a) + t\Pi_W^H(b) \in W$ para todo $t \in [0, 1]$. Así, como $\Pi_W^H(a)$ es la proyección, entonces

$$\|(1-t)\Pi_W^H(a) + t\Pi_W^H(b) - a\| \geq \|\Pi_W^H(a) - a\| \quad \forall t \in [0, 1]$$

Por lo tanto, el óptimo de $\|(1-t)\Pi_W^H(a) + t\Pi_W^H(b) - a\|$ se alcanza en $t = 0$, por lo que en dicho valor la función debe ser no decreciente. Así, usando la regla del producto de las derivadas (y posteriormente evaluando en 0)

$$0 \leq \frac{d}{dt} \|(1-t)\Pi_W^H(a) + t\Pi_W^H(b) - a\|^2 \Big|_{t=0} = 2 \langle \Pi_W^H(b) - \Pi_W^H(a), \Pi_W^H(a) - a \rangle \quad (1)$$

De forma análoga con b

$$\langle \Pi_W^H(a) - \Pi_W^H(b), \Pi_W^H(b) - b \rangle \geq 0 \quad (2)$$

Ahora consideremos la función

$$d(t) = \|(1-t)\Pi_W^H(a) + ta - ((1-t)\Pi_W^H(b) + tb)\|^2$$

De (1) y (2) concluimos que

$$d'(0) = 2 \langle \Pi_W^H(a) - \Pi_W^H(b), a - \Pi_W^H(a) - b + \Pi_W^H(b) \rangle \geq 0$$

Por lo tanto, d es creciente en $[0, \infty)$. En particular, $d(1) \geq d(0)$, lo cual significa que $\|a - b\| \geq \|\Pi_W^H(a) - \Pi_W^H(b)\|$, tal como queríamos demostrar.

- (b) Utilizando la actualización del algoritmo tenemos que $w_{t+1} = \Pi_W^H(w_t - \eta H_t^{-1} g_t)$. Además, usando la propiedad demostrada en el punto anterior (y notando que $\Pi_W^H(w^*) = w^*$ puesto que $w^* \in W$) tenemos que

$$\|w_{t+1} - w^*\|_{H_t}^2 = \|\Pi_W^H(w_t - \eta H_t^{-1} g_t) - w^*\|_{H_t}^2 \leq \|w_t - \eta H_t^{-1} g_t - w^*\|_{H_t}^2$$

Esto es

$$\|w_{t+1} - w^*\|_{H_t}^2 \leq (w_t - \eta H_t^{-1} g_t - w^*)^T H_t (w_t - \eta H_t^{-1} g_t - w^*)$$

Distribuyendo convenientemente

$$\|w_{t+1} - w^*\|_{H_t}^2 \leq (w_t - w^*)^T H_t (w_t - w^*) + \eta^2 g_t^T H_t^{-1} g_t - \eta g_t^T H (w_t - w^*) - (w_t - w^*)^T H \eta g_t$$

Esto es

$$\|w_{t+1} - w^*\|_{H_t}^2 \leq \|w_t - w^*\|_H^2 + \eta^2 \|g_t\|_{H_t^{-1}}^2 - 2\eta \langle g_t, w_t - w^* \rangle_{H_t}$$

Demostrando lo pedido

(c) Dado que g_t es un gradiente estocástico de f en w_t , por definición

$$\mathbb{E}[g_t | \mathcal{F}_t] = \nabla f(w_t)$$

Ahora, como f es una función convexa, para cualquier $w \in W$, se cumple que

$$f(w) \geq f(w_t) + \langle \nabla f(w_t), w - w_t \rangle$$

En particular, para $w = w^*$ obtenemos

$$f(w^*) \geq f(w_t) + \langle \nabla f(w_t), w^* - w_t \rangle$$

Usando la relación de convexidad anterior y el hecho de que $\mathbb{E}[g_t | \mathcal{F}_t] = \nabla f(w_t)$, tenemos que

$$f(w^*) - f(w_t) \geq \langle \mathbb{E}[g_t | \mathcal{F}_t], w^* - w_t \rangle = \mathbb{E}[\langle g_t, w^* - w_t \rangle | \mathcal{F}_t]$$

Demostrando la primera parte de la pregunta. Reordenando el resultado de (b) tenemos que

$$2\eta \langle g_t, w_t - w^* \rangle \leq \|w_t - w^*\|_{H_t}^2 - \|w_{t+1} - w^*\|_{H_t}^2 + \eta^2 \|g_t\|_{H_t^{-1}}^2$$

