MÉTODOS DE PRIMER ORDEN?

ANÁLISIS DE CONVERGENCIA??

Universidad de Buenos Aires

Facultad de Ciencias Exactas y Naturales
Departamento de Matemática

Tesis de Licenciatura
Director de Tesis: Dr. Pablo Amster
Septiembre 2018 – version 0.1

Axel Sirota: *Métodos de primer orden?,* Análisis de convergencia??, © Septiembre 2018

ABSTRACT

Aca va a ir el abstract cuando lo tengamos

We have seen that computer programming is an art,
because it applies accumulated knowledge to the world,
because it requires skill and ingenuity, and especially
because it produces objects of beauty.

— Donald E. Knuth [1]

AGRADECIMIENTOS

Agradecimientos para todos

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LISTINGS

ACRÓNIMOS

Part I

Introducción

1

INTRODUCCIÓN

De lo dicho en [4] y [5]

Part II

Algoritmos de tipo Batch

En esta parte vamos a analizar los tipos de convergencia de los diferentes algoritmos de primer orden usados en Machine Learning

2

INTUICIÓN

Usemos un caso modelo para ejemplificar porque no es probable que los metodos de primer orden (entre ellos *gradient descent*) convergan a puntos silla. Sea $f : \mathbb{R}^n \to \mathbb{R}^n$ dada por $f(x) = \frac{1}{2}x^T Hx$ con $H = \mathbf{diag}(\lambda_1, \ldots, \lambda_n)$; supongamos además que $\lambda_1, \ldots, \lambda_k > 0$ y $\lambda_{k+1}, \ldots, \lambda_n < 0$.

Si usamos en la base canónica de $\mathbb{R}^n$ $\left\{e^1, \ldots, e^n\right\}$ entonces:

$$f(x) = f(x^1, \ldots, x^n) = \frac{1}{2}\left(\lambda_1 {x_1}^2 + \cdots + \lambda_n {x_n}^2\right)$$

Por lo tanto:

$$\nabla f(x) \;=\; \lambda_i x_i e^i = 0 \iff x = x_1 e^1 = 0$$

Y tenemos que en el único punto crítico el Hessiano de f es $\nabla^2 f(0) = H$.

Recordemos que si $g(x) = x - \alpha \nabla f(x)$ entonces *gradient descent* está dado por la iteración $x_{t+1} = g(x_t) := g^t(x_0)$ con $t \in \mathbb{N}$ y $x_0 \in \mathbb{R}^n$, y en este caso esta representado por:

$$\begin{aligned} x_{t+1} =& g(x_t) \\ =& x_t - \alpha \nabla f(x_t) \\ =& \left(1 - \alpha\lambda_i\right) x_{it} e^i \\ =& \left(1 - \alpha\lambda_i\right) \left\langle x_t, e^i \right\rangle e^i \end{aligned}$$

Por lo tanto por inducción es fácil probar que:

$$x_{t+1} = \left(1 - \alpha\lambda_i\right)^t \left\langle x_o, e^i \right\rangle e^i$$

Sea $L = \max_i |\lambda_i|$ y supongamos que $\alpha < \frac{1}{L}$, luego:

$$\begin{aligned} 1 - \alpha\lambda_i < 1 \quad \text{Si } i \leq k \\ 1 - \alpha\lambda_i > 1 \quad \text{Si } i > k \end{aligned}$$

Con lo que concluímos que:

$$\lim_t x_t = \begin{cases} 0 & \text{Si } x \in E_s := \left\langle e^1, \ldots, e^k \right\rangle \\ \infty & \text{Si no} \end{cases}$$

Finalmente, si $k < n$ entonces concluímos que:

$$P_{\mathbb{R}^n}(\left\{ x \in \mathbb{R}^n \;/\; \lim_t g^t(x) = 0 \right\}) = |E_s| = 0$$

Para notar este fenómeno en un ejemplo no cuadrático consideremos $f(x,y) = \frac{1}{2}x^2 + \frac{1}{4}y^4 - \frac{1}{2}y^2$, reproduciendo los calculos anteriores:

$$\begin{aligned} \nabla f &= (x, y^3 - y) \\ g &= \left((1-\alpha)x, (1+\alpha)y - \alpha y^3\right) \\ \nabla^2 f &= \begin{pmatrix} 1 & 0 \\ 0 & 3y^2 - 1 \end{pmatrix} \end{aligned} \tag{1}$$

