PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
Escuela de ingeniería
ICS3153 Optimización Avanzada
Profesor: Cristóbal Guzmán (crguzmanp@uc.cl)
Alumno: Alberto Valdés (anvaldes@uc.cl)

Tarea 5

Pregunta 1

a) Notemos que:

$$\|x^{t+1} - x^*\|_{\mathbf{E}}^2 \overset{x^{t+1}=x^t-\eta_t \cdot g(x^t)}{=} \|x^t - \eta_t \cdot g(x^t) - x^*\|_{\mathbf{E}}^2 = \|[x^t - x^*] - \eta_t \cdot g(x^t)\|_{\mathbf{E}}^2 =$$

$$\|x^t - x^*\|_{\mathbf{E}}^2 - 2\eta_t \langle g(x^t), x^t - x^* \rangle + {\eta_t}^2 \cdot \|g(x^t)\|_{\mathbf{E}}^2 = \|x^t - x^*\|_{\mathbf{E}}^2 + 2\eta_t \langle g(x^t), x^* - x^t \rangle + {\eta_t}^2 \cdot \|g(x^t)\|_{\mathbf{E}}^2$$

$\Rightarrow$

$$\boxed{\|x^{t+1} - x^*\|_{\mathbf{E}}^2 = \|x^t - x^*\|_{\mathbf{E}}^2 + 2\eta_t \langle g(x^t), x^* - x^t \rangle + {\eta_t}^2 \cdot \|g(x^t)\|_{\mathbf{E}}^2 \quad (1)}$$

Notemos que como $g(x^t) \in \partial f(x^t)$, entonces:

$$f(x^t) + \langle g(x^t), u - x^t \rangle \leq f(u) \ \forall u \in E$$

Reemplazando en particular con $u = x^*$ tenemos que:

$$f(x^t) + \langle g(x^t), x^* - x^t \rangle \leq f(x^*)$$

$\Rightarrow$

$$\langle g(x^t), x^* - x^t \rangle \leq f(x^*) - f(x^t)$$

$\Rightarrow$

$$2\eta_t \langle g(x^t), x^* - x^t \rangle \leq 2\eta_t [f(x^*) - f(x^t)]$$

Y reemplazando en (1) llegamos a que:

$$\|x^{t+1} - x^*\|_{\mathbf{E}}^2 = \|x^t - x^*\|_{\mathbf{E}}^2 + 2\eta_t \langle g(x^t), x^* - x^t \rangle + {\eta_t}^2 \cdot \|g(x^t)\|_{\mathbf{E}}^2$$

$$\leq \|x^t - x^*\|_{\mathbf{E}}^2 + 2\eta_t [f(x^*) - f(x^t)] + {\eta_t}^2 \cdot \|g(x^t)\|_{\mathbf{E}}^2$$

$\Rightarrow$

$$\boxed{\|x^{t+1} - x^*\|_{\mathbf{E}}^2 \leq \|x^t - x^*\|_{\mathbf{E}}^2 + 2\eta_t[f(x^*) - f(x^t)] + {\eta_t}^2 \cdot \|g(x^t)\|_{\mathbf{E}}^2 \quad (2)}$$

Ahora considerando que $\eta_t = \frac{f(x^t)-f(x^*)}{\|g(x^t)\|_{\mathbf{E}}^2}$, y reemplazando en (2) llegamos a que:

$$\|x^{t+1} - x^*\|_{\mathbf{E}}^2 \leq \|x^t - x^*\|_{\mathbf{E}}^2 + 2\eta_t[f(x^*) - f(x^t)] + {\eta_t}^2 \cdot \|g(x^t)\|_{\mathbf{E}}^2$$

$$= \|x^t - x^*\|_{\mathbf{E}}^2 + 2\left(\frac{f(x^t) - f(x^*)}{\|g(x^t)\|_{\mathbf{E}}^2}\right)[f(x^*) - f(x^t)] + \left(\frac{f(x^t) - f(x^*)}{\|g(x^t)\|_{\mathbf{E}}^2}\right)^2 \cdot \|g(x^t)\|_{\mathbf{E}}^2$$

$$= \|x^t - x^*\|_{\mathbf{E}}^2 - 2 \cdot \frac{[f(x^t) - f(x^*)]^2}{\|g(x^t)\|_{\mathbf{E}}^2} + \frac{[f(x^t) - f(x^*)]^2}{\|g(x^t)\|_{\mathbf{E}}^4} \cdot \|g(x^t)\|_{\mathbf{E}}^2$$

$$= \|x^t - x^*\|_{\mathbf{E}}^2 - 2 \cdot \frac{[f(x^t) - f(x^*)]^2}{\|g(x^t)\|_{\mathbf{E}}^2} + \frac{[f(x^t) - f(x^*)]^2}{\|g(x^t)\|_{\mathbf{E}}^2} = \|x^t - x^*\|_{\mathbf{E}}^2 - \frac{[f(x^t) - f(x^*)]^2}{\|g(x^t)\|_{\mathbf{E}}^2}$$

$\Rightarrow$

$$\boxed{\|x^{t+1} - x^*\|_{\mathbf{E}}^2 \leq \|x^t - x^*\|_{\mathbf{E}}^2 - \frac{[f(x^t) - f(x^*)]^2}{\|g(x^t)\|_{\mathbf{E}}^2} \quad (3)}$$

Y (3) es justo lo que nos piden demostrar.

Ahora reordenando (3) tenemos que:

$$\frac{[f(x^t) - f(x^*)]^2}{\|g(x^t)\|_{\mathbf{E}}^2} \leq \|x^t - x^*\|_{\mathbf{E}}^2 - \|x^{t+1} - x^*\|_{\mathbf{E}}^2$$

Aplicando sumatoria a ambos lados tenemos que:

$$\sum_{t=0}^{T-1} \frac{[f(x^t) - f(x^*)]^2}{\|g(x^t)\|_{\mathbf{E}}^2} \leq \sum_{t=0}^{T-1} \left(\|x^t - x^*\|_{\mathbf{E}}^2 - \|x^{t+1} - x^*\|_{\mathbf{E}}^2\right) = \sum_{t=0}^{T-1} \|x^t - x^*\|_{\mathbf{E}}^2 - \sum_{t=0}^{T-1} \|x^{t+1} - x^*\|_{\mathbf{E}}^2$$

$$= \sum_{t=0}^{T-1} \|x^t - x^*\|_{\mathbf{E}}^2 - \sum_{t=1}^{T} \|x^t - x^*\|_{\mathbf{E}}^2 = \|x^0 - x^*\|_{\mathbf{E}}^2 + \sum_{t=1}^{T-1} \|x^t - x^*\|_{\mathbf{E}}^2 - \sum_{t=1}^{T-1} \|x^t - x^*\|_{\mathbf{E}}^2 - \|x^T - x^*\|_{\mathbf{E}}^2$$

$$= \|x^0 - x^*\|_{\mathbf{E}}^2 - \|x^T - x^*\|_{\mathbf{E}}^2 \overset{-\|x^T - x^*\|_{\mathbf{E}}^2 \leq 0}{\leq} \|x^0 - x^*\|_{\mathbf{E}}^2$$

$\Rightarrow$

$$\boxed{\sum_{t=0}^{T-1} \frac{[f(x^t) - f(x^*)]^2}{\|g(x^t)\|_{\mathbf{E}}^2} \leq \|x^0 - x^*\|_{\mathbf{E}}^2 \quad (4)}$$

Ahora notemos que es claro que $x^* \in \mathcal{X}$, pues $x^* \in argmin\{f(x) : x \in \mathcal{X}\}$.

Por otro lado, también notemos que $x^0 \in \mathcal{X}$ pues $x^0 \in argmin\{\Phi(x) : x \in \mathcal{X}\}$.

De esta forma:

$$\|x^0 - x^*\|_{\mathbf{E}}^2 \leq \sup_{x,y \in \mathcal{X}} \|x - y\|_{\mathbf{E}}^2 = diam_{\|\cdot\|_{\mathbf{E}}}^2(\mathcal{X})$$

Así reemplazando esto en (4) llegamos a que:

$$\sum_{t=0}^{T-1} \frac{[f(x^t) - f(x^*)]^2}{\|g(x^t)\|_{\mathbf{E}}^2} \leq \|x^0 - x^*\|_{\mathbf{E}}^2 \leq diam_{\|\cdot\|_{\mathbf{E}}}^2(\mathcal{X})$$

$\Rightarrow$

$$\boxed{\sum_{t=0}^{T-1} \frac{[f(x^t) - f(x^*)]^2}{\|g(x^t)\|_{\mathbf{E}}^2} \leq diam_{\|\cdot\|_{\mathbf{E}}}^2(\mathcal{X}) \quad (5)}$$

De esta manera, como demostramos (3) y como a partir de (3) se concluyo (5) entonces se demostró todo lo pedido en este inciso.

---

b) Lo primero que tenemos que tener en cuenta es la desigualdad a), es decir:

$$\boxed{\sum_{t=0}^{T-1} \frac{[f(x^t) - f(x^*)]^2}{\|g(x^t)\|_{\mathbf{E}}^2} \leq diam_{\|\cdot\|_{\mathbf{E}}}^2(\mathcal{X}) \quad (1)}$$

Por otro lado, notemos que como $f \in \mathcal{F}_{E,\|\cdot\|_{\mathbf{E}}}^0(L)$, entonces:

$$\|g(x^t)\|_{\mathbf{E},*} \leq L$$

Pero como sabemos que $\|\cdot\|_{\mathbf{E},*} = \|\cdot\|_{\mathbf{E}}$, entonces:

$$\|g(x^t)\|_{\mathbf{E}} \leq L \quad \Rightarrow \quad 1 \leq \frac{L}{\|g(x^t)\|_{\mathbf{E}}} \quad \Rightarrow \quad 1 \leq \frac{L^2}{\|g(x^t)\|_{\mathbf{E}}^2}$$

$\Rightarrow$

$$\boxed{1 \leq \frac{L^2}{\|g(x^t)\|_{\mathbf{E}}^2} \quad (2)}$$

Ahora usando (1) y (2) llegamos a que:

$$\sum_{t=0}^{T-1}[f(x^t) - f(x^*)]^2 \overset{Por(2)}{\leq} \sum_{t=0}^{T-1} \frac{L^2}{\|g(x^t)\|_{\mathbf{E}}^2}[f(x^t) - f(x^*)]^2 = L^2 \cdot \sum_{t=0}^{T-1} \frac{[f(x^t) - f(x^*)]^2}{\|g(x^t)\|_{\mathbf{E}}^2}$$

$$\overset{Por(1)}{\leq} L^2 \cdot diam^2_{\|\cdot\|_{\mathbf{E}}}(\mathcal{X})$$

$\Rightarrow$

$$\boxed{\sum_{t=0}^{T-1}[f(x^t) - f(x^*)]^2 \leq L^2 \cdot diam^2_{\|\cdot\|_{\mathbf{E}}}(\mathcal{X}) \quad (3)}$$

Ahora definiremos la siguiente función:

$$h(x) = [f(x) - f^*]^2$$

Y también definiendo las funciones $g(u) = u^2$ y $w(x) = f(x) - f^*$, podemos decir que:

$$h(x) = g(w(x))$$

Además como $w(x) = f(x) - f^* = f(x) - \text{mín}_{u \in E} f(u) \geq 0 \ \Rightarrow \ w(x) \geq 0 \ \forall x \in E$.

Ahora usando que $f(x)$ es convexa demostraremos que $w(x)$ también lo es.

Sean $x, y \in E$ arbitrarios y $\lambda \in [0, 1]$ también arbitrario, entonces:

$$w(\lambda x + [1-\lambda]y) = f(\lambda x + [1-\lambda]y) - f^* \overset{Convexidad\ de\ f}{\leq} \lambda f(x) + [1-\lambda]f(y) - f^* = \lambda f(x) + [1-\lambda]f(y) - [\lambda + (1-\lambda)]f^*$$

$$= \lambda f(x) - \lambda f^* + [1-\lambda]f(y) - [1-\lambda]f^* = \lambda \cdot (f(x) - f^*) + [1-\lambda] \cdot (f(y) - f^*) \overset{Definicion\ de\ w}{=} \lambda \cdot w(x) + [1-\lambda] \cdot w(y)$$

$\Rightarrow$

$$w(\lambda x + [1-\lambda]y) \leq \lambda \cdot w(x) + [1-\lambda] \cdot w(y)$$

Y como se demostró para un $\lambda \in [0,1]$ arbitrario y para $x, y \in E$ arbitrarios también, entonces:

$$\boxed{w(\lambda x + [1-\lambda]y) \leq \lambda \cdot w(x) + [1-\lambda] \cdot w(y) \quad \forall \lambda \in [0,1], \forall x, y \in E \quad (4)}$$

$\Rightarrow$

$$\boxed{w \text{ es convexa} \quad (5)}$$

Notemos que como $g(u) = u^2 \;\Rightarrow\; g'(u) = 2u \;\Rightarrow\; g''(u) = 2$ lo cual implica que g es convexa y no-decreciente para todo $u \geq 0$.

Ahora consideremos $x, y \in E$ arbitrarios junto con un $\lambda \in [0,1]$ también arbitrario.

Por (4) tenemos que:

$$w(\lambda \cdot x + [1-\lambda] \cdot y) \leq \lambda \cdot w(x) + [1-\lambda] \cdot w(y)$$

Y como dijimos que $w(x) \geq 0 \; \forall x \in E$ y como g es creciente para todo $u \geq 0$, entonces:

$$\boxed{g(w(\lambda \cdot x + [1-\lambda] \cdot y)) \leq g(\lambda \cdot w(x) + [1-\lambda] \cdot w(y)) \quad (6)}$$

Ahora como g es convexa, entonces:

$$\boxed{g(\lambda \cdot w(x) + [1-\lambda] \cdot w(y)) \leq \lambda \cdot g(w(x)) + [1-\lambda] \cdot g(w(y)) \quad (7)}$$

Ahora usando (6) y (7) tenemos que:

$$h(\lambda \cdot x + [1-\lambda] \cdot y) = g(w(\lambda \cdot x + [1-\lambda] \cdot y)) \overset{Por\ (6)}{\leq} g(\lambda \cdot w(x) + [1-\lambda] \cdot w(y))$$

$$\overset{Por\ (7)}{\leq} \lambda \cdot g(w(x)) + [1-\lambda] \cdot g(w(y)) \overset{Definicion\ de\ h}{=} \lambda \cdot h(x) + [1-\lambda] \cdot h(y)$$

$\Rightarrow$

$$h(\lambda \cdot x + [1-\lambda] \cdot y) \leq \lambda \cdot h(x) + [1-\lambda] \cdot h(y)$$

Como esto se demostró para $x, y \in E$ arbitrarios y para $\lambda \in [0,1]$ arbitrario, entonces:

$$\boxed{h(\lambda \cdot x + [1-\lambda] \cdot y) \leq \lambda \cdot h(x) + [1-\lambda] \cdot h(y) \quad \forall x, y \in E, \forall \lambda \in [0,1] \quad (8)}$$

Por lo tanto:

$$\boxed{h \text{ es convexa} \quad (9)}$$

Ahora usando la convexidad de h y la desigualdad de Jensen, tenemos que:

$$\left[f\left(\frac{1}{T}\cdot\sum_{t=0}^{T-1}x^t\right)-f^*\right]^2 = h\left(\frac{1}{T}\cdot\sum_{t=0}^{T-1}x^t\right) = h\left(\sum_{t=0}^{T-1}\frac{1}{T}\cdot x^t\right) \overset{Por\ Jensen}{\leq} \sum_{t=0}^{T-1}\frac{1}{T}\cdot h(x^t)$$

$$= \frac{1}{T}\cdot\sum_{t=0}^{T-1}h(x^t) \overset{Definicion\ de\ h}{=} \frac{1}{T}\cdot\sum_{t=0}^{T-1}[f(x^t)-f^*]^2 \overset{Por\ (3)}{\leq} \frac{L^2\cdot diam^2_{\|\cdot\|_{\mathbf{E}}}(\mathcal{X})}{T}$$

$\Rightarrow$

$$\left[f\left(\frac{1}{T}\cdot\sum_{t=0}^{T-1}x^t\right)-f^*\right]^2 \leq \frac{L^2\cdot diam^2_{\|\cdot\|_{\mathbf{E}}}(\mathcal{X})}{T}$$

Ahora como sabemos que tanto $f\left(\frac{1}{T}\cdot\sum_{t=0}^{T-1}x^t\right)-f^*$ como $\dfrac{L^2\cdot diam^2_{\|\cdot\|_{\mathbf{E}}}(\mathcal{X})}{T}$ son cantidades positivas, entonces a partir de lo anterior podemos deducir que:

$$\boxed{f\left(\frac{1}{T}\cdot\sum_{t=0}^{T-1}x^t\right)-f^* \leq \frac{L\cdot diam_{\|\cdot\|_{\mathbf{E}}}(\mathcal{X})}{\sqrt{T}} \quad (10)}$$

Y así se demostró justamente lo pedido.