Tomamos la expectativa condicional en ambos lados

$$2\eta \mathbb{E}[\langle g_t, w_t - w^* \rangle | \mathcal{F}_t] \leq \mathbb{E}\left[\|w_t - w^*\|_{H_t}^2 - \|w_{t+1} - w^*\|_{H_t}^2 + \eta^2 \|g_t\|_{H_t^{-1}}^2 | \mathcal{F}_t\right]$$

Antes demostramos que $\mathbb{E}[\langle g_t, w^* - w_t \rangle | \mathcal{F}_t] \leq f(w^*) - f(w_t)$, por lo que multiplicando por -1 podemos concluir que $f(w_t) - f(w^*) \leq \mathbb{E}[\langle g_t, w_t - w^* \rangle | \mathcal{F}_t]$. Usando esta ultima expresión resulta que

$$2\eta(f(w_t) - f(w^*)) \leq \mathbb{E}\left[\|w_t - w^*\|_{H_t}^2 - \|w_{t+1} - w^*\|_{H_t}^2 + \eta^2 \|g_t\|_{H_t^{-1}}^2 | \mathcal{F}_t\right]$$

Aplicando esperanza en ambos lados nuevamente (en el lado derecho se pierde la condicional al aplicar esperanza nuevamente)

$$\mathbb{E}[2\eta(f(w_t) - f(w^*))] \leq \mathbb{E}\left[\|w_t - w^*\|_{H_t}^2 - \|w_{t+1} - w^*\|_{H_t}^2 + \eta^2 \|g_t\|_{H_t^{-1}}^2\right]$$

Multiplicamos a ambos lados por $\frac{1}{2\eta T}$ y sacamos $f(w^*)$ de la esperanza puesto que no es aleatorio

$$\frac{1}{T}(\mathbb{E}[f(w_t)] - f(w^*)) \leq \mathbb{E}\left[\frac{1}{2\eta T}\|w_t - w^*\|_{H_t}^2 - \frac{1}{2\eta T}\|w_{t+1} - w^*\|_{H_t}^2 + \frac{\eta}{2T}\|g_t\|_{H_t^{-1}}^2\right]$$

Sumamos sobre todos los tiempos t , en el lado izquierdo resulta

$$\sum_{t=0}^{T-1} \frac{1}{T} (\mathbb{E}[f(w_t)] - f(w^*)) = \mathbb{E}\left[\frac{1}{T} \sum_{t=0}^{T-1} f(w_t)\right] - T \cdot \frac{1}{T} f(w^*) \geq \mathbb{E}[f(\bar{w}^T)] - f(w^*)$$

Donde la última desigualdad se obtiene de la convexidad de f . En el lado derecho nos resulta

$$\sum_{t=0}^{T-1} \mathbb{E} \left[\frac{1}{2\eta T} \|w_t - w^*\|_{H_t}^2 - \frac{1}{2\eta T} \|w_{t+1} - w^*\|_{H_t}^2 + \frac{\eta}{2T} \|g_t\|_{H_t^{-1}}^2 \right]$$

Usando la linealidad de la esperanza

$$\mathbb{E} \left[\sum_{t=0}^{T-1} \left(\frac{1}{2\eta T} \|w_t - w^*\|_{H_t}^2 - \frac{1}{2\eta T} \|w_{t+1} - w^*\|_{H_t}^2 + \frac{\eta}{2T} \|g_t\|_{H_t^{-1}}^2 \right) \right]$$

Reescribiendo la sumatoria (haciendo que el t pase de iterar sobre H_t a iterar sobre w_t) y el último w_{t+1} realmente no existe y solo abusamos de la notación, resulta en

$$\mathbb{E} \left[\frac{1}{2\eta T} \left(\|w_0 - w^*\|_{H_0}^2 + \sum_{t=1}^{T-1} (\|w_t - w^*\|_{H_t}^2 - \|w_t - w^*\|_{H_{t-1}}^2) \right) + \frac{\eta}{2T} \sum_{t=0}^{T-1} \|g_t\|_{H_t^{-1}}^2 \right]$$

Y en conjunto con lo anterior, demuestra lo pedido (lo separe en las dos partes de la desigualdad para que quepa en los márgenes)

(d) Comencemos de la definición de D_∞ , tenemos que

$$\|w_t - w^*\|_\infty \leq \sup_{w \in W} \|w - w^*\|_\infty = D_\infty$$

Pero como H_t es definida positiva podemos acotar la norma con respecto a dicha matriz de la forma siguiente

$$\|w_t - w^*\|_{H_t}^2 \leq D_\infty^2 \cdot \text{Tr}(H_t)$$

Entonces tenemos que

$$\sum_{t=1}^{T-1} (\|w_t - w^*\|_{H_t}^2 - \|w_t - w^*\|_{H_{t-1}}^2) \leq \sum_{t=1}^{T-1} (D_\infty^2 \cdot \text{Tr}(H_t) - D_\infty^2 \cdot \text{Tr}(H_{t-1}))$$