De lo que vemos que los puntos críticos son:

$$z_1 = (0,0) \qquad z_2 = (0,1) \qquad z_3 = (0,-1)$$

Y del critério del Hessiano concluímos que z_2, z_3 son mínimos locales mientras que z_1 es un punto silla. De la intuición previa, como en z_1 el autovector asociado al autovalor positivo es e^1 podemos intuir que:

Lema 2.1 *Para $f(x,y) = \frac{1}{2}x^2 + \frac{1}{4}y^4 - \frac{1}{2}y^2$ resulta que $E_s = \left\langle t * e^1 \ / \ t \in \mathbb{R} \right\rangle := W_s$*

Asumiendo el resultado por un momento, dado que $\dim_{\mathbb{R}^2}(E_s) = 1 < 2$ entonces $P_{\mathbb{R}^2}(E_s) = 0$ que es lo que queríamos verificar. Demostremos el lema ahora:

Demostración Del lema Sea $x_0 \in \mathbb{R}^n$ y g la iteración de *gradient descent* dada por 2, luego:

$$(x_t, y_t) = g^t(x,y) = \begin{pmatrix} (1-\alpha)^t x_0 \\ g_y^t(y_0) \end{pmatrix} \underset{(t\to\infty)}{\longrightarrow} \begin{pmatrix} 0 \\ \lim_t g_y^t(y_0) \end{pmatrix}$$

Por lo que todo depende de y_0. Analizando $\dfrac{dg_y}{dy} = 1 + \alpha - 3\alpha y^2$ notemos que:

$$\begin{aligned} \left|\frac{dg_y}{dy}\right| < 1 \iff & \left|1 + \alpha - 3\alpha y^2\right| < 1 \\ \iff & -1 < 1 + \alpha - 3\alpha y^2 < 1 \\ \iff & -2 - \alpha < -3\alpha y^2 < -\alpha \\ \iff & \sqrt{\frac{2+\alpha}{3\alpha}} > |y| > \sqrt{\frac{1}{3}} \\ \iff & \sqrt{\frac{1+\frac{2}{\alpha}}{3}} > |y| > \sqrt{\frac{1}{3}} \end{aligned}$$

Por lo que por el Teorema de Punto Fijo de Banach:

$$\lim_t g_y^t(y_0) = \begin{cases} 1 & \text{Si } \sqrt{\dfrac{1+\frac{2}{\alpha}}{3}} > y_0 > \sqrt{\dfrac{1}{3}} \\ -1 & \text{Si } \sqrt{\dfrac{1+\frac{2}{\alpha}}{3}} < -y_0 < \sqrt{\dfrac{1}{3}} \end{cases}$$

Si analizamos simplemente los signos de g y $\frac{dg_y}{dy}$ en los otros intervalos podemos conluir que:

$$\lim_t g_y^t(y_0) = \begin{cases} -\infty & \text{Si } y_0 > \sqrt{\frac{1+\frac{2}{\alpha}}{3}} \\ 1 & \text{Si } \sqrt{\frac{1+\frac{2}{\alpha}}{3}} > y_0 > 0 \\ -1 & \text{Si } -\sqrt{\frac{1+\frac{2}{\alpha}}{3}} < y_0 < 0 \\ \infty & \text{Si } y_0 < -\sqrt{\frac{1+\frac{2}{\alpha}}{3}} \end{cases}$$

Dedujimos entonces que $(x,y) \in E_s \iff (x,y) = (t,0)\ t \in \mathbb{R} \iff (x,y) \in W_s$. ∎

3

TEOREMA DE LA VARIEDAD ESTABLE Y LOS PUNTOS FIJOS INESTABLES

3.1 RESULTADOS PREVIOS

Esto quizas deberia ir en prerequisitos cuando lo tengamos

Por el resto del documento, $g : \chi \to \chi$ y χ es una $d-$variedad sin borde.