Pregunta 2

a) Notemos que por definición de subgradiente tenemos que:

$$f(x^t) + \langle g(x^t), y - x^t \rangle \leq f(y) \quad \forall y \in E$$

Por lo tanto:

$$f(x^t) + \langle g(x^t), y - x^t \rangle \leq f(y) \quad \forall y \in \mathcal{X}$$

Así tenemos que sea $u \in \mathcal{X}$ arbitrario, entonces:

$$f(x^t) + \langle g(x^t), u - x^t \rangle \leq f(u)$$

$\Rightarrow$

$$f(x^t) - f(u) \leq -\langle g(x^t), u - x^t \rangle = \langle g(x^t), x^t - u \rangle = \langle g(x^t), x^t - x^{t+1} + x^{t+1} - u \rangle = \langle g(x^t), x^t - x^{t+1} \rangle + \langle g(x^t), x^{t+1} - u \rangle$$

$\Rightarrow$

$$\boxed{f(x^t) - f(u) \leq \langle g(x^t), x^t - x^{t+1} \rangle + \langle g(x^t), x^{t+1} - u \rangle \quad (1)}$$

Ahora recordemos como se obtiene x^{t+1}.

$$x^{t+1} = argmin_{w \in \mathcal{X}} \{ f(x^t) + \langle g(x^t), w - x^t \rangle + \frac{1}{\eta_t} V_{x^t}(w) \}$$

Notemos que si definimos:

$$F_{x^t}(w) = f(x^t) + \langle g(x^t), w - x^t \rangle + \frac{1}{\eta_t} V_{x^t}(w)$$

Entonces la definición de x^{t+1} nos permite decir que:

$$0 \in \{ \nabla F_{x^t}(x^{t+1}) \} + N_{\mathcal{X}}(x^{t+1})$$

$\Rightarrow$

$$-\nabla F_{x^t}(x^{t+1}) \in N_{\mathcal{X}}(x^{t+1})$$

Y usando la definición de $F_{x^t}(w)$ tenemos que:

$$\nabla F_{x^t}(w) = g(x^t) + \frac{1}{\eta_t} \nabla V_{x^t}(w) = g(x^t) + \frac{1}{\eta_t} \cdot \left[\nabla \Phi(w) - \nabla \Phi(x^t) \right]$$

De esta forma, tenemos que:

$$-\left(g(x^t) + \frac{1}{\eta_t} \cdot \left[\nabla \Phi(x^{t+1}) - \nabla \Phi(x^t) \right] \right) \in N_{\mathcal{X}}(x^{t+1})$$

$\Rightarrow$

$$-\langle g(x^t) + \frac{1}{\eta_t} \cdot \left[\nabla\Phi(x^{t+1}) - \nabla\Phi(x^t)\right], a - x^{t+1}\rangle \leq 0 \quad \forall a \in \mathcal{X}$$

$\Rightarrow$

$$\langle g(x^t) + \frac{1}{\eta_t} \cdot \left[\nabla\Phi(x^{t+1}) - \nabla\Phi(x^t)\right], a - x^{t+1}\rangle \geq 0 \quad \forall a \in \mathcal{X}$$

$\Rightarrow$

$$\langle g(x^t), a - x^{t+1}\rangle \geq -\frac{1}{\eta_t} \cdot \langle \left[\nabla\Phi(x^{t+1}) - \nabla\Phi(x^t)\right], a - x^{t+1}\rangle \quad \forall a \in \mathcal{X}$$

$\Rightarrow$

$$\langle g(x^t), x^{t+1} - a\rangle \leq \frac{1}{\eta_t} \cdot \langle \left[\nabla\Phi(x^{t+1}) - \nabla\Phi(x^t)\right], a - x^{t+1}\rangle \quad \forall a \in \mathcal{X}$$

Y como tenemos que $u \in \mathcal{X}$ entonces la desigualdad anterior se da también reemplazando con $a = u$.

De este modo:

$$\boxed{\langle g(x^t), x^{t+1} - u\rangle \leq \frac{1}{\eta_t} \cdot \langle \left[\nabla\Phi(x^{t+1}) - \nabla\Phi(x^t)\right], u - x^{t+1}\rangle \quad (2)}$$

Ahora usando (1) y (2) tenemos que:

$$f(x^t) - f(u) \leq \langle g(x^t), x^t - x^{t+1}\rangle + \langle g(x^t), x^{t+1} - u\rangle \overset{Por\ (2)}{\leq} \langle g(x^t), x^t - x^{t+1}\rangle + \frac{1}{\eta_t} \cdot \langle \left[\nabla\Phi(x^{t+1}) - \nabla\Phi(x^t)\right], u - x^{t+1}\rangle$$

$\Rightarrow$

$$\boxed{f(x^t) - f(u) \leq \langle g(x^t), x^t - x^{t+1}\rangle + \frac{1}{\eta_t} \cdot \langle \left[\nabla\Phi(x^{t+1}) - \nabla\Phi(x^t)\right], u - x^{t+1}\rangle \quad (3)}$$

Ahora notemos que por el Lema de 3 puntos tenemos que:

$$\langle \nabla\Phi(x^{t+1}) - \nabla\Phi(x^t), u - x^{t+1}\rangle = V_{x^t}(u) - V_{x^{t+1}}(u) - V_{x^t}(x^{t+1})$$

Reemplazando esto en (3) llegamos a que:

$$f(x^t) - f(u) \leq \langle g(x^t), x^t - x^{t+1}\rangle + \frac{1}{\eta_t} \cdot \left[V_{x^t}(u) - V_{x^{t+1}}(u) - V_{x^t}(x^{t+1})\right]$$

$$= \frac{1}{\eta_t} \cdot \left[V_{x^t}(u) - V_{x^{t+1}}(u)\right] + \left(\langle g(x^t), x^t - x^{t+1}\rangle - \frac{1}{\eta_t} \cdot V_{x^t}(x^{t+1})\right)$$

$\Rightarrow$

$$f(x^t) - f(u) \leq \frac{1}{\eta_t} \cdot \left[V_{x^t}(u) - V_{x^{t+1}}(u)\right] + \left(\langle g(x^t), x^t - x^{t+1}\rangle - \frac{1}{\eta_t} \cdot V_{x^t}(x^{t+1})\right)$$

Y como $u \in \mathcal{X}$ es arbitrario, entonces podemos tomar $u = x^* \in \mathcal{X}$.

De este modo:

$$\boxed{f(x^t) - f^* \leq \frac{1}{\eta_t} \cdot [V_{x^t}(x^*) - V_{x^{t+1}}(x^*)] + \left(\langle g(x^t), x^t - x^{t+1} \rangle - \frac{1}{\eta_t} \cdot V_{x^t}(x^{t+1}) \right)} \quad (4)$$

Ahora nos centraremos en acotar $\left(\langle g(x^t), x^t - x^{t+1} \rangle - \frac{1}{\eta_t} \cdot V_{x^t}(x^{t+1}) \right)$ y lo que haremos es usar en primer lugar que $V_{x^t}(x^{t+1})$ es $1-$fuertemente convexa, por ende:

$$V_{x^t}(x^{t+1}) \geq \frac{1}{2} \|x^t - x^{t+1}\|^2$$

$\Rightarrow$

$$-V_{x^t}(x^{t+1}) \leq -\frac{1}{2} \|x^t - x^{t+1}\|^2$$

$\Rightarrow$

$$\langle g(x^t), x^t - x^{t+1} \rangle - \frac{1}{\eta_t} \cdot V_{x^t}(x^{t+1}) \leq \langle g(x^t), x^t - x^{t+1} \rangle - \frac{1}{2\eta_t} \|x^t - x^{t+1}\|^2$$

$$\overset{Usando\ que\ \eta_t = \frac{1}{\mu}}{=} \langle g(x^t), x^t - x^{t+1} \rangle - \frac{\mu}{2} \|x^t - x^{t+1}\|^2$$

$\Rightarrow$

$$\boxed{\langle g(x^t), x^t - x^{t+1} \rangle - \frac{1}{\eta_t} \cdot V_{x^t}(x^{t+1}) \leq \langle g(x^t), x^t - x^{t+1} \rangle - \frac{\mu}{2} \|x^t - x^{t+1}\|^2} \quad (5)$$

Usando (4) y (5) llegamos a que:

$$f(x^t) - f^* \leq \frac{1}{\eta_t} \cdot [V_{x^t}(x^*) - V_{x^{t+1}}(x^*)] + \left(\langle g(x^t), x^t - x^{t+1} \rangle - \frac{1}{\eta_t} \cdot V_{x^t}(x^{t+1}) \right)$$

$$\overset{Por\ (5)}{\leq} \frac{1}{\eta_t} \cdot [V_{x^t}(x^*) - V_{x^{t+1}}(x^*)] + \langle g(x^t), x^t - x^{t+1} \rangle - \frac{\mu}{2} \|x^t - x^{t+1}\|^2$$

$\Rightarrow$

$$\boxed{f(x^t) - f^* \leq \frac{1}{\eta_t} \cdot [V_{x^t}(x^*) - V_{x^{t+1}}(x^*)] + \langle g(x^t), x^t - x^{t+1} \rangle - \frac{\mu}{2} \|x^t - x^{t+1}\|^2} \quad (6)$$

Ahora notemos que por propiedades de μ-suavidad tenemos que:

$$f(y) \leq f(x^t) + \langle g(x^t), y - x^t \rangle + \frac{\mu}{2} \cdot \|x^t - y\|^2 \quad \forall y \in \mathcal{X}$$

Y en particular tomando $y = x^{t+1}$ el cual pertenece a $\mathcal{X}$ por definición de como se obtiene x^{t+1}.

Así tenemos que:

$$f(x^{t+1}) \leq f(x^t) + \langle g(x^t), x^{t+1} - x^t \rangle + \frac{\mu}{2} \cdot \|x^t - x^{t+1}\|^2$$

$\Rightarrow$

$$f(x^{t+1}) - f(x^t) \leq \langle g(x^t), x^{t+1} - x^t \rangle + \frac{\mu}{2} \cdot \|x^t - x^{t+1}\|^2$$

$\Rightarrow$

$$f(x^t) - f(x^{t+1}) \geq -\langle g(x^t), x^{t+1} - x^t \rangle - \frac{\mu}{2} \cdot \|x^t - x^{t+1}\|^2 = \langle g(x^t), x^t - x^{t+1} \rangle - \frac{\mu}{2} \cdot \|x^t - x^{t+1}\|^2$$

$\Rightarrow$

$$f(x^t) - f(x^{t+1}) \geq \langle g(x^t), x^t - x^{t+1} \rangle - \frac{\mu}{2} \cdot \|x^t - x^{t+1}\|^2$$

$\Rightarrow$

$$\boxed{\langle g(x^t), x^t - x^{t+1} \rangle - \frac{\mu}{2} \cdot \|x^t - x^{t+1}\|^2 \leq f(x^t) - f(x^{t+1}) \quad (7)}$$

Usando (6) y (7) tenemos que:

$$f(x^t) - f^* \leq \frac{1}{\eta_t} \cdot [V_{x^t}(x^*) - V_{x^{t+1}}(x^*)] + \langle g(x^t), x^t - x^{t+1} \rangle - \frac{\mu}{2} \|x^t - x^{t+1}\|^2$$

$$\overset{Por\ (7)}{\leq} \frac{1}{\eta_t} \cdot [V_{x^t}(x^*) - V_{x^{t+1}}(x^*)] + f(x^t) - f(x^{t+1})$$

$\Rightarrow$

$$f(x^t) - f^* \leq \frac{1}{\eta_t} \cdot [V_{x^t}(x^*) - V_{x^{t+1}}(x^*)] + f(x^t) - f(x^{t+1})$$

$\Rightarrow$

$$\boxed{f(x^{t+1}) - f^* \leq \frac{1}{\eta_t} \cdot [V_{x^t}(x^*) - V_{x^{t+1}}(x^*)] \quad (8)}$$

Lo cual es justamente lo que nos pedían demostrar.

---

b) Notemos que por a) tenemos que:

$$f(x^{t+1}) - f^* \leq \frac{1}{\eta_t}[V_{x^t}(x^*) - V_{x^{t+1}}(x^*)] \qquad \forall t = 0, ..., T-1$$

Entonces esto es lo mismo a decir que:

$$f(x^t) - f^* \leq \frac{1}{\eta_{t-1}}[V_{x^{t-1}}(x^*) - V_{x^t}(x^*)] \qquad \forall t = 1, ..., T$$

De este modo si aplicamos sumatoria desde $t = 1$ hasta $t = T$ obtenemos que:

$$\sum_{t=1}^{T}[f(x^t) - f^*] \leq \sum_{t=1}^{T} \frac{1}{\eta_{t-1}}[V_{x^{t-1}}(x^*) - V_{x^t}(x^*)] \overset{Como\ \eta_{t-1}=\frac{1}{\mu}}{=} \sum_{t=1}^{T} \frac{1}{\frac{1}{\mu}}[V_{x^{t-1}}(x^*) - V_{x^t}(x^*)]$$

$$= \sum_{t=1}^{T} \mu \cdot [V_{x^{t-1}}(x^*) - V_{x^t}(x^*)] = \mu \cdot \sum_{t=1}^{T}[V_{x^{t-1}}(x^*) - V_{x^t}(x^*)] = \mu \cdot \left(\sum_{t=1}^{T} V_{x^{t-1}}(x^*) - \sum_{t=1}^{T} V_{x^t}(x^*) \right)$$

$$= \mu \cdot \left(\sum_{t=0}^{T-1} V_{x^t}(x^*) - \sum_{t=1}^{T} V_{x^t}(x^*) \right) = \mu \cdot \left(V_{x^0}(x^*) + \sum_{t=1}^{T-1} V_{x^t}(x^*) - \sum_{t=1}^{T-1} V_{x^t}(x^*) - V_{x^T}(x^*) \right)$$

$$= \mu \cdot [V_{x^0}(x^*) - V_{x^T}(x^*)]$$

$\Rightarrow$

$$\boxed{\sum_{t=1}^{T}[f(x^t) - f^*] \leq \mu \cdot [V_{x^0}(x^*) - V_{x^T}(x^*)] \quad (1)}$$

Ahora utilizando la desigualdad de Jensen ya que f es convexa, entonces tenemos que:

$$f\left(\frac{1}{T} \sum_{t=1}^{T} x^t \right) = f\left(\sum_{t=1}^{T} \left[\frac{1}{T} \right] \cdot x^t \right) \leq \sum_{t=1}^{T} \left(\left[\frac{1}{T} \right] \cdot f(x^t) \right) = \frac{1}{T} \cdot \sum_{t=1}^{T} f(x^t)$$

$\Rightarrow$

$$\boxed{f\left(\frac{1}{T} \sum_{t=1}^{T} x^t \right) \leq \frac{1}{T} \cdot \sum_{t=1}^{T} f(x^t) \quad (2)}$$

De esta forma:

$$f\left(\frac{1}{T} \sum_{t=1}^{T} x^t \right) - f^* \overset{Por\ (2)}{\leq} \frac{1}{T} \cdot \sum_{t=1}^{T} f(x^t) - f^* = \frac{1}{T} \cdot \sum_{t=1}^{T} f(x^t) - \frac{T \cdot f^*}{T} = \frac{1}{T} \cdot \sum_{t=1}^{T} f(x^t) - \frac{\sum_{t=1}^{T} f^*}{T}$$

$$= \frac{1}{T} \cdot \sum_{t=1}^{T} f(x^t) - \frac{1}{T} \sum_{t=1}^{T} f^* = \frac{1}{T} \sum_{t=1}^{T}[f(x^t) - f^*] \overset{Por\ (1)}{\leq} \frac{1}{T} \cdot \mu \cdot [V_{x^0}(x^*) - V_{x^T}(x^*)]$$