Podemos factorizar en el lado derecho

$$\sum_{t=1}^{T-1} (\|w_t - w^*\|_{H_t}^2 - \|w_t - w^*\|_{H_{t-1}}^2) \leq \sum_{t=1}^{T-1} D_\infty^2 \cdot \text{Tr}(H_t - H_{t-1})$$

Ahora podemos simplificar la suma telescopica, de lo que resulta

$$\sum_{t=1}^{T-1} (\|w_t - w^*\|_{H_t}^2 - \|w_t - w^*\|_{H_{t-1}}^2) \leq D_\infty^2 \cdot \text{Tr}(H_{T-1} - H_0)$$

Demostrando la primera parte. Para la segunda parte recordemos que la actualización de las matrices es de la forma

- (a) $R_t = R_{t-1} + g_t g_t^\top$
(b) $H_t = \text{Diag}(R_t)^{1/2}$

Podemos escribir la norma por definición

$$\|g_t\|_{H_t^{-1}}^2 = g_t^\top H_t^{-1} g_t = \sum_{i=1}^d \frac{(g_t)_i^2}{\sqrt{(R_t)_{ii}}} = \sum_{i=1}^d \frac{(g_t)_i^2}{\sqrt{(R_{t-1})_{ii} + (g_t)_i^2}} \leq \sum_{i=1}^d |(g_t)_i|$$

Entonces, la suma de estos términos desde $t = 0$ hasta $t = T - 1$ es

$$\sum_{t=0}^{T-1} g_t^\top H_t^{-1} g_t \leq \sum_{t=0}^{T-1} \sum_{i=1}^d |(g_t)_i|$$

Por otro lado, podemos ver que la traza de H_{T-1} , dada por

$$\text{tr}(H_{T-1}) = \sum_{i=1}^d \sqrt{(R_{T-1})_{ii}} = \sum_{i=1}^d \sqrt{\sum_{t=0}^{T-1} (g_t)_i^2} \geq \sum_{t=0}^{T-1} \sum_{i=1}^d \sqrt{(g_t)_i^2} = \sum_{t=0}^{T-1} \sum_{i=1}^d |(g_t)_i|$$

Lo que corresponde con la cota superior de $\|g_t\|_{H_t^{-1}}^2$, tal como queríamos demostrar

- (e) Consideremos $\eta = D_\infty$. De lo demostrado en la parte (c) consideremos el lado derecho (el cual es el que depende de η). Reemplazando nos resulta que

$$\mathbb{E} \left[\frac{1}{2D_\infty T} \left(\|w_0 - w^*\|_{H_0}^2 + \sum_{t=1}^{T-1} (\|w_t - w^*\|_{H_t}^2 - \|w_t - w^*\|_{H_{t-1}}^2) \right) + \frac{D_\infty}{2T} \sum_{t=0}^{T-1} \|g_t\|_{H_t^{-1}}^2 \right]$$

Además podemos remplazar por las cotas obtenidas en la parte (d)

$$\mathbb{E} \left[\frac{1}{2D_\infty T} (\|w_0 - w^*\|_{H_0}^2 + D_\infty^2 \cdot \text{Tr}(H_{T-1} - H_0)) + \frac{D_\infty}{2T} \text{Tr}(H_{T-1}) \right]$$

Pero notemos que por la propiedad de los espacios de Hilbert en dimensión finita tenemos que $\|w_0 - w^*\|_{H_0}^2 \leq D_\infty^2 \text{Tr}(H_0)$. De lo que resulta

$$\mathbb{E} \left[\frac{1}{2D_\infty T} (D_\infty^2 \text{Tr}(H_0) + D_\infty^2 \cdot \text{Tr}(H_{T-1} - H_0)) + \frac{D_\infty}{2T} \text{Tr}(H_{T-1}) \right]$$

Factorizando y sumando las trazas

$$\mathbb{E} \left[\frac{1}{2D_\infty T} (D_\infty^2 \cdot \text{Tr}(H_{T-1})) + \frac{D_\infty}{2T} \text{Tr}(H_{T-1}) \right]$$

De lo que resulta

$$\mathbb{E} \left[\frac{D_\infty}{2T} \text{Tr}(H_{T-1}) + \frac{D_\infty}{2T} \text{Tr}(H_{T-1}) \right] = \mathbb{E} \left[\frac{D_\infty}{T} \text{Tr}(H_{T-1}) \right]$$

Y recordando el lado izquierdo de la parte (c), se concluye que

$$\mathbb{E} [f(\bar{w}^T)] - f(w^*) \leq \mathbb{E} \left[\frac{D_\infty}{T} \text{Tr}(H_{T-1}) \right]$$

Obteniendo lo pedido. En comparación con la cota obtenida en clases hay 3 observaciones importantes

- (a) En función de T esta cota decae en el orden de $\frac{1}{T}$, mientras que la vista en clases decae en el orden de $\frac{1}{\sqrt{T}}$.
- (b) La cota vista en clases es determinística, mientras que esta cota depende sobre la esperanza de la traza de H_{T-1}
- (c) La cota vista en clases depende de la cota del segundo momento ν^2 mientras que esta no, lo cual puede ser beneficioso si dicha cota es grande.