Definición Dada una variedad de dimensión d χ y el espacio de medida $\left(\mathbb{R}^d, \mathcal{B}, \mu\right)$, decimos que $E \subset \chi$ tiene *medida cero* si existe un atlas $\mathcal{A} = \left\{U_i, \phi^i\right\}_{i\in\mathbb{N}}$ tal que $\mu\left(\phi^i\left(E \cap U_i\right)\right) = 0$. En este caso usamos el abuso de notación $\mu(E) = 0$.

Lema 3.1 *Sea $E \subset \chi$ tal que $\mu(E) = 0$; si $\det\left(Dg(x)\right) \neq 0$ para todo $x \in \chi$, luego $\mu\left(g^{-1}(E)\right) = 0$*

Demostración Sea $h = g^{-1}$ y $\left(V_i, \psi^i\right)$ una colección de cartas en el dominio de g, si verificamos que $\mu\left(h\left(E\right) \cap V_i\right) = 0$ para todo $i \in \mathbb{N}$ entonces:

$$\mu(h(E)) = \mu\left(\bigcup_{i\in\mathbb{N}} h(E) \cap V_i\right) \leqslant \sum_{i\in\mathbb{N}} \mu\left(h(E) \cap V_i\right) = 0$$

Sin pérdida de generalidad podemos asumir que $h(E) \subseteq V$ con $(V, \psi) \in \left\{(V_i, \phi^i)\right\}$ una carta determinada. Sea $\mathcal{A} := \left\{\left(U_i, \varphi^i\right)\right\}$ un atlas de χ y notemos $E_i = E \cap U_i$; luego $E = \bigcup\limits_{i\in\mathbb{N}} E_i = \bigcup\limits_{i\in\mathbb{N}} {\varphi^i}^{-1} \circ \varphi^i\left(E_i\right)$ por lo que:

$$\begin{aligned} \mu\left(\psi \circ h(E)\right) = & \quad \mu\left(\psi \circ h\left(\bigcup_{i\in\mathbb{N}} {\varphi^i}^{-1} \circ \varphi^i\left(E_i\right)\right)\right) \\ \leqslant & \quad \sum_{i\in\mathbb{N}} \mu\left(\psi \circ h \circ {\varphi^i}^{-1}\left(\varphi^i(E_i)\right)\right) \end{aligned}$$

Uso Teorema de la funcion inversa en variedades y que localmente Lipshitz preserva medida

Por hipótesis $\varphi^i(E_i)$ es de medida cero, luego como g es difeomorfismo local por **??** entonces $\psi \circ h \circ {\varphi^i}^{-1} \in C^1$. Como si $f \in C^1(\mathbb{R}^d)$ entonces es localmente Lipshitz, ergo f preserva la medida, concluímos que $\mu\left(\psi \circ h \circ {\varphi^i}^{-1}\left(\varphi^i(E_i)\right)\right) = 0$ para todo $i \in \mathbb{N}$. ∎

3.2 PUNTOS FIJOS INESTABLES

Definición Sea:

$$\mathcal{A}_g^* := \left\{x \ : \ g(x) = x \quad \max_i \left|\lambda_i\left(Dg(x)\right) > 1\right|\right\}$$

El conjunto de puntos fijos de g cuyo diferencial en ese punto tiene algún autovalor mayor que 1. A este conjunto lo llamaremos el conjunto de *puntos fijos inestables*

Este teorema debería ir en prerequisitos

Teorema 3.1 *Sea x^* un punto fijo de $g \in C^r(\chi)$ un difeomorfismo local. Supongamos que $E = E_s \oplus E_u$ donde*

$$\begin{aligned} E_s = &\quad \langle \{v_i \ / \ Dg(x^*)v_i = \lambda_i v_i \quad , \quad \lambda_i \leqslant 1\} \rangle \\ E_u = &\quad \langle \{v_i \ / \ Dg(x^*)v_i = \lambda_i v_i \quad , \quad \lambda_i > 1\} \rangle \end{aligned}$$

Entonces existe $W^{cs}_{loc} \hookrightarrow \chi$ un embedding C^r local tangente a E_s en x^* llamado la variedad local estable central que cumple que existe $B \ni x^*$ entorno tal que $g\left(W^{cs}_{loc}\right) \cap B \subseteq W^{cs}_{loc}$ y $\bigcap\limits_{k\in\mathbb{N}} g^{-k}(B) \subseteq W^{cs}_{loc}$