$\Rightarrow$

$$f\left(\frac{1}{T}\sum_{t=1}^{T}x^t\right)-f^*\leq\frac{\mu}{T}\cdot[V_{x^0}(x^*)-V_{x^T}(x^*)]$$

$\Rightarrow$

$$\boxed{f\left(\frac{1}{T}\sum_{t=1}^{T}x^t\right)-f^*\leq\frac{\mu}{T}\cdot[V_{x^0}(x^*)-V_{x^T}(x^*)]\quad(3)}$$

Ahora notemos que por definición de divergencia de Bregman, tenemos que:

$$V_{x^T}(x^*)=\Phi(x^*)-\Phi(x^T)-\langle\nabla\Phi(x^T),x^*-x^T\rangle$$

Y como $\Phi(\cdot)$ es 1-fuertemente convexa, entonces:

$$\Phi(\lambda x+[1-\lambda]y)\leq\lambda\Phi(x)+[1-\lambda]\Phi(y)-\frac{1}{2}\lambda(1-\lambda)\|x-y\|_{\mathbf{E}}^2\quad\forall x,y\in E,\forall\lambda\in[0,1]$$

Y como $\lambda\in[0,1]$ y además como por definición de norma tenemos que es siempre positiva, entonces:

$$-\frac{1}{2}\lambda(1-\lambda)\|x-y\|_{\mathbf{E}}^2\leq 0$$

De este modo, usando que $\Phi(\cdot)$ es 1-fuertemente convexa y lo recién demostrado tenemos que:

$$\Phi(\lambda x+[1-\lambda]y)\leq\lambda\Phi(x)+[1-\lambda]\Phi(y)-\frac{1}{2}\lambda(1-\lambda)\|x-y\|_{\mathbf{E}}^2\leq\lambda\Phi(x)+[1-\lambda]\Phi(y)\quad\forall x,y\in E,\forall\lambda\in[0,1]$$

$\Rightarrow$

$$\Phi(\lambda x+[1-\lambda]y)\leq\lambda\Phi(x)+[1-\lambda]\Phi(u)\quad\forall x,y\in E,\forall\lambda\in[0,1]$$

$\Rightarrow$

$$\Phi(\cdot)\text{ es convexa}$$

De este modo se tiene que:

$$\Phi(x^T)+\langle\nabla\Phi(x^T),u-x^T\rangle\leq\Phi(u)\quad\forall u\in E$$

Y en particular esto se cumple para $u=x^*$ de este modo:

$$\Phi(x^T)+\langle\nabla\Phi(x^T),x^*-x^T\rangle\leq\Phi(x^*)$$

$\Rightarrow$

$$0\leq\Phi(x^*)-\Phi(x^T)-\langle\nabla\Phi(x^T),x^*-x^T\rangle$$

Y por definición de divergencia de Bregman, tenemos que esto implica que:

$$0 \leq V_{x^T}(x^*)$$

$\Rightarrow$

$$-V_{x^T}(x^*) \leq 0$$

Usando esto en (3) llegamos a que:

$$f\left(\frac{1}{T}\sum_{t=1}^{T} x^t\right) - f^* \leq \frac{\mu}{T} \cdot [V_{x^0}(x^*) - V_{x^T}(x^*)] \leq \frac{\mu}{T} \cdot V_{x^0}(x^*)$$

$$\overset{Definicion\ de\ divegencia\ de\ Bregman}{=} \frac{\mu}{T} \cdot [\Phi(x^*) - \Phi(x^0) - \langle \nabla\Phi(x^0), x^* - x^0\rangle]$$

$\Rightarrow$

$$\boxed{f\left(\frac{1}{T}\sum_{t=1}^{T} x^t\right) - f^* \leq \frac{\mu}{T}[\Phi(x^*) - \Phi(x^0) - \langle \nabla\Phi(x^0), x^* - x^0\rangle] \quad (4)}$$

Ahora como $x^0 \in argmin\{\Phi(x) : x \in \mathcal{X}\}$, entonces:

$$0 \in N_{\mathcal{X}}(x^0) + \partial\Phi(x^0)$$

Y como $\partial\Phi(x^0) = \{\nabla\Phi(x^0)\}$, entonces:

$$\boxed{\exists g \in N_{\mathcal{X}}(x^0) : g = -\nabla\Phi(x^0) \quad (5)}$$

Y como sabemos que:

$$N_{\mathcal{X}}(x^0) = \{g \in E : \langle g, u - x^0\rangle \leq 0 \ \ \forall u \in \mathcal{X}\}$$

Entonces usando (5) tenemos que:

$$\langle -\nabla\Phi(x^0), u - x^0\rangle \leq 0 \ \ \forall u \in \mathcal{X} \quad \Rightarrow \quad -\langle \nabla\Phi(x^0), u - x^0\rangle \leq 0 \ \ \forall u \in \mathcal{X}$$

Como $x^* \in argmin\{f(x) : x \in \mathcal{X}\} \ \Rightarrow \ x^* \in \mathcal{X}$, así podemos usar lo demostrado recién para $u = x^*$.

De este modo:

$$\boxed{-\langle \nabla\Phi(x^0), x^* - x^0\rangle \leq 0 \quad (6)}$$

Así usando (6) en (4) llegamos a que:

$$f\left(\frac{1}{T}\sum_{t=1}^{T}x^t\right)-f^*\leq\frac{\mu}{T}[\Phi(x^*)-\Phi(x^0)-\langle\nabla\Phi(x^0),x^*-x^0\rangle]\overset{Por\ (6)}{\leq}\frac{\mu}{T}[\Phi(x^*)-\Phi(x^0)]$$

$\Rightarrow$

$$\boxed{f\left(\frac{1}{T}\sum_{t=1}^{T}x^t\right)-f^*\leq\frac{\mu}{T}[\Phi(x^*)-\Phi(x^0)]\quad(7)}$$

Como $x^0\in argmin\{\Phi(x):x\in\mathcal{X}\}\ \Rightarrow\ x^0\in\mathcal{X}$ y como $x^*\in argmin\{f(x):x\in\mathcal{X}\}\ \Rightarrow\ x^*\in\mathcal{X}$.

De esta forma:

$$[\Phi(x^*)-\Phi(x^0)]\leq\sup_{x,y\in\mathcal{X}}[\Phi(x)-\Phi(y)]=D_\Phi(\mathcal{X})$$

$\Rightarrow$

$$\boxed{[\Phi(x^*)-\Phi(x^0)]\leq D_\Phi(\mathcal{X})\quad(8)}$$

Ahora usando (7) y (8) llegamos a que:

$$f\left(\frac{1}{T}\sum_{t=1}^{T}x^t\right)-f^*\leq\frac{\mu}{T}[\Phi(x^*)-\Phi(x^0)]\overset{Por\ (8)}{\leq}\frac{\mu\cdot D_\Phi(\mathcal{X})}{T}$$

$\Rightarrow$

$$\boxed{f\left(\frac{1}{T}\sum_{t=1}^{T}x^t\right)-f^*\leq\frac{\mu\cdot D_\Phi(\mathcal{X})}{T}\quad(9)}$$

Y esto es justamente lo que nos pedían demostrar.

Pregunta 3

Primero que todo notemos que el algoritmo del mirror descense nos dice que si queremos resolver el problema mín$\{f(x) : x \in \mathcal{X}\}$, debemos aplicar los siguientes pasos:

1. $x^0 \in argmin\{\Phi(x) : x \in \mathcal{X}\}$

2. For $t = 0, ..., T-1$:

$g(x^t) \in \partial f(x^t)$

$x^{t+1} = argmin_{y \in \mathcal{X}}\{f(x^t) + \langle g(x^t), y - x^t\rangle + \frac{1}{\eta_t} V_{x^t}(y)\}$

End for.

3. Return $\bar{x} = \dfrac{\sum_{t=0}^{T-1} \eta_t \cdot x^t}{\sum_{t=0}^{T-1} \eta_t}$

Ahora comenzaremos a ver cada uno de los casos.

i. Lo primero que haremos será encontrar el subgradiente de $F(x)$.

$$\frac{\partial F(x)}{\partial x_j} = \frac{\partial}{\partial x_j}\left(\frac{1}{n}\sum_{i=1}^{n}|\langle a_i, x\rangle - b_i|\right) = \frac{1}{n}\cdot\frac{\partial}{\partial x_j}\left(\sum_{i=1}^{n}|\langle a_i, x\rangle - b_i|\right) = \frac{1}{n}\cdot\sum_{i=1}^{n}\frac{\partial}{\partial x_j}\left(|\langle a_i, x\rangle - b_i|\right)$$

$$= \frac{1}{n}\cdot\sum_{i=1}^{n} sign(\langle a_i, x\rangle - b_i)\cdot\frac{\partial}{\partial x_j}\left(\langle a_i, x\rangle - b_i\right) = \frac{1}{n}\cdot\sum_{i=1}^{n} sign(\langle a_i, x\rangle - b_i)\cdot\frac{\partial}{\partial x_j}\left(\sum_{k=1}^{d} a_{i,k}\cdot x_k - b_i\right)$$

$$= \frac{1}{n}\cdot\sum_{i=1}^{n} sign(\langle a_i, x\rangle - b_i)\cdot\frac{\partial}{\partial x_j}\left(\sum_{k=1}^{d} a_{i,k}\cdot x_k\right) = \frac{1}{n}\cdot\sum_{i=1}^{n} sign(\langle a_i, x\rangle - b_i)\cdot\frac{\partial}{\partial x_j}\left(\sum_{k=1:k\neq j}^{d} a_{i,k}\cdot x_k + a_{i,j}\cdot x_j\right)$$

$$= \frac{1}{n}\cdot\sum_{i=1}^{n} sign(\langle a_i, x\rangle - b_i)\cdot a_{i,j}$$

$\Rightarrow$

$$\nabla F(x) = \begin{pmatrix} \frac{\partial F(x)}{\partial x_1} \\ \vdots \\ \frac{\partial F(x)}{\partial x_d} \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\cdot\sum_{i=1}^{n} sign(\langle a_i, x\rangle - b_i)\cdot a_{i,1} \\ \vdots \\ \frac{1}{n}\cdot\sum_{i=1}^{n} sign(\langle a_i, x\rangle - b_i)\cdot a_{i,d} \end{pmatrix} = \frac{1}{n}\cdot\sum_{i=1}^{n}\begin{pmatrix} sign(\langle a_i, x\rangle - b_i)\cdot a_{i,1} \\ \vdots \\ sign(\langle a_i, x\rangle - b_i)\cdot a_{i,d} \end{pmatrix}$$

$$= \frac{1}{n} \cdot \sum_{i=1}^{n} sign(\langle a_i, x\rangle - b_i) \cdot \begin{pmatrix} a_{i,1} \\ \vdots \\ a_{i,d} \end{pmatrix} = \frac{1}{n} \cdot \sum_{i=1}^{n} sign(\langle a_i, x\rangle - b_i) \cdot a_i$$

De esta forma, tenemos que:

$$\boxed{g(x) = \frac{1}{n} \cdot \sum_{i=1}^{n} sign(\langle a_i, x\rangle - b_i) \cdot a_i \quad (1)}$$

Este resultado es muy importante y se utilizará en los otros algortimos también.

Por otro lado notemos que como estamos hablando del método del subgradiente proyectado, entonces estamos utilizando mirror descense con $\Phi(x) = \frac{\|x\|_{\mathbf{E}}^2}{2}$.

De este modo, como se debe cumplir que $x^0 \in argmin\{\Phi(x) : \|x\| \leq 1\}$, entonces $\boxed{x^0 = 0 \quad (2)}$

Por otro lado notemos que cuando usamos $\Phi(x) = \frac{\|x\|_{\mathbf{E}}^2}{2}$ se nos da que:

$$V_x(y) = \Phi(y) - \Phi(x) - \langle \nabla\Phi(x), y - x\rangle = \frac{\|y\|_{\mathbf{E}}^2}{2} - \frac{\|x\|_{\mathbf{E}}^2}{2} - \langle x, y - x\rangle = \frac{\|y\|_{\mathbf{E}}^2}{2} - \frac{\|x\|_{\mathbf{E}}^2}{2} - \langle x, y\rangle + \langle x, x\rangle$$

$$= \frac{\|y\|_{\mathbf{E}}^2}{2} - \frac{\|x\|_{\mathbf{E}}^2}{2} - \langle x, y\rangle + \|x\|_{\mathbf{E}}^2 = \frac{\|y\|_{\mathbf{E}}^2}{2} + \frac{\|x\|_{\mathbf{E}}^2}{2} - \langle x, y\rangle = \frac{1}{2} \cdot \left(\|x\|_{\mathbf{E}}^2 - 2\langle x, y\rangle + \|y\|_{\mathbf{E}}^2\right)$$

$$= \frac{1}{2}\|x - y\|_{\mathbf{E}}^2$$

$\Rightarrow$

$$V_x(y) = \frac{1}{2}\|x - y\|_{\mathbf{E}}^2$$

De este modo, tenemos que:

$$x^{t+1} = argmin_{y\in\mathcal{X}}\{f(x^t) + \langle g(x^t), y - x^t\rangle + \frac{1}{\eta_t}V_{x^t}(y)\} = argmin_{y\in\mathcal{X}}\{\langle g(x^t), y - x^t\rangle + \frac{1}{\eta_t}V_{x^t}(y)\}$$

$$= argmin_{y\in\mathcal{X}}\{2\eta_t \cdot \langle g(x^t), y - x^t\rangle + 2 \cdot V_{x^t}(y)\} = argmin_{y\in\mathcal{X}}\{\langle 2\eta_t \cdot g(x^t), y - x^t\rangle + 2 \cdot \frac{\|y - x^t\|_{\mathbf{E}}^2}{2}\}$$

$$= argmin_{y\in\mathcal{X}}\{\langle 2\eta_t \cdot g(x^t), y - x^t\rangle + \|y - x^t\|_{\mathbf{E}}^2\} = argmin_{y\in\mathcal{X}}\{\|\eta_t \cdot g(x^t)\|_{\mathbf{E}}^2 + \langle 2\eta_t \cdot g(x^t), y - x^t\rangle + \|y - x^t\|_{\mathbf{E}}^2\}$$

$$= argmin_{y\in\mathcal{X}}\{\|\eta_t \cdot g(x^t) + y - x^t\|_{\mathbf{E}}^2\} = argmin_{y\in\mathcal{X}}\{\|y - [x^t - \eta_t \cdot g(x^t)]\|_{\mathbf{E}}^2\}$$

$$= \Pi_{\mathcal{X}}(x^t - \eta_t \cdot g(x^t))$$

$$\boxed{x^{t+1} = \Pi_{\mathcal{X}}(x^t - \eta_t \cdot g(x^t)) \quad (3)}$$

Ahora tenemos que ver como podemos proyectar sobre $\mathcal{X} = \{x \in E : \|x\|_1 \leq R\}$, es decir, tenemos que resolver el siguiente problema:

$$\Pi_{\mathcal{X}}(\bar{x}) = min\{\|x - \overline{x}\|_2^2 : \|x\|_1 \leq R\}$$

Que es lo mismo a que:

$$min\{\sum_{i=1}^{d}[x_i - \overline{x_i}]^2 : \sum_{i=1}^{d}|x_i| \leq R\}$$

Si hacemos la sustitución $x_i = signo(x_i) \cdot t_i$ con $t_i \geq 0 \ \forall i \in \{1, ..., d\}$ entonces nos queda que:

$|x_i| = |signo(x_i) \cdot t_i| = |signo(x_i)| \cdot |t_i| = t_i$

Así la restricción queda como:

$$\sum_{i=1}^{d} t_i \leq R$$

Ademas, notemos que la función objetivo queda como sigue:

$$\sum_{i=1}^{d}[x_i - \overline{x_i}]^2 = \sum_{i=1}^{d}[signo(x_i) \cdot t_i - \overline{x_i}]^2 = \sum_{i=1}^{d}[signo(x_i) \cdot t_i - \overline{x_i}]^2$$

$$= \sum_{i=1}^{d}([signo(x_i)]^2 \cdot {t_i}^2 - 2signo(x_i) \cdot t_i \cdot \overline{x_i} + [\overline{x_i}]^2)$$

$$= \sum_{i=1}^{d}({t_i}^2 - 2signo(x_i) \cdot t_i \cdot \overline{x_i} + [\overline{x_i}]^2)$$