Pregunta 2

Aplicamos Jensen:

$$f(\bar{w}_T) - f(w^*) \leq \frac{1}{\eta T} \sum_{t=0}^{T-1} \eta(f(w_t) - f(w^*))$$

Ahora, descomponemos la suma en tres términos:

$$\sum_{t=0}^{T-1} \eta(f(w_t) - f(w^*)) = \underbrace{\sum_{t=0}^{T-1} \eta(f(w_t, \xi_{t+1}) - f(w^*, \xi_{t+1}))}_{(A)} \quad (1)$$

$$+ \underbrace{\sum_{t=0}^{T-1} \eta(f(w_t) - f(w_t, \xi_{t+1}))}_{(B)} \quad (2)$$

$$+ \underbrace{\sum_{t=0}^{T-1} \eta(f(w^*) - f(w^*, \xi_{t+1}))}_{(C)} \quad (3)$$

El término A se acota usando la convexidad de f y la definición de la actualización w_{t+1} en el algoritmo.

$$(A) \leq \frac{1}{2} \left(\|w_0 - w^*\|_{H_0}^2 + \sum_{t=0}^{T-1} \eta_t \|G(w_t, \xi_{t+1})\|_{H_t^{-1}}^2 \right)$$

Debido a que $\|G(w_t, \xi_{t+1})\|_2 \leq \hat{L}$, podemos acotar el segundo término por \hat{L}^2 .

$$(A) \leq \frac{1}{2} \left(\|w_0 - w^*\|_{H_0}^2 + \hat{L}^2 \sum_{t=0}^{T-1} \eta_t^2 \right).$$

Para el término (C), notamos que es una suma de variables aleatorias independientes y acotadas, lo que permite aplicar la desigualdad de Hoeffding. Definimos $Z_t = \eta(f(w^*) - f(w^*, \xi_{t+1}))$, donde $|Z_t| \leq 2M$. Entonces, aplicando la desigualdad de Hoeffding

$$P \left(\sum_{t=0}^{T-1} Z_t > \epsilon_1 \right) \leq \exp \left(-\frac{\epsilon_1^2}{8M^2 T \eta^2} \right)$$

Para el término (B), usamos que es una suma de diferencias de martingalas acotadas $d_t = \eta(f(w_t) - f(w_t, \xi_{t+1}))$, donde $|d_t| \leq 2M$ y su valor esperado es cero. Aplicando la desigualdad de concentración para martingalas, obtenemos:

$$P \left(\sum_{t=0}^{T-1} d_t > \epsilon_2 \right) \leq \exp \left(-\frac{\epsilon_2^2}{8M^2 T \eta^2} \right).$$

Para obtener la cota final en alta probabilidad, combinamos las probabilidades de concentración de los términos (B) y (C) y despejamos ϵ_1 y ϵ_2 para obtener una cota total con probabilidad al menos $1 - \delta$

$$f(\bar{w}_T) - f(w^*) \leq \frac{1}{2\eta T} \left(\|w_0 - w^*\|_{H_0}^2 + \hat{L}^2 T \eta^2 \right) + \frac{4}{T\eta} \sqrt{2M^2 \log(2/\delta) T \eta^2}$$

Lo cual es análogo a lo encontrado en clases, ahora bien, con consideremos que $\|w_0 - w^*\|_{H_0}^2 \leq D_\infty^2 Tr(H_0)$ y $\eta = D_\infty$, tal como lo hicimos antes, resulta en que con probabilidad al menos $1 - \delta$

$$f(\bar{w}_T) - f(w^*) \leq \frac{1}{2D_\infty T} \left(D_\infty^2 Tr(H_0) + \hat{L}^2 T D_\infty^2 \right) + \frac{4}{T D_\infty} \sqrt{2M^2 \log(2/\delta) T D_\infty^2}$$

Desarrollando resulta que

$$f(\bar{w}_T) - f(w^*) \leq \frac{D_\infty Tr(H_0)}{2T} + \frac{D_\infty \hat{L}^2}{2} + \frac{6M}{\sqrt{T}} \sqrt{\log(2/\delta)}$$

Siendo esta la mejor cota de alta probabilidad que pude encontrar