Con todos estos resultados demostremos el teorema principal:

Teorema 3.2 *Sea $g \in C^1(\chi)$ tal que $\det(Dg(x)) \neq 0$ para todo $x \in \chi$, luego el conjunto de puntos iniciales que convergen por g a un punto fijo inestable tiene medida cero, i. e.:*

$$\mu\left(\left\{x_0 : \lim_k g^k(x_0) \in \mathcal{A}^*_g\right\}\right) = 0$$

Demostración Para cada $x^* \in \mathcal{A}^*_g$ por 3.2 existe B_{x^*} un entorno abierto; es más, $\bigcup\limits_{x^*\in\mathcal{A}^*_g} B_{x^*}$ forma un cubrimiento abierto del cual existe un subcubrimiento numerable pues X es variedad, i. e.

$$\bigcup_{x^*\in\mathcal{A}^*_g} B_{x^*} = \bigcup_{i\in\mathbb{N}} B_{x^*_i}$$

Usamos que en una variedad se cumple la propiedad de Lindeloff

Primero si $x_0 \in \chi$ sea:

$$\begin{aligned} x_k =& g^k(x_0) \\ =& \underbrace{g \circ \cdots \circ g}_{k \text{ veces}}(x_0) \end{aligned}$$

la sucesión del flujo de g evaluado en x_0, entonces si $W := \left\{x_0 : \lim\limits_k x_k \in \mathcal{A}^*_g\right\}$ queremos ver que $\mu(W) = 0$.

Sea $x_0 \in W$, luego como $x_k \to x^* \in \mathcal{A}^*_g$ entonces existe $T \in \mathbb{N}$ tal que para todo $t \geqslant T$, $x_t \in \bigcup\limits_{i\in\mathbb{N}} B_{x^*_i}$ por lo que $x_t \in B_{x^*_i}$ para algún $x^*_i \in \mathcal{A}^*_g$ y $t \geqslant T$. Afirmo que:

Pablo: Hace falta demostrar esto??

Lema 3.3 *$x_t \in \bigcap\limits_{k\in\mathbb{N}} g^{-k}(B_{x^*_i})$ para todo $t \geqslant T$*

Si notamos $S_i \overset{\triangle}{=} \bigcap\limits_{k\in\mathbb{N}} g^{-k}(B_{x^*_i})$, entonces por 3.1 sabemos por un lado que es una subvariedad de W^{cs}_{loc} y por el otro que $\dim(S_i) \leqslant \dim(W^{cs}_{loc}) = \dim(E_s) < d - 1$ [1]; por lo que $\mu(S_i) = 0$.

Usamos que la dimension de la variedad es la de su tangente

Finalmente como $x_T \in S_i$ para algún T entonces $x_0 \in \bigcup_{k\in\mathbb{N}} g^{-k}(S_i)$ por lo que $W \subseteq \bigcup_{i\in\mathbb{N}} \bigcup_{k\in\mathbb{N}} g^{-k}(S_i)$. Concluímos:

Usamos que una subvariedad de dimension menor tiene medida o

$$\begin{aligned} \mu(W) &\leqslant \mu\left(\bigcup_{i\in\mathbb{N}} \bigcup_{k\in\mathbb{N}} g^{-k}(S_i)\right) \\ &\leqslant \sum_{i\in\mathbb{N}} \sum_{k\in\mathbb{N}} \mu\left(g^{-k}(S_i)\right) \\ &\overset{3.1}{=} 0 \end{aligned}$$

∎

Para finalizar veamos un caso simple que nos encontraremos seguido:

Corolario 3.4 *Bajo las mismas hipótesis que en 3.2 si agregamos que* $\chi* \subseteq \mathcal{A}_g^*$ *entonces* $\mu(W_g) = 0$

Demostración Como $\chi^* \subseteq \mathcal{A}_g^*$ entonces $W_g \subseteq W$, luego $\mu(W_g) \leqslant \mu(W) \overset{3.2}{=} 0$. ∎