Notemos que elegimos $signo(x_i) = signo(\overline{x_i})$ para así minimizar, pues es el valor que hace que la expresión del medio siempre sea negativa. De este modo, nos queda:

$$= \sum_{i=1}^{d}({t_i}^2 - 2 \cdot t_i \cdot |\overline{x_i}| + [|\overline{x_i}|]^2) = \sum_{i=1}^{d}(t_i - |\overline{x_i}|)^2$$

Así el problema es equivalente a:

$$min\{\sum_{i=1}^{d}(t_i - |\overline{x_i}|)^2 : \sum_{i=1}^{d} t_i \leq R, t \geq 0\}$$

Planteamos el lagrangeano:

$$L(\lambda, t) = \sum_{i=1}^{d}(t_i - |\overline{x_i}|)^2 + \lambda \left(\sum_{i=1}^{d} t_i - R\right)$$

Ahora calculamos la derivada parcial respecto a t_k y nos queda que:

$$\frac{\partial L(\lambda, t)}{\partial t_k} = 2 \cdot (t_k - |\overline{x_k}|) + \lambda$$

De esta forma en el optimo:

$$2 \cdot (t_k - |\overline{x_k}|) + \lambda = 0 \;\Rightarrow\; 2 \cdot (t_k - |\overline{x_k}|) = -\lambda \;\Rightarrow\; t_k - |\overline{x_k}| = -\frac{\lambda}{2} \;\Rightarrow\; t_k = |\overline{x_k}| - \frac{\lambda}{2}$$

Pero como $t \geq 0$, entonces:

$$t_k(\lambda) = max\{|\overline{x_k}| - \frac{\lambda}{2}, 0\}$$

Y además como en el optimo se da que:

$$\sum_{i=1}^{d} t_i = R \;\Rightarrow\; \sum_{i=1}^{d} max\{|\overline{x_i}| - \frac{\lambda}{2}, 0\} = R$$

Así para encontrar el optimo debemos encontrar $\lambda > 0$ tal que:

$$\sum_{i=1}^{d} max\{|\overline{x_i}| - \frac{\lambda}{2}, 0\} - R = 0$$

Ahora definiremos:

$$g(\lambda) = \sum_{i=1}^{d} max\{|\overline{x_i}| - \frac{\lambda}{2}, 0\} - R$$

Y lo que uno esta buscando es que:

$$g(\lambda) = 0$$

Ahora notemos que $g(\lambda)$ es decreciente por ende, esto lo resolveremos por bisección.

Como sabemos que debe existir al menos un valor de t_i que sea positivo, entonces:

$$max_{i=1,\ldots,d}\left(|\overline{x_i}| - \frac{\lambda}{2}\right) > 0$$

$\Rightarrow$

$$max_{i=1,\ldots,d}\left(|\overline{x_i}|\right) > \frac{\lambda}{2}$$

$\Rightarrow$

$$2 \cdot max_{i=1,\dots,d}\left(|\overline{x_i}|\right) > \lambda$$

Así, nuestro método de bisección lo haremos en el intervalo:

$$[0, 2 \cdot max_{i=1,\dots,d}\left(|\overline{x_i}|\right)]$$

Y de esta manera, programaremos la proyección.

Por otro lado, queremos demostrar que $F(x)$ es Lipschitz, y para esto partiremos utilizando que:

$$|x+y| \leq |x| + |y| \;\Rightarrow\; |x+y| - |x| \leq |y|$$

Y si tomamos $u = x + y$, entonces tenemos que $|u| - |x| \leq |u - x|$

$$\boxed{|u| - |x| \leq |u - x| \quad (4)}$$

Ahora:

$$F(y) - F(x) = \frac{1}{n}\sum_{i=1}^{n}|\langle a_i, y\rangle - b_i| - \frac{1}{n}\sum_{i=1}^{n}|\langle a_i, x\rangle - b_i| = \frac{1}{n}\sum_{i=1}^{n}[|\langle a_i, y\rangle - b_i| - |\langle a_i, x\rangle - b_i|]$$

$$\overset{Por\ (4)}{\leq} \frac{1}{n}\sum_{i=1}^{n}|\langle a_i, y\rangle - b_i - \langle a_i, x\rangle + b_i| = \frac{1}{n}\sum_{i=1}^{n}|\langle a_i, y\rangle - \langle a_i, x\rangle| = \frac{1}{n}\sum_{i=1}^{n}|\langle a_i, y - x\rangle|$$

$$\overset{Cauchy-Schwarz}{\leq} \frac{1}{n}\sum_{i=1}^{n}\|y - x\|_{\mathbf{E}}\|a_i\|_{\mathbf{E}} = \|y - x\|_{\mathbf{E}}\left(\frac{1}{n}\sum_{i=1}^{n}\|a_i\|_{\mathbf{E}}\right)$$

$\Rightarrow$

$$F(x) - F(y) \leq \|y - x\|_{\mathbf{E}}\left(\frac{1}{n}\sum_{i=1}^{n}\|a_i\|_{\mathbf{E}}\right)$$

$\Rightarrow$

$$|F(x) - F(y)| \leq \left(\frac{1}{n}\sum_{i=1}^{n}\|a_i\|_{\mathbf{E}}\right) \cdot \|y - x\|_{\mathbf{E}}\ \forall x, y \in E$$

De este modo:

$$\boxed{F \in \mathcal{F}^0_{\|\cdot\|_{\mathbf{E}},\mathbf{E}}(L) \text{ con } L = \frac{1}{n}\sum_{i=1}^{n}\|a_i\|_{\mathbf{E}} \quad (5)}$$

Gracias a esto, de clases sabemos que es posible elegir $\eta_t = \eta = \sqrt{\frac{2V_{x^0}(x^*)}{T \cdot L^2}}$ y no solo en este algortimo sino que también en el algoritmo **ii** y el algoritmo **iii**.

Ahora notemos que:

$$V_{x^0}(x^*) = \frac{\|x^*\|_{\mathbf{E}}^2}{2} - \frac{\|x^0\|_{\mathbf{E}}^2}{2} - \langle \nabla\Phi(x^0), x^* - x^0\rangle = \frac{\|x^*\|_{\mathbf{E}}^2}{2} - \frac{\|0\|_{\mathbf{E}}^2}{2} - \langle x^0, x^* - 0\rangle = \frac{\|x^*\|_{\mathbf{E}}^2}{2} - \frac{\|0\|_{\mathbf{E}}^2}{2} - \langle 0, x^* - 0\rangle = \frac{\|x^*\|_{\mathbf{E}}^2}{2}$$

De esta forma tenemos que:

$$\eta_t = \eta = \sqrt{\frac{2 \cdot \frac{\|x^*\|_{\mathbf{E}}^2}{2}}{T \cdot L^2}} = \sqrt{\frac{\|x^*\|_{\mathbf{E}}^2}{T \cdot L^2}} = \frac{\|x^*\|_{\mathbf{E}}}{L} \cdot \frac{1}{\sqrt{T}}$$

$\Rightarrow$

$$\eta_t = \eta = \frac{\|x^*\|_{\mathbf{E}}}{L} \cdot \frac{1}{\sqrt{T}}$$

Ahora notemos que como $\mathcal{X} = \{x \in E : \|x\|_1 \leq R\}$ y como también es claro que:

$$B_{\|\cdot\|_1}[0, R] \subseteq B_{\|\cdot\|_{\mathbf{E}}}[0, R]$$

$\Rightarrow$

$$\boxed{\eta_t = \eta = \frac{R}{L} \cdot \frac{1}{\sqrt{T}} \quad (6)}$$

Así finalmente el algoritmo queda como sigue:

1. $x^0 = 0$

2. For $t = 0, ..., T-1$:

$$g(x^t) = \frac{1}{n} \cdot \sum_{i=1}^{n} sign(\langle a_i, x^t\rangle - b_i) \cdot a_i$$

$$L = \frac{1}{n} \sum_{i=1}^{n} \|a_i\|_{\mathbf{E}}$$

$$\eta_t = \frac{R}{L} \cdot \frac{1}{\sqrt{T}}$$

$$x^{t+1} = \Pi_X(x^t - \eta_t \cdot g(x^t))$$

3. Return:

$$\bar{x} = \frac{\sum_{t=0}^{T-1} \eta_t \cdot x^t}{\sum_{t=0}^{T-1} \eta_t} = \frac{\sum_{t=0}^{T-1} \eta \cdot x^t}{\sum_{t=0}^{T-1} \eta} = \frac{\eta \cdot \sum_{t=0}^{T-1} x^t}{\eta \cdot T} = \frac{\sum_{t=0}^{T-1} x^t}{T}$$

Y este sería el algoritmo del gradiente proyectado para la regresión robusta.

---

ii. Lo primero que debemos notar es que aquí utilizaremos a:

$$\Phi(x) = \frac{e \cdot ln(d)}{p(d)} \sum_{k=1}^{d} |x_k|^{p(d)}$$

Ahora notemos que en el método de mirror descense debemos resolver el siguiente problema:

$$x^{t+1} = argmin_{y \in \mathcal{X}} \{ f(x^t) + \langle g(x^t), y - x^t \rangle + \frac{1}{\eta_t} V_{x^t}(y) \} = argmin_{y \in \mathcal{X}} \{ \langle g(x^t), y - x^t \rangle + \frac{1}{\eta_t} V_{x^t}(y) \}$$

$$= argmin_{y \in \mathcal{X}} \{ \langle g(x^t), y \rangle + \frac{1}{\eta_t} V_{x^t}(y) \} = argmin_{y \in \mathcal{X}} \{ \langle g(x^t), y \rangle + \frac{1}{\eta_t} \cdot \big[\Phi(y) - \Phi(x^t) - \langle \nabla \Phi(x^t), y - x^t \rangle \big] \}$$

$$= argmin_{y \in \mathcal{X}} \{ \langle g(x^t), y \rangle + \frac{1}{\eta_t} \cdot \big[\Phi(y) - \langle \nabla \Phi(x^t), y \rangle \big] \} = argmin_{y \in \mathcal{X}} \{ \langle g(x^t), y \rangle - \frac{1}{\eta_t} \cdot \langle \nabla \Phi(x^t), y \rangle + \frac{1}{\eta_t} \cdot \Phi(y) \}$$

$$= argmin_{y \in \mathcal{X}} \{ \langle g(x^t) - \frac{1}{\eta_t} \cdot \nabla \Phi(x^t), y \rangle + \frac{1}{\eta_t} \cdot \Phi(y) \}$$

$\Rightarrow$

$$\boxed{x^{t+1} = argmin_{y \in \mathcal{X}} \{ \langle g(x^t) - \frac{1}{\eta_t} \cdot \nabla \Phi(x^t), y \rangle + \frac{1}{\eta_t} \cdot \Phi(y) \} \quad (1)}$$

Ahora lo que haremos será calcular $\nabla \Phi(\cdot)$ para así evaluarlo en x^t.

Y para esto lo que debemos calcular son las derivadas parciales.

$$\frac{\partial \Phi(x)}{\partial x_j} = \frac{\partial}{\partial x_j} \left(\frac{e \cdot ln(d)}{p(d)} \sum_{k=1}^{d} |x_k|^{p(d)} \right) = \frac{e \cdot ln(d)}{p(d)} \cdot \frac{\partial}{\partial x_j} \left(\sum_{k=1}^{d} |x_k|^{p(d)} \right)$$

$$= \frac{e \cdot ln(d)}{p(d)} \cdot \frac{\partial}{\partial x_j} \left(\sum_{k=1:k \neq j}^{d} |x_k|^{p(d)} + |x_j|^{p(d)} \right) = \frac{e \cdot ln(d)}{p(d)} \cdot \left(\frac{\partial}{\partial x_j} \left[\sum_{k=1:k \neq j}^{d} |x_k|^{p(d)} \right] + \frac{\partial}{\partial x_j} \left[|x_j|^{p(d)} \right] \right)$$

$$= \frac{e \cdot ln(d)}{p(d)} \cdot \left(\frac{\partial}{\partial x_j} \left[|x_j|^{p(d)} \right] \right) = \frac{e \cdot ln(d)}{p(d)} \cdot p(d) \cdot |x_j|^{p(d)-1} \cdot sign(x_j) = e \cdot ln(d) \cdot |x_j|^{\frac{1}{ln(d)}} \cdot sign(x_j)$$

$\Rightarrow$

$$\boxed{[\nabla\Phi(x)]_j = \frac{\partial\Phi(x)}{\partial x_j} = e \cdot ln(d) \cdot |x_j|^{\frac{1}{ln(d)}} \cdot sign(x_j) \quad \forall j \in \{1, ..., d\} \quad (2)}$$

Como ya calculamos $\nabla\Phi(x)$ entonces podemos resolver (1) asumiendo conocido el valor de $\nabla\Phi(x^t)$.

Ahora nos centramos en resolver a:

$$x^{t+1} = argmin_{\sum_{j=1}^{d}|y_j|\leq R}\{\langle g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t), y\rangle + \frac{1}{\eta_t} \cdot \Phi(y)\}$$

De este modo:

$$x^{t+1} = argmin_{\sum_{j=1}^{d}|y_j|\leq R}\{\sum_{j=1}^{d}\left[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t)\right]_j \cdot y_j + \frac{1}{\eta_t} \cdot \frac{e \cdot ln(d)}{p(d)}\sum_{j=1}^{d}|y_j|^{p(d)}\}$$

Ahora realizamos el siguiente cambio de variable: $y_j = signo(y_j) \cdot t_j$ con $t_j \geq 0$ $\forall j \in \{1, ..., d\}$.

De este modo, el problema a resolver nos queda como:

$$argmin_{\sum_{j=1}^{d} t_j\leq R, t_1,...,t_d\geq 0}\{\sum_{j=1}^{d}\left[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t)\right]_j \cdot signo(y_j) \cdot t_j + \frac{1}{\eta_t} \cdot \frac{e \cdot ln(d)}{p(d)}\sum_{j=1}^{d} {t_j}^{p(d)}\}$$

Así notemos que es debe ocurrir que:

$$\boxed{signo(y_j) = -signo\left(\left[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t)\right]_j\right) \quad \forall j \in \{1, ..., d\} \quad (3)}$$

De este modo, el problema a resolver nos queda como:

$$argmin_{\sum_{j=1}^{d} t_j\leq R, t_1,...,t_d\geq 0}\{-\sum_{j=1}^{d}|\left[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t)\right]_j| \cdot t_j + \frac{1}{\eta_t} \cdot \frac{e \cdot ln(d)}{p(d)}\sum_{j=1}^{d} {t_j}^{p(d)}\}$$

Ahora plantearemos el lagrangeano:

$$L(t, \lambda) = -\sum_{j=1}^{d}|\left[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t)\right]_j| \cdot t_j + \frac{1}{\eta_t} \cdot \frac{e \cdot ln(d)}{p(d)}\sum_{j=1}^{d} {t_j}^{p(d)} + \lambda \cdot \left(\sum_{j=1}^{d} t_j - R\right)$$

Ahora lo que sabemos es que la derivada parcial de $L(t, \lambda)$ respecto a t_k:

$$\frac{\partial L(t, \lambda)}{\partial t_k} = -|\left[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t)\right]_k| + \lambda + \frac{1}{\eta_t} \cdot \frac{e \cdot ln(d)}{p(d)} \cdot p(d) \cdot {t_k}^{p(d)-1}$$

$$= -|\left[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t)\right]_k| + \lambda + \frac{1}{\eta_t} \cdot e \cdot ln(d) \cdot t_k{}^{\frac{1}{ln(d)}}$$

Y como esta derivada debe ser igual a 0 debe ocurrir que:

$$\boxed{\frac{1}{\eta_t} \cdot e \cdot ln(d) \cdot t_k{}^{\frac{1}{ln(d)}} = |\left[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t)\right]_k| - \lambda \quad \forall k \in \{1, ..., d\} \quad (4)}$$

Ahora nos debemos poner en 2 casos. El caso en que $\lambda = 0$ y el caso en que $\sum_{j=1} t_j = R \;\; (\lambda > 0)$.

$\underline{\lambda = 0:}$ Usando (4) tenemos que:

$$\frac{1}{\eta_t} \cdot e \cdot ln(d) \cdot t_k{}^{\frac{1}{ln(d)}} = |\left[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t)\right]_k|$$

$\Rightarrow$

$$t_k{}^{\frac{1}{ln(d)}} = |\left[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t)\right]_k| \cdot \frac{\eta_t}{e \cdot ln(d)}$$

$\Rightarrow$

$$\boxed{\bar{t_k} = \left[|\left[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t)\right]_k| \cdot \frac{\eta_t}{e \cdot ln(d)}\right]^{ln(d)} \quad (5)}$$

Notemos que $\bar{t_k} \geq 0 \;\; \forall k \in \{1, ..., d\}$

Y si $\sum_{j=1}^{d} \bar{t_j} \leq R$, entonces el optimo esta acá en este resultado.

Y así tenemos que:

$$\boxed{x_k^{t+1} = -signo\left(\left[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t)\right]_k\right) \cdot \left[|\left[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t)\right]_k| \cdot \frac{\eta_t}{e \cdot ln(d)}\right]^{ln(d)} \quad (6)}$$

$\underline{\lambda > 0:}$ Notemos que por holguras complementarias, esto implica que:

$$\boxed{\sum_{k=1}^{d} t_k = R \;\; (7)}$$

Ahora notemos que si $|\left[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t)\right]_k| - \lambda \geq 0$ tenemos por (4) que:

$$\bar{t_k} = \left[\left(|\left[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t)\right]_k| - \lambda\right) \cdot \frac{\eta_t}{e \cdot ln(d)}\right]^{ln(d)}$$

Ahora si $|\Big[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t)\Big]_k| - \lambda < 0$ y recordando que:

$$\boxed{\frac{\partial L(t,\lambda)}{\partial t_k} = -|\Big[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t)\Big]_k| + \lambda + \frac{1}{\eta_t} \cdot e \cdot ln(d) \cdot t_k{}^{\frac{1}{ln(d)}} \quad (8)}$$

(resultado justo antes de la ecuación (4))

Así de este modo tenemos que para todo $t_k \geq 0$ por (8) se cumple que:

$$\frac{\partial L(t,\lambda)}{\partial t_k}(t_k) \geq -|\Big[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t)\Big]_k| + \lambda \overset{Como\ |\Big[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t)\Big]_k| - \lambda < 0}{> 0}$$

$\Rightarrow$

$$\boxed{\frac{\partial L(t,\lambda)}{\partial t_k}(t_k) \geq 0 \quad \forall t_k \geq 0 \quad (9)}$$

Y considerando (4) y (9) tenemos que si $|\Big[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t)\Big]_k| - \lambda < 0$ entonces $\boxed{\bar{t_k} = 0 \quad (10)}$

De esta forma notemos que para el caso de $\lambda > 0$ tenemos que:

$$\bar{t_k}(\lambda) = \begin{cases} \Big[\Big(|\Big[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t)\Big]_k| - \lambda\Big) \cdot \frac{\eta_t}{e \cdot ln(d)}\Big]^{ln(d)} & \text{si } |\Big[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t)\Big]_k| - \lambda \geq 0 \\ 0 & \text{si } |\Big[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t)\Big]_k| - \lambda < 0 \end{cases}$$

Así podemos decir que el resultado es:

$$\boxed{\bar{t_k}(\lambda) = \Big[\Big(max\{|\Big[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t)\Big]_k| - \lambda, 0\}\Big) \cdot \frac{\eta_t}{e \cdot ln(d)}\Big]^{ln(d)} \quad (11)}$$

Ahora notemos que por (7) debe ocurrir que:

$$\boxed{\sum_{k=1}^{d} \bar{t_k}(\lambda) = R \quad (12)}$$

Así debemos encontrar ese λ y lo haremos a través de busqueda binaria.

De este modo, definiremos $h(\lambda)$ de la siguiente manera:

$$h(\lambda) = \sum_{k=1}^{d} \Big[\Big(max\{|\Big[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t)\Big]_k| - \lambda, 0\}\Big) \cdot \frac{\eta_t}{e \cdot ln(d)}\Big]^{ln(d)} - R$$

Y lo que buscamos es un λ^* tal que $h(\lambda^*) = 0$.

Así notemos que:

$$h(0) = \sum_{k=1}^{d} \left[\left(max\{ | \left[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t) \right]_k |, 0\} \right) \cdot \frac{\eta_t}{e \cdot ln(d)} \right]^{ln(d)} - R$$

$$= \sum_{k=1}^{d} \left[\left(| \left[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t) \right]_k | \right) \cdot \frac{\eta_t}{e \cdot ln(d)} \right]^{ln(d)} - R > 0$$

Y por otro lado, si definimos $\lambda' = \text{máx}_{i=1,...d}\{ | \left[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t) \right]_i | \}$

Tenemos que:

$$h(\lambda') = -R < 0$$

Así, tenemos que la busqueda binaria se debe realizar en el intervalo:

$$\boxed{\left[0, \max_{i=1,...d} \{ | \left[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t) \right]_i | \} \right] \quad (13)}$$

Ahora queremos demostrar que $h(\lambda)$ es monotona, de hecho, más en especifico queremos demostrar que es no creciente y para esto es suficiente demostrar que $\bar{t_k}(\lambda)$ es no creciente $\forall k \in \{1, ..., d\}$.

Ahora recordando el valor $\bar{t_k}(\lambda)$ tenemos que:

$$\bar{t_k}(\lambda) = \left[\left(max\{ | \left[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t) \right]_k | - \lambda, 0\} \right) \cdot \frac{\eta_t}{e \cdot ln(d)} \right]^{ln(d)}$$

Es claro que para el intervalo $\left[| \left[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t) \right]_k |, +\infty \right)$ ocurre que $\bar{t_k}(\lambda) = 0$ por ende en ese intervalo la función es no creciente

Ahora tomemos un λ_1 arbitrario tal que $0 \leq \lambda_1 < | \left[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t) \right]_k |$ y sea $\lambda_2 > \lambda_1$ entonces tenemos dos posibilidades:

i. $\lambda_2 \geq | \left[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t) \right]_k |$**:**

De esta forma $\bar{t_k}(\lambda_2) = 0$ y también tenemos que $\bar{t_k}(\lambda_1) > 0$ y así tenemos que:

$$\bar{t_k}(\lambda_1) > \bar{t_k}(\lambda_2)$$

ii. $0 \leq \lambda_2 < | \left[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t) \right]_k |$**:**

De esta forma:

$$-\lambda_2 < -\lambda_1$$

$\Rightarrow$

$$|\left[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t)\right]_k| - \lambda_2 < |\left[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t)\right]_k| - \lambda_1$$

$\Rightarrow$

$$\text{máx}\{|\left[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t)\right]_k| - \lambda_2, 0\} < \text{máx}\{|\left[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t)\right]_k| - \lambda_1, 0\}$$

$\Rightarrow$

$$\frac{\eta_t}{e \cdot ln(d)} \cdot \left(\text{máx}\{|\left[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t)\right]_k| - \lambda_2, 0\}\right) < \frac{\eta_t}{e \cdot ln(d)} \cdot \left(\text{máx}\{|\left[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t)\right]_k| - \lambda_1, 0\}\right)$$

$\Rightarrow$

$$\left[\frac{\eta_t}{e \cdot ln(d)} \cdot \left(\text{máx}\{|\left[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t)\right]_k| - \lambda_2, 0\}\right)\right]^{ln(d)}$$
$$< \left[\frac{\eta_t}{e \cdot ln(d)} \cdot \left(\text{máx}\{|\left[g(x^t) - \frac{1}{\eta_t} \cdot \nabla\Phi(x^t)\right]_k| - \lambda_1, 0\}\right)\right]^{ln(d)}$$

$\Rightarrow$

$$\bar{t_k}(\lambda_2) < \bar{t_k}(\lambda_1)$$

Así notamos que en todos los casos tenemos que $\bar{t_k}(\lambda)$ es una función no creciente $\forall k \in \{1, ..., d\}$

De esta forma $h(\lambda)$ es una función no creciente, por ende h es monotona y así efectivamente podemos usar el método de la bisección.

Ahora debemos encontrar el paso que utilizaremos para este algoritmo y nosotros sabemos que (resultado del analisis del algoritmo **i**):

$$\eta_t = \eta = \sqrt{\frac{2 \cdot V_{x^0}(x^*)}{T \cdot L^2}}$$

Ahora debemos encontrar el valor de $V_{x^0}(x^*)$ y para esto debemos utilizar el valor de $\Phi(x)$, de este modo:

$$V_{x^0}(x^*) = \Phi(x^*) - \Phi(x^0) - \langle \nabla\Phi(x^0), x^* - x^0 \rangle$$

Y como $x^0 = 0 \ \Rightarrow \ \Phi(x^0) = 0 \wedge \nabla\Phi(x^0) = 0$

De esta forma:

$$V_{x^0}(x^*) = \Phi(x^*) = \frac{e \cdot ln(d)}{p(d)} \cdot \sum_{k=1}^{d} |x_k^*|^{p(d)} = \frac{e \cdot ln(d)}{p(d)} \cdot \|x^*\|_{p(d)}^{p(d)} = \frac{e \cdot ln(d)}{1 + \frac{1}{ln(d)}} \cdot [\|x^*\|_{p(d)}]^{\left(1+\frac{1}{ln(d)}\right)}$$

Así tenemos que el paso a utilizar es:

$$\eta_t = \eta = \frac{\sqrt{2}}{\sqrt{T} \cdot L} \cdot \sqrt{\frac{e \cdot ln(d)}{\left(1 + \frac{1}{ln(d)}\right)} \cdot R^{\left(1+\frac{1}{ln(d)}\right)}}$$

Ahora el algoritmo queda como sigue:

1. $x^0 = 0$

2. For $t = 0, ..., T-1$:

$$g(x^t) = \frac{1}{n} \cdot \sum_{i=1}^{n} sign(\langle a_i, x^t \rangle - b_i) \cdot a_i$$

$$[\nabla \Phi(x^t)]_j = e \cdot ln(d) \cdot |x_j^t|^{\frac{1}{ln(d)}} \cdot sign(x_j^t) \quad \forall j \in \{1, ..., d\}$$

$$L = \frac{1}{n} \sum_{i=1}^{n} \|a_i\|_{\mathbf{E}}$$

$$\eta_t = \eta = \frac{\sqrt{2}}{\sqrt{T} \cdot L} \cdot \sqrt{\frac{e \cdot ln(d)}{\left(1 + \frac{1}{ln(d)}\right)} \cdot R^{\left(1+\frac{1}{ln(d)}\right)}}$$

Lo primero que haremos será decir que:

$$x_k^{t+1} = -signo\left(\left[g(x^t) - \frac{1}{\eta_t} \cdot \nabla \Phi(x^t)\right]_k\right) \cdot \left[|\left[g(x^t) - \frac{1}{\eta_t} \cdot \nabla \Phi(x^t)\right]_k| \cdot \frac{\eta_t}{e \cdot ln(d)}\right]^{ln(d)}$$

$\forall k \in \{1, ..., d\}$

Si $\|x^{t+1}\|_1 \leq R$ entonces pasamos a la siguiente iteración. De lo contrario, definimos:

$$h(\lambda) = \sum_{k=1}^{d} \left[\left(max\{|\left[g(x^t) - \frac{1}{\eta_t} \cdot \nabla \Phi(x^t)\right]_k| - \lambda, 0\}\right) \cdot \frac{\eta_t}{e \cdot ln(d)}\right]^{ln(d)} - R$$

Notando que la función es no creciente, entonces usaremos método de la bisección

para encontrar λ tal que $h(\lambda) = 0$

Esa busqueda se realizará en el intervalo:

$$\left[0, \max_{i=1,...d}\{|\left[g(x^t) - \frac{1}{\eta_t}\cdot\nabla\Phi(x^t)\right]_i|\}\right]$$

$$x_k^{t+1} = -signo\left(\left[g(x^t) - \frac{1}{\eta_t}\cdot\nabla\Phi(x^t)\right]_k\right)\cdot\left[\left(max\{|\left[g(x^t) - \frac{1}{\eta_t}\cdot\nabla\Phi(x^t)\right]_k| - \lambda, 0\}\right)\cdot\frac{\eta_t}{e\cdot ln(d)}\right]^{ln(d)}$$

$\forall k \in \{1, ..., d\}$

y pasamos a la siguiente iteración.

3. Return (como los η_t son constantes para todas las iteraciones):

$$\bar{x} = \frac{\sum_{t=0}^{T-1} x^t}{T}$$

iii. Lo primero que debemos notar que es que definiremos a $y = (y^+, y^-)$

Ahora para el algoritmo de mirror descense se debe cumplir que $y^0 \in argmin(\Phi(x) : x \in \Delta_{2d})$

Y como tenemos que $\Phi(y) = \sum_{j=1}^{2d} y_j \cdot log(y_j)$ entonces para encontrar y^0 debemos encontrar el minimo de $\Phi(y)$ tal que $y \in \Delta_{2d}$ y para esto usaremos multiplicadores de lagrange:

$$L(y, \mu) = \sum_{j=1}^{2d} y_j \cdot log(y_j) + \mu\left(\sum_{j=1}^{2d} y_j - 1\right)$$

Ahora notemos que como el gradiente respecto a y de $L(y, \mu)$ debe ser igual a 0 entonces:

$$\frac{\partial L(y,\mu)}{\partial y_k} = y_k \cdot \frac{1}{y_k} + ln(y_k) + \mu = 0$$

$\Rightarrow$

$$1 + ln(y_k) + \mu = 0 \;\Rightarrow\; ln(y_k) = -\mu - 1 \;\Rightarrow\; e^{-\mu-1} = y_k \;\Rightarrow\; y_k = A$$

Y además como se debe cumplir que:

$$\sum_{j=1}^{2d} y_j = 1 \;\Rightarrow\; \sum_{j=1}^{2d} A = 1 \;\Rightarrow\; A \cdot 2d = 1 \;\Rightarrow\; A = \frac{1}{2d} \;\Rightarrow\; y_k = \frac{1}{2d}$$

De esta forma el y^0 que utilizaremos para este algoritmo será:

$$\boxed{y^0 = \left(\frac{1}{2d}, ..., \frac{1}{2d}\right)^T \quad (1)}$$

Ahora lo que tendremos que hacer es plantear el caso de los pesos multiplicativas para la situación actual.

Lo primero es notar que:

$$H(y^+, y^-) = F(x = y^+ - y^-) = \frac{1}{n} \cdot \sum_{j=1}^{n} |\langle a_j, y^+ - y^- \rangle - b_j| = \frac{1}{n} \cdot \sum_{j=1}^{n} |\langle a_j, y^+ \rangle - \langle a_j, y^- \rangle - b_j|$$

Notemos que:

$$\frac{\partial H(y^+, y^-)}{\partial y_k^+} = \frac{1}{n} \cdot \sum_{j=1}^{n} \frac{\partial}{\partial y_k^+} \left(|\langle a_j, y^+ \rangle - \langle a_j, y^- \rangle - b_j| \right)$$

$$= \frac{1}{n} \cdot \sum_{j=1}^{n} signo(\langle a_j, y^+ \rangle - \langle a_j, y^- \rangle - b_j) \cdot \frac{\partial}{\partial y_k^+} \left(\langle a_j, y^+ \rangle - \langle a_j, y^- \rangle - b_j \right)$$

$$= \frac{1}{n} \cdot \sum_{j=1}^{n} signo(\langle a_j, y^+ \rangle - \langle a_j, y^- \rangle - b_j) \cdot a_{j,k}$$

$\Rightarrow$

$$\frac{\partial H(y^+, y^-)}{\partial y_k^+} = \frac{1}{n} \cdot \sum_{j=1}^{n} signo(\langle a_j, y^+ \rangle - \langle a_j, y^- \rangle - b_j) \cdot a_{j,k}$$

Por otro lado:

$$\frac{\partial H(y^+, y^-)}{\partial y_k^-} = \frac{1}{n} \cdot \sum_{j=1}^{n} \frac{\partial}{\partial y_k^-} \left(|\langle a_j, y^+ \rangle - \langle a_j, y^- \rangle - b_j| \right)$$

$$= \frac{1}{n} \cdot \sum_{j=1}^{n} signo(\langle a_j, y^+ \rangle - \langle a_j, y^- \rangle - b_j) \cdot \frac{\partial}{\partial y_k^-} \left(\langle a_j, y^+ \rangle - \langle a_j, y^- \rangle - b_j \right)$$

$$= \frac{1}{n} \cdot \sum_{j=1}^{n} signo(\langle a_j, y^+ \rangle - \langle a_j, y^- \rangle - b_j) \cdot [- \cdot a_{j,k}]$$

$\Rightarrow$

$$\frac{\partial H(y^+, y^-)}{\partial y_k^-} = -\frac{1}{n} \cdot \sum_{j=1}^{n} signo(\langle a_j, y^+ \rangle - \langle a_j, y^- \rangle - b_j) \cdot a_{j,k}$$

De este modo, denotaremos $\nabla H(y)$ al subgradiente de $H(y^+, y^-)$ por simplicidad de notación.

Así tenemos que:

$$\nabla H(y^+, y^-) = \begin{pmatrix} \frac{\partial H(x)}{\partial y_1^+} \\ \vdots \\ \frac{\partial H(x)}{\partial y_d^+} \\ \frac{\partial H(x)}{\partial y_1^-} \\ \vdots \\ \frac{\partial H(x)}{\partial y_d^-} \end{pmatrix} = \begin{pmatrix} \frac{1}{n} \cdot \sum_{j=1}^{n} signo(\langle a_j, y^+ \rangle - \langle a_j, y^- \rangle - b_j) \cdot a_{j,1} \\ \vdots \\ \frac{1}{n} \cdot \sum_{j=1}^{n} signo(\langle a_j, y^+ \rangle - \langle a_j, y^- \rangle - b_j) \cdot a_{j,d} \\ -\frac{1}{n} \cdot \sum_{j=1}^{n} signo(\langle a_j, y^+ \rangle - \langle a_j, y^- \rangle - b_j) \cdot a_{j,1} \\ \vdots \\ -\frac{1}{n} \cdot \sum_{j=1}^{n} signo(\langle a_j, y^+ \rangle - \langle a_j, y^- \rangle - b_j) \cdot a_{j,d} \end{pmatrix}$$

$$= \frac{1}{n} \cdot \sum_{j=1}^{n} signo(\langle a_j, y^+ \rangle - \langle a_j, y^- \rangle - b_j) \cdot \begin{pmatrix} a_{j,1} \\ \vdots \\ a_{j,d} \\ -a_{j,1} \\ \vdots \\ -a_{j,d} \end{pmatrix} = \frac{1}{n} \cdot \sum_{j=1}^{n} signo(\langle a_j, y^+ \rangle - \langle a_j, y^- \rangle - b_j) \cdot \begin{pmatrix} a_j \\ -a_j \end{pmatrix}$$

$\Rightarrow$

$$\boxed{\nabla H(y^+, y^-) = \frac{1}{n} \cdot \sum_{j=1}^{n} signo(\langle a_j, y^+ \rangle - \langle a_j, y^- \rangle - b_j) \cdot \begin{pmatrix} a_j \\ -a_j \end{pmatrix} \quad (2)}$$

Ahora para esta situación en particular podemos encontrar el valor de L y lo calcularemos usando la propiedad de que la norma de subgradiente siempre deben ser menores que L con L la constante de *Lipschitz*.

Entonces:

$$\|\nabla H(y^+, y^-)\|_2 = \|\frac{1}{n} \cdot \sum_{j=1}^{n} signo(\langle a_j, y^+ \rangle - \langle a_j, y^- \rangle - b_j) \cdot \begin{pmatrix} a_j \\ -a_j \end{pmatrix}\|_2$$

$$\leq \frac{1}{n} \cdot \sum_{j=1}^{n} \|signo(\langle a_j, y^+ \rangle - \langle a_j, y^- \rangle - b_j) \cdot \begin{pmatrix} a_j \\ -a_j \end{pmatrix}\|_2$$

$$= \frac{1}{n} \cdot \sum_{j=1}^{n} \|\begin{pmatrix} a_j \\ -a_j \end{pmatrix}\|_2$$

Por otro lado, notemos que:

$$\|\begin{pmatrix} a_j \\ -a_j \end{pmatrix}\|_2^2 = \sum_{j=1}^{d} {a_j}^2 + \sum_{j=1}^{d} (-a_j)^2 = \sum_{j=1}^{d} {a_j}^2 + \sum_{j=1}^{d} {a_j}^2 = 2 \cdot \sum_{j=1}^{d} {a_j}^2 = 2 \cdot \|a_j\|_2^2$$

$\Rightarrow$

$$\|\begin{pmatrix} a_j \\ -a_j \end{pmatrix}\|_2 = \sqrt{2} \cdot \|a_j\|_2$$

$\Rightarrow$

$$\|\nabla H(y^+, y^-)\|_2 \leq \frac{1}{n} \cdot \sum_{j=1}^{n} \|\begin{pmatrix} a_j \\ -a_j \end{pmatrix}\|_2 \leq \frac{1}{n} \cdot \sum_{j=1}^{n} \left(\sqrt{2} \cdot \|a_j\|_2\right) = \frac{\sqrt{2}}{n} \cdot \sum_{j=1}^{n} \|a_j\|_2$$

$\Rightarrow$

$$\boxed{L_1 = \frac{\sqrt{2}}{n} \cdot \sum_{j=1}^{n} \|a_j\|_2 \quad (3)}$$

Ahora debemos calcular el paso $\eta_t = eta$, entonces tomamos:

$$\boxed{\eta_t = \eta = \frac{R}{L_1 \cdot \sqrt{T}} \quad (4)}$$

De esta forma, el algoritmo queda como sigue:

1. $y^0 = \left(\frac{1}{2d}, ..., \frac{1}{2d}\right)^T$

2. For $t = 0, ..., T-1$:

$$\nabla H(y^{t,+}, y^{t,-}) = \frac{1}{n} \cdot \sum_{j=1}^{n} signo(\langle a_j, y^{t,+} \rangle - \langle a_j, y^{t,-} \rangle - b_j) \cdot \begin{pmatrix} a_j \\ -a_j \end{pmatrix}$$

$$L_1 = \frac{\sqrt{2}}{n} \sum_{i=1}^{n} \|a_i\|_{\mathbf{E}}$$

$$\eta_t = \frac{R}{L} \cdot \frac{1}{\sqrt{T}}$$

$$y_k^{t+1} = \frac{y_k^t \cdot exp\left(-\eta_t \cdot [\nabla H(y^{t,+}, y^{t,-})]_k\right)}{\sum_{j=1}^{2d} y_j^t \cdot exp\left(-\eta_t \cdot [\nabla H(y^{t,+}, y^{t,-})]_j\right)}$$

3. Return:

$$\bar{y} = \frac{\sum_{t=0}^{T-1} y^t}{T}$$

Resultados computacionales

Para todos los algoritmos lo que saldra como output será $f^{best} = \text{mín}_{k=0,...,T}\{f(x^k)\}$ y $f(\bar{x})$.

Y realizaremos todas las comparaciones considerando la forma que nos provea una mejor solución.

Ahora para los diferentes datasets realizaremos las comparación en cuanto a datos computacionales (n), número de iteraciones y cercanía al optimo.

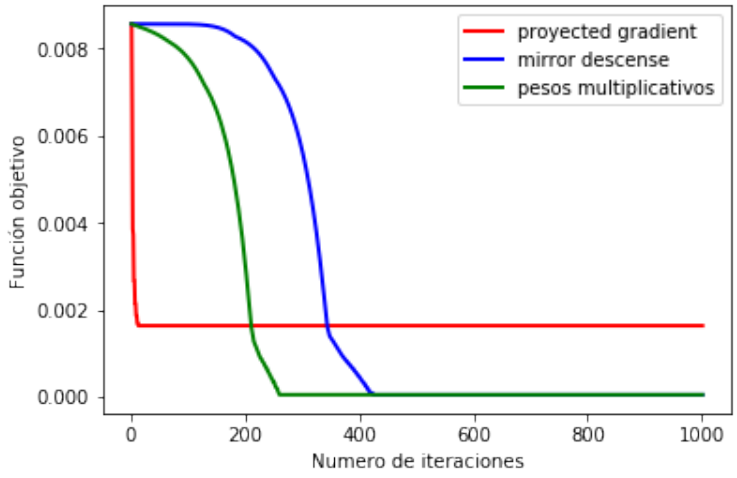
Para comenzar diremos que utilizaremos el termino DSX para referirnos al dataset X con $X = 1, ..., 6$

Por otro lado, notemos que la cantidad de datos para los diferentes datasets son:

$$n_1 = 7, n_2 = 12, n_3 = 22, n_4 = 1000, n_5 = 1000, n_6 = 1000$$

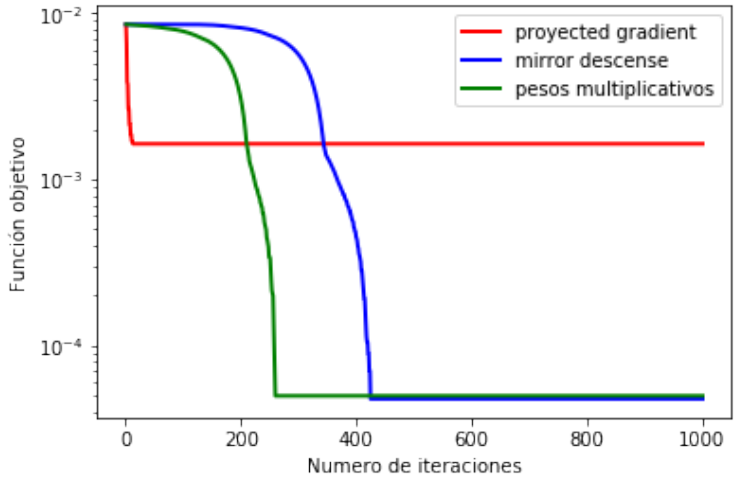
DATASET 1:

i. Función objetivo vs iteraciones (ESCALA NORMAL):



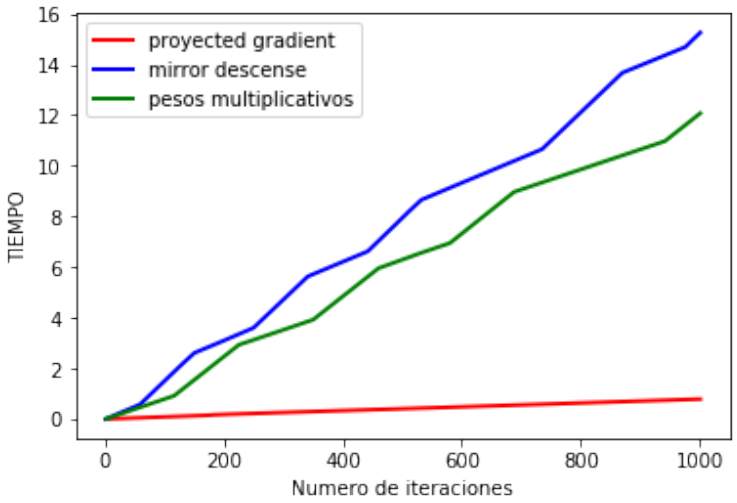
Comentarios: Notemos que si bien el algoritmo de subgradiente proyectado converge rapidamente, luego de la iteración 400 tanto el método de mirror descense como el de los pesos multiplicativos tienen errores menores.

ii. Función objetivo vs iteraciones (ESCALA LOGARITMICA):



Comentarios: Complementando el analisis anterior, gracias a la escala logaritmica se puede observar que el método del gradiente proyectado tiene un error mayor a 10^{-3} mientras que los métodos de los pesos multiplicativos como el de mirror descence tienen erorres menores a 10^{-4} lo que es bastante bueno.

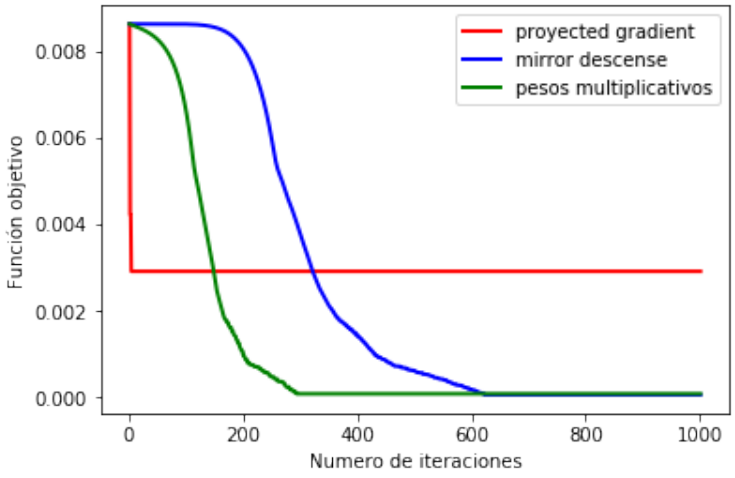
iii. Tiempo de ejecución vs iteraciones:



Comentarios: Como se observa el método del gradiente proyectado tiene tiempos de ejecución mucho menores a diferencia de los métodos de pesos multiplicativos y de mirror descense, los cuales como ya dijimos, poseen errores menores. Así se puede evidanciar un trade-off.

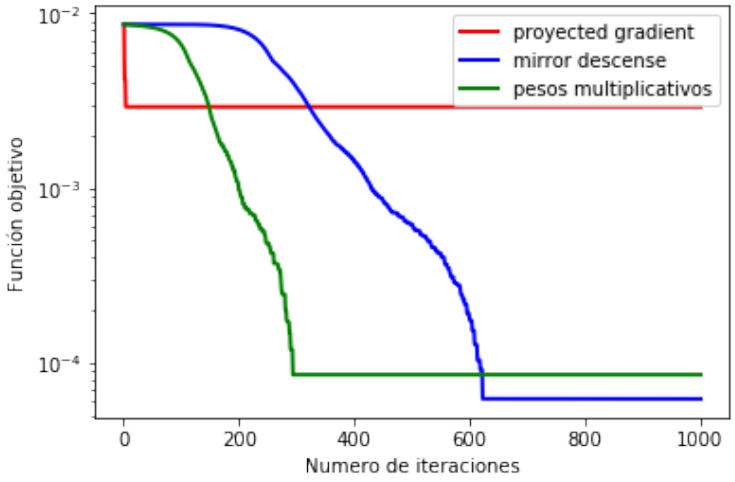
DATASET 2:

i. Función objetivo vs iteraciones (ESCALA NORMAL):



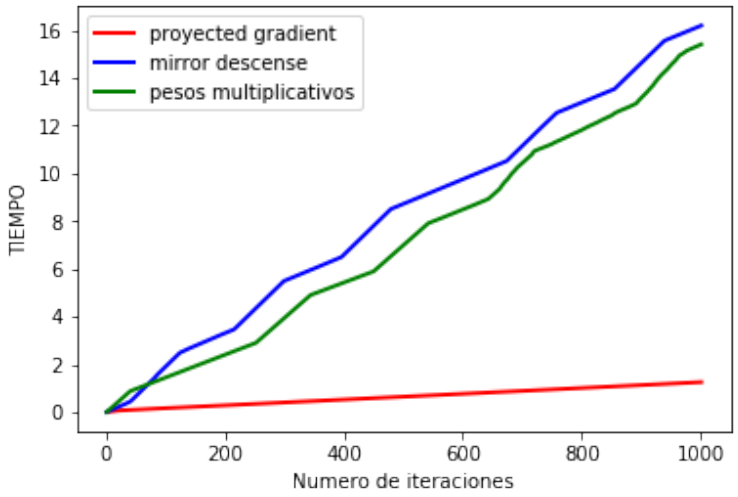
Comentarios: Notemos que si bien el algoritmo de subgradiente proyectado converge rapidamente, luego de la iteración 600 tanto el método de mirror descense como el de los pesos multiplicativos tienen errores menores.

ii. Función objetivo vs iteraciones (ESCALA LOGARITMICA):



Comentarios: Complementando el analisis anterior, gracias a la escala logaritmica se puede observar que el método del gradiente proyectado tiene un error mayor a 10^{-3} mientras que los métodos de los pesos multiplicativos como el de mirror descence tienen erorres menores a 10^{-4} lo que es bastante bueno.

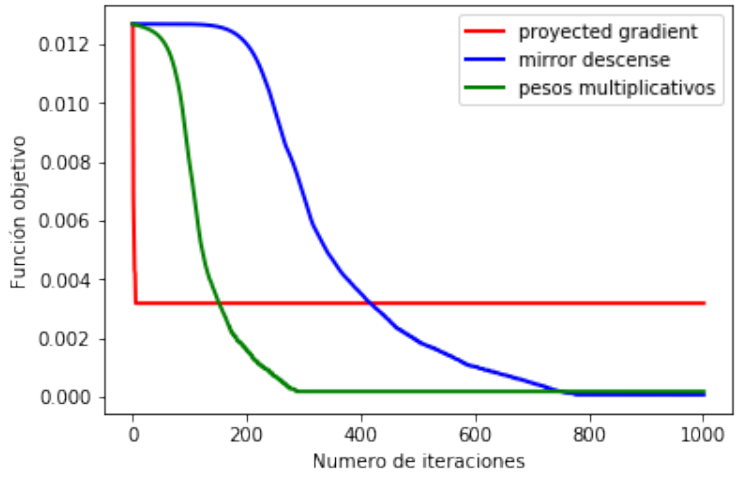
iii. Tiempo de ejecución vs iteraciones:



Comentarios: Como se observa el método del gradiente proyectado tiene tiempos de ejecución mucho menores a diferencia de los métodos de pesos multiplicativos y de mirror descense, los cuales como ya dijimos, poseen errores menores. Así se puede evidanciar un trade-off.

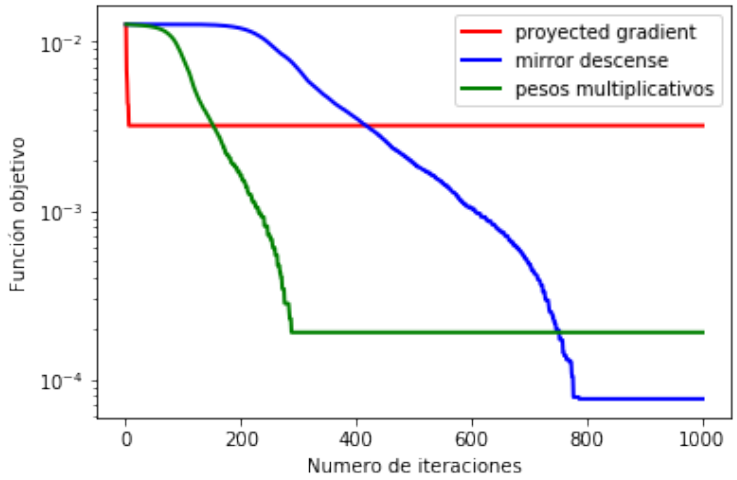
DATASET 3:

i. Función objetivo vs iteraciones (ESCALA NORMAL):



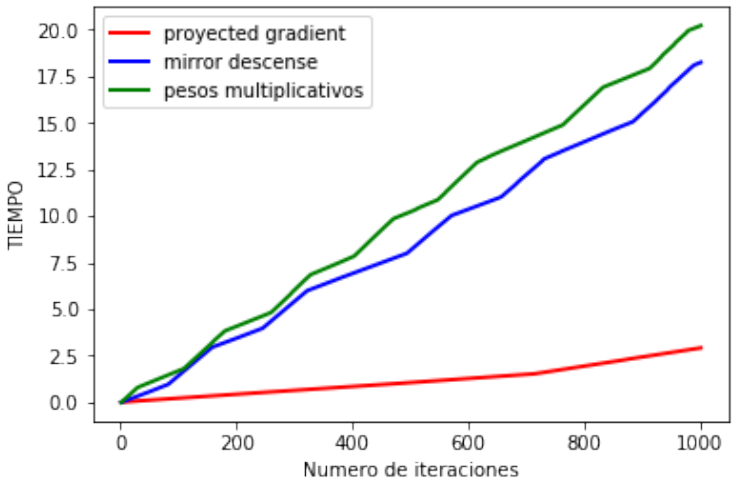
Comentarios: Notemos que si bien el algoritmo de subgradiente proyectado converge rapidamente, luego de la iteración 800 tanto el método de mirror descense como el de los pesos multiplicativos tienen errores menores.

ii. Función objetivo vs iteraciones (ESCALA LOGARITMICA):



Comentarios: Complementando el analisis anterior, gracias a la escala logaritmica se puede observar que el método del gradiente proyectado tiene un error mayor a 10^{-3} mientras que los métodos de los pesos multiplicativos tiene errores menores a 10^{-3} y el de mirror descence tiene erorres menores a 10^{-4} lo que es bastante bueno.

iii. Tiempo de ejecución vs iteraciones:

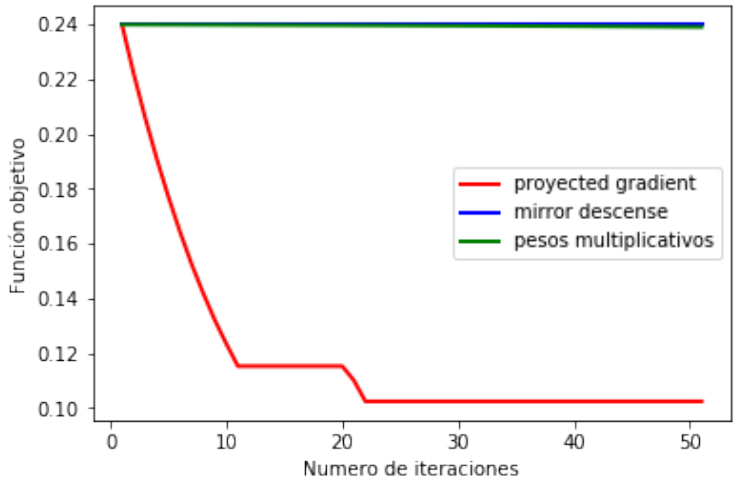


Comentarios: Como se observa el método del gradiente proyectado tiene tiempos de ejecución mucho menores a diferencia de los métodos de pesos multiplicativos y de mirror descense, los cuales como ya dijimos, poseen errores menores. Así se puede evidanciar un trade-off.

Ahora es importante tonar que para los proximos datasets realizamos 50 iteraciones pues teniamos muchos datos $n = 1000$ y el computador no tiraba outputs si colocabamos más iteraciones.

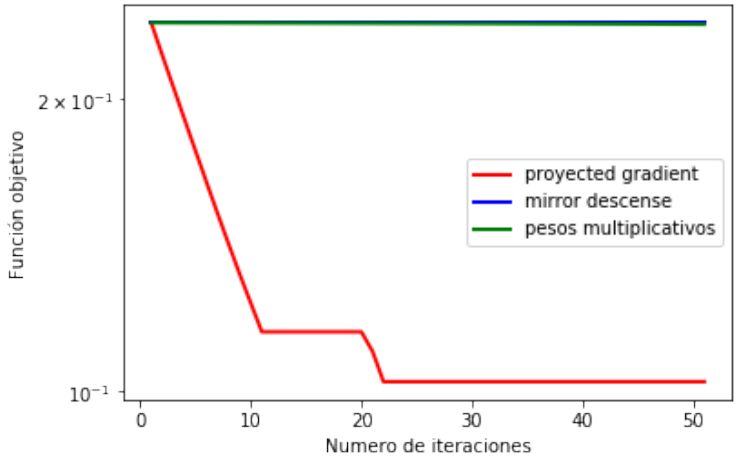
DATASET 4:

i. Función objetivo vs iteraciones (ESCALA NORMAL):



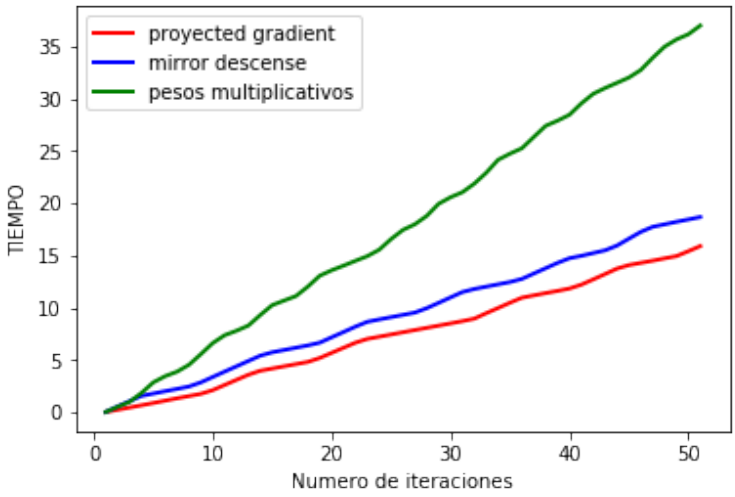
Comentarios: Se observa que el método del gradiente proyectado tiene errores menores que los métodos de los pesos multiplicativos y que el de mirror descense.

ii. Función objetivo vs iteraciones (ESCALA LOGARITMICA):



Comentarios: Complementando lo anterior podemos decir que los errores del gradiente proyectado son menores a 10^{-1} mientras que los métodos de los pesos multiplicativos y de mirror descense son mayores a 10^{-1}.

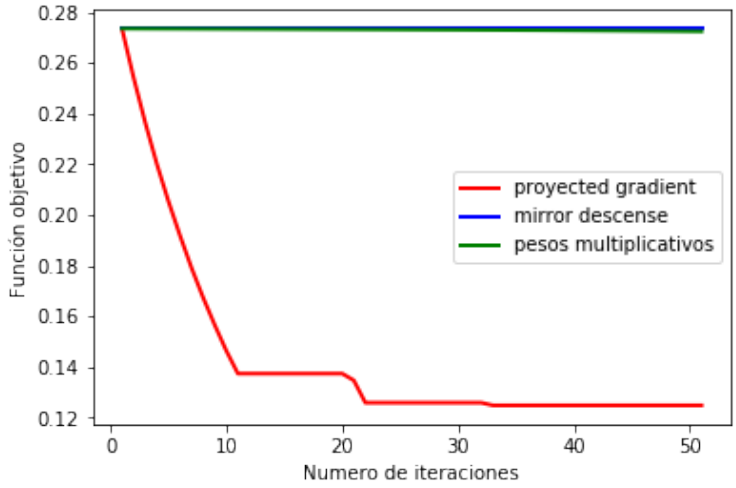
iii. Tiempo de ejecución vs iteraciones:



Comentarios: Respecto a los tiempos, nuevamente gradiente proyectado tiene tiempos menores que los algoritmos de gradiente proyectado y de pesos multiplicativos.

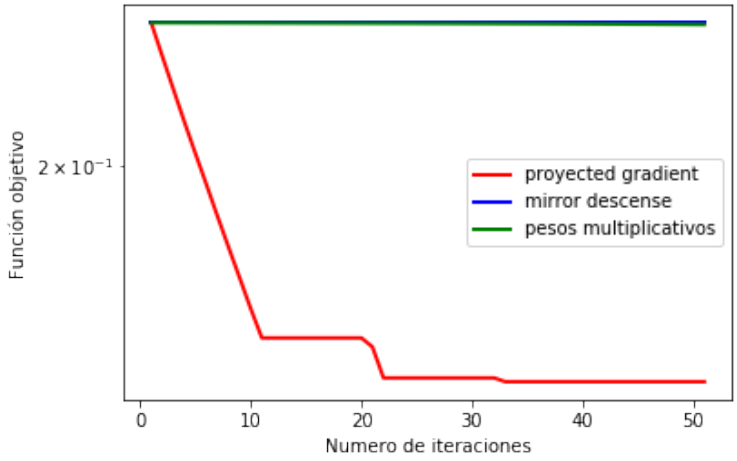
DATASET 5:

i. Función objetivo vs iteraciones (ESCALA NORMAL):



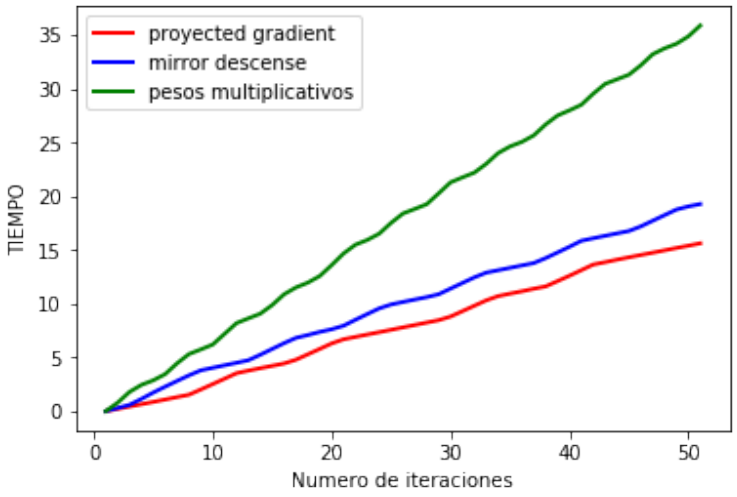
Comentarios: Se observa que el método del gradiente proyectado tiene errores menores que los métodos de los pesos multiplicativos y que el de mirror descense.

ii. Función objetivo vs iteraciones (ESCALA LOGARITMICA):



Comentarios: Complementando lo anterior podemos decir que los errores del gradiente proyectado son menores a 10^{-1} mientras que los métodos de los pesos multiplicativos y de mirror descense son mayores a 10^{-1}.

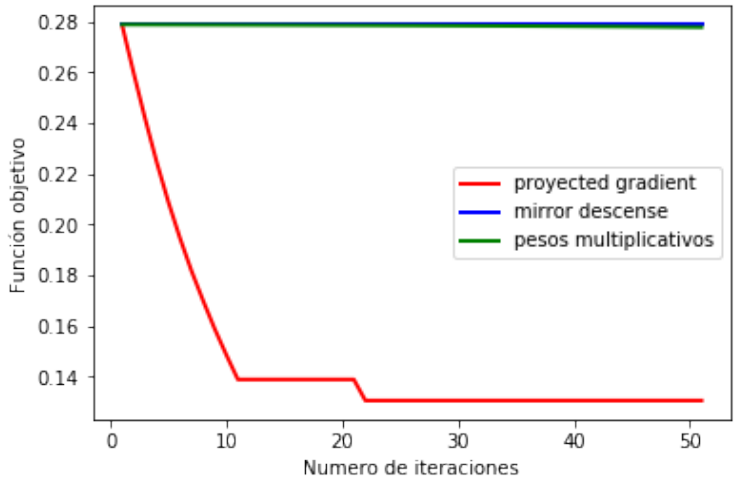
iii. Tiempo de ejecución vs iteraciones:



Comentarios: Respecto a los tiempos, nuevamente gradiente proyectado tiene tiempos menores que los algoritmos de gradiente proyectado y de pesos multiplicativos.

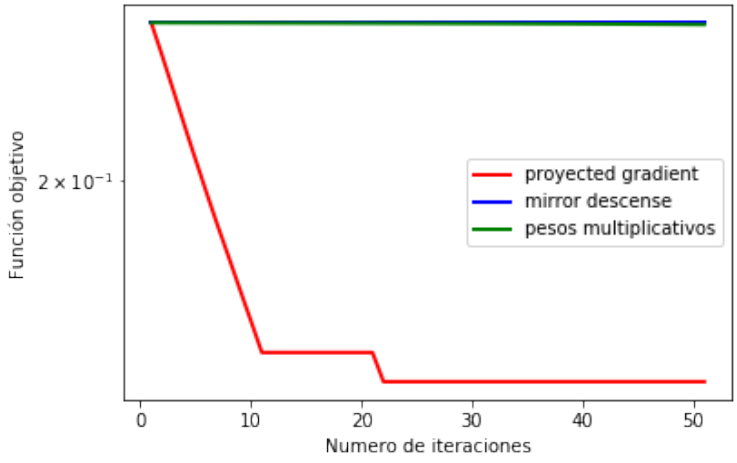
DATASET 6:

i. Función objetivo vs iteraciones (ESCALA NORMAL):



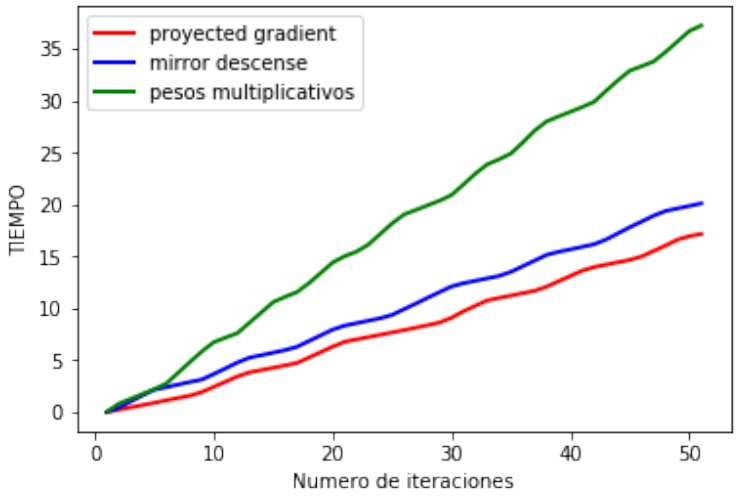
Comentarios: Se observa que el método del gradiente proyectado tiene errores menores que los métodos de los pesos multiplicativos y que el de mirror descense.

ii. Función objetivo vs iteraciones (ESCALA LOGARITMICA):



Comentarios: Complementando lo anterior podemos decir que los errores del gradiente proyectado son menores a 10^{-1} mientras que los métodos de los pesos multiplicativos y de mirror descense son mayores a 10^{-1}.

iii. Tiempo de ejecución vs iteraciones:

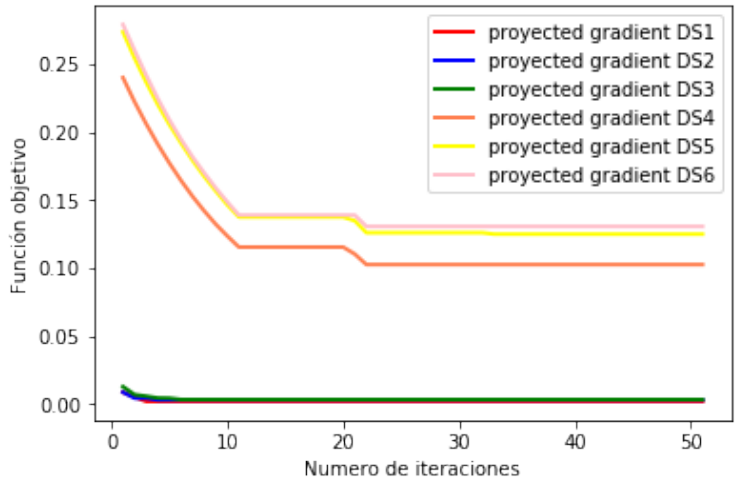


Comentarios: Respecto a los tiempos, nuevamente gradiente proyectado tiene tiempos menores que los algoritmos de gradiente proyectado y de pesos multiplicativos.

Ahora lo que tenemos que hacer es comparar todos los métodos para los diferentes data sets.

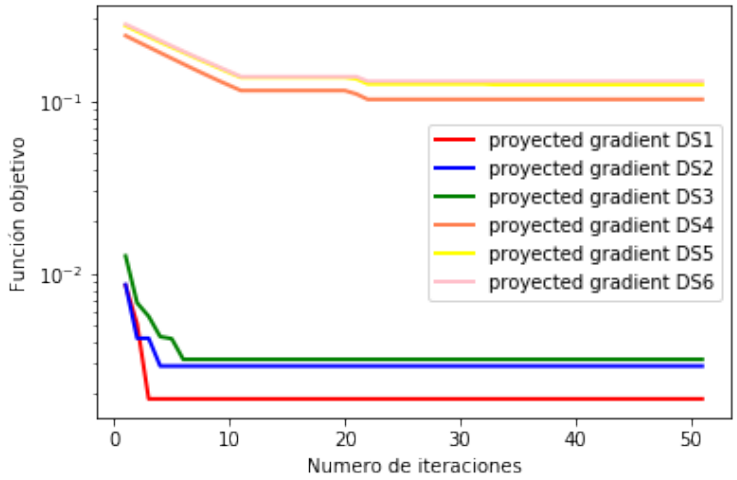
GRADIENTE PROYECTADO:

i. Función objetivo vs iteraciones (ESCALA NORMAL):



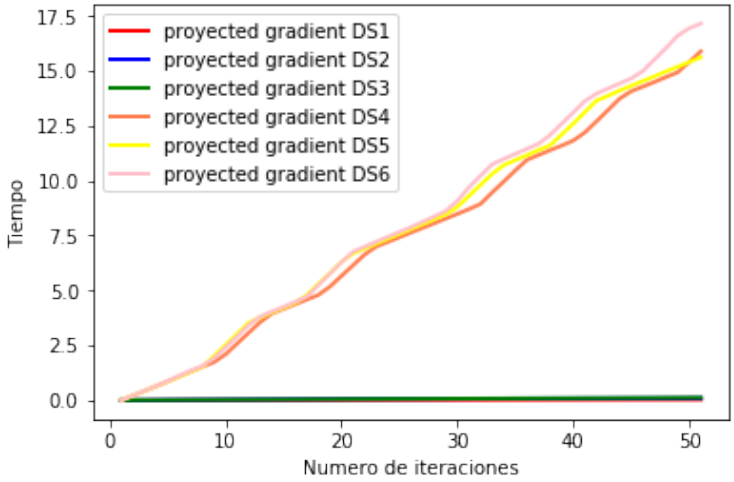
Comentarios: Notemos que los errores de DS1,DS2 y DS3 se comportan de manera similar, al igual que los errores de los DS4, DS5 y DS6. Y además tenemos que los datasets con menores datos tienen errores mucho menores.

ii. Función objetivo vs iteraciones (ESCALA LOGARITMICA):



Comentarios: Complementando lo dicho anteriormente, podemos ver que los errores de los datasets con menos datos son menores que 10^{-2} mientras que los errores de los datasets con más datos tienen errores del orden 10^{-1}.

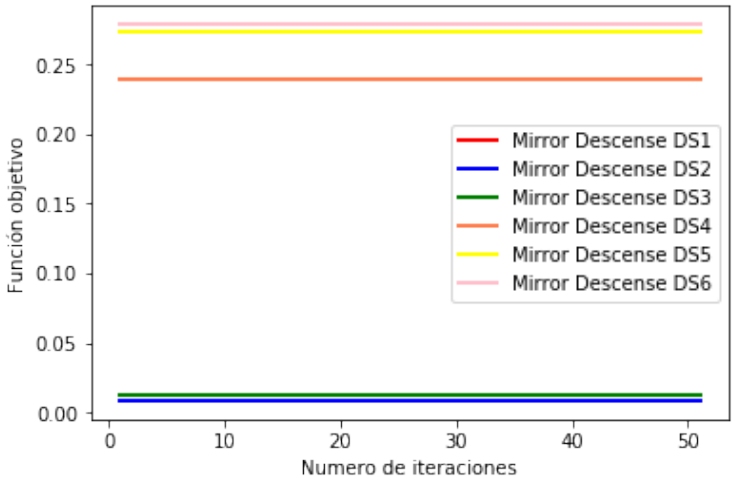
iii. Tiempo de ejecución vs iteraciones:



Comentarios: Se observa claramente que los datasets con más datos poseen tiempos de ejecución mucho más grandes que los datasets con menos datos.

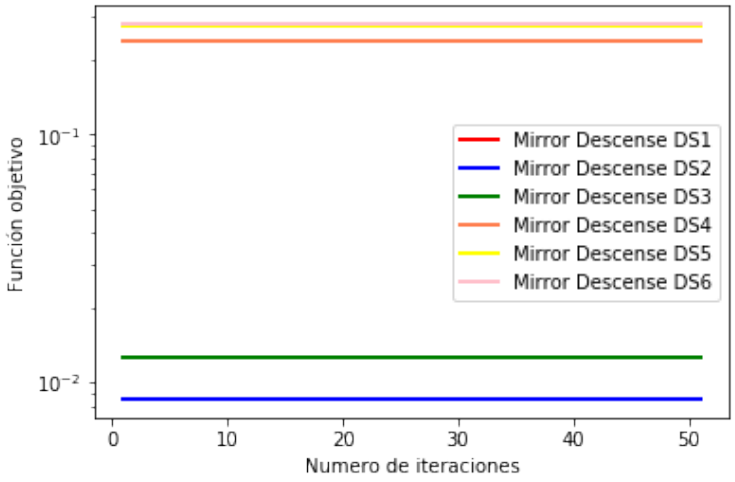
MIRROR DESCENSE:

i. Función objetivo vs iteraciones (ESCALA NORMAL):



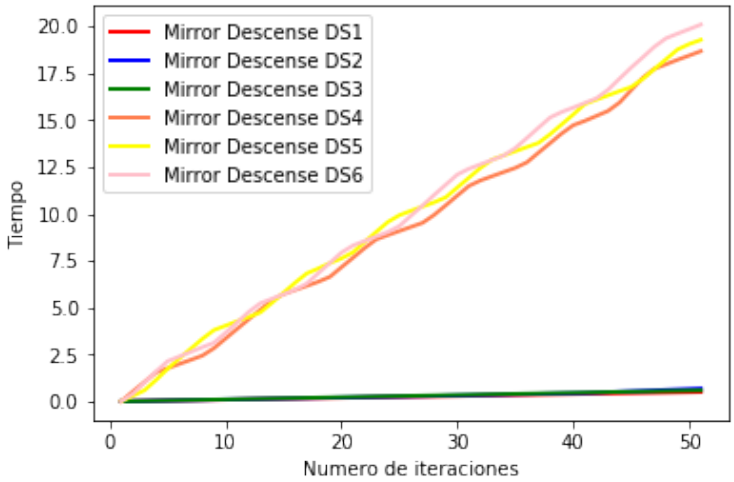
Comentarios: Notemos que aquí los algoritmos se mantienen constantes durante el tiempo y además como era de esperar, los datasets con menos datos poseen errores menores.

ii. Función objetivo vs iteraciones (ESCALA LOGARITMICA):



Comentarios: Complementando lo anterior podemos decir que los errores de los datasets más pequeños son del orden 10^{-2} mientras que los errores de los datasets más grandes son mayores a 10^{-1}.

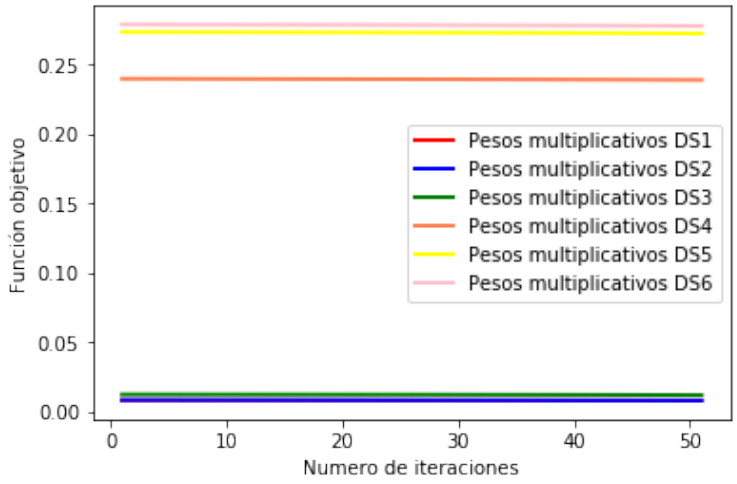
iii. Tiempo de ejecución vs iteraciones:



Comentarios: Como era de esperarse, los datasets con más datos se demoran más tiempo en ejecutarse que los datasets con menos datos.

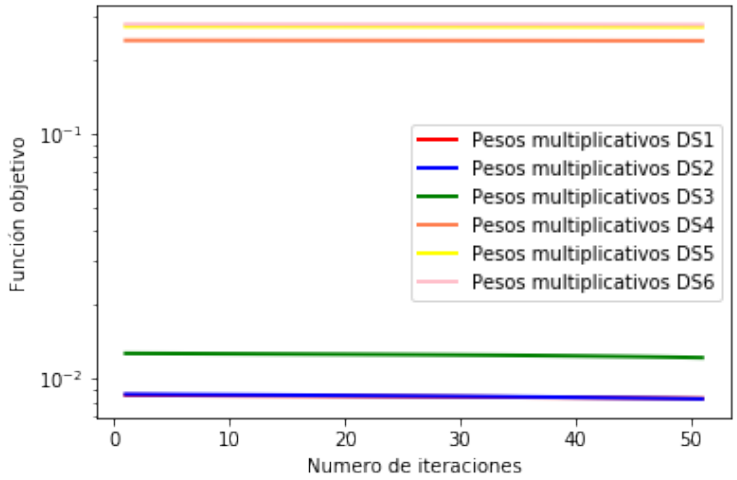
PESOS MULTIPLICATIVOS:

i. Función objetivo vs iteraciones (ESCALA NORMAL):



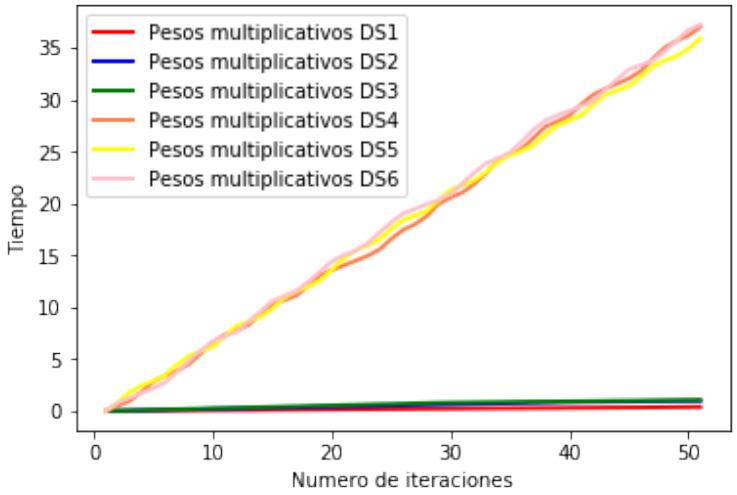
Comentarios: Notemos que aquí los algoritmos se mantienen constantes durante el tiempo y además como era de esperar, los datasets con menos datos poseen errores menores.

ii. Función objetivo vs iteraciones (ESCALA LOGARITMICA):



Comentarios: Complementando lo anterior podemos decir que los errores de los datasets más pequeños son del orden 10^{-2} mientras que los errores de los datasets más grandes son mayores a 10^{-1}.

iii. Tiempo de ejecución vs iteraciones:



Comentarios: Como era de esperarse, los datasets con más datos se demoran más tiempo en ejecutarse que los datasets con menos datos.

Ahora lo que haremos será realizar comentarios generales sobre los distintos métodos y sobre los distintos datasets.

Comentarios generales:

El método del gradiente en todos los datasets siempre presento tiempos menores de ejecución respecto a los métodos de los pesos multiplicativos y de mirror descense. Y además el método del mirror descense posee menores tiempos de ejecución que el método de los pesos multiplicativos. Pero para datasets con n chicos ocurre que el método de los pesos multiplicativos es el más eficiente, seguido por el método de mirror descense, mientras que el método del subgradiente proyectado es el más ineficiente. Por otro lado, cuando tenemos datasets más grandes, el método del subgradiente proyectado es el que posee los menores errores, y el método de los pesos multiplicativos y de mirror descense poseen errores similars.

Es también bueno notar que los comportamientos del método de mirror descense y el de pesos multiplicativos poseen comportamientos similares en cuanto a tiempo de ejecución y precisión de la soluciones, a pesar de que si se observa la formulación de los algoritmos poseen forma de ejecutarse muy distintas, pero podriamos decir o conjeturar que este comportamiento similar tiene un motivo más de fondo, que sería que ambos métodos son de mirror descense solo que con diferentes funciones $\Phi(\cdot)$ pero para ambos casos, esta función es 1-fuertemente convexa.