1 Por que???

4

APLICACIÓNES

4.1 GRADIENT DESCENT

Como una aplicación del teorema en 3.2 demostremos que *gradient descent* tiene probabilidad cero de converger a puntos silla. Consideremos *gradient descent* con *learning rate* α:

$$x_{k+1} = g(x_k) \stackrel{\triangle}{=} x_k - \alpha \nabla f(x_k) \tag{2}$$

Hipótesis 1 Asumamos que $f \in \mathcal{C}^2$ y $\left\|\nabla^2 f(x)\right\|_2 \leqslant L$

Proposición 4.1 *Todo punto silla estricto de f es un punto fijo inestable de g, i. e.$\chi^* \subseteq \mathcal{A}_g^*$.*

Demostración Es claro que un punto crítico de f es punto fijo de g; si $x^* \in \chi^*$ entonces $Dg(x^*) = Id - \alpha \nabla^2 f(x^*)$ y entonces los autovalores de Dg son $\left\{1 - \alpha\lambda_i \ : \ \lambda_i \in \left\{\mu \ : \ \nabla^2 f(x^*)v = \mu v \quad \text{para algún } v \neq 0\right\}\right\}$. Como $x^* \in \chi^*$ existe $\lambda_{j^*} < 0$ por lo que $1 - \alpha\lambda_{j^*} > 1$; concluímos que $x^* \in \mathcal{A}_g^*$. ∎

Usamos que f(A) tiene autovalores $f(\{\lambda_i\})$

Proposición 4.2 *Bajo 4.1 y $\alpha < \frac{1}{L}$ entonces* $\det\left(Dg(x)\right) \neq 0$.

Demostración Como ya sabemos $Dg(x) = Id - \alpha \nabla^2 f(x)$ por lo que:

$$\det\left(Dg(x)\right) = \prod_{i \in \{1,\dots,d\}} (1 - \alpha\lambda_i)$$

Luego por 4.1 tenemos que $\alpha < \frac{1}{|\lambda_i|}$ y entonces $1 - \alpha\lambda_i > 0$ para todo $i \in \{1, \dots, d\}$; concluímos que $\det\left(Dg(x)\right) > 0$. ∎

Corolario 4.3 *Sea g dada por Gradient descent en 2, bajo 4.1 y $\alpha < \frac{1}{L}$ se tiene que $\mu\left(W_g\right) = 0$.*

Demostración Por 4.1 y 4.2 tenemos que vale 3.4 y concluímos que $\mu\left(W_g\right) = 0$. ∎

4.2 PUNTO PRÓXIMO

El algoritmo de punto próximo esta dado por la iteración:

$$x_{k+1} = g(x_k) \stackrel{\triangle}{=} \arg\min_{z \in \chi} f(z) + \frac{1}{2\alpha} \|x_k - z\|_2^2 \tag{3}$$

Proposición 4.1 *Bajo 4.1 y $\alpha < \frac{1}{L}$ entonces vale:*

1. $\det(Dg(x)) \neq 0$
2. $\chi^* \subseteq \mathcal{A}_g^*$

Probamos esto? Me parece un poco claro

Demostración Veamos primero el siguiente lema:

Lema 4.2 *Bajo 4.1, $\alpha < \frac{1}{L}$ y $x \in \chi$ entonces $f(z) + \frac{1}{2\alpha}\|x - z\|_2^2$ es estrictamente convexa, por lo que $g \in \mathcal{C}^1(\chi)$*

Por lo tanto por 4.2 podemos tomar límite, i. e.

$$\begin{aligned}
x_{k+1} = \quad g(x_k) = \quad & \arg\min_{z\in\chi} f(z) + \frac{1}{2\alpha}\|x_k - z\|_2^2 \\
\downarrow \qquad \downarrow \qquad & \qquad\qquad\qquad\qquad \downarrow \\
x = \quad g(x) = \quad & \arg\min_{z\in\chi} f(z) + \frac{1}{2\alpha}\|x - z\|_2^2 \\
\iff & \nabla_z \left(f(z) + \frac{1}{2\alpha}\|x - z\|^2 \right)(g(x)) = 0 \\
\iff & \nabla f(g(x)) - \frac{1}{\alpha}(x - g(x)) = 0 \\
\iff & g(x) + \alpha\nabla f(g(x)) = x
\end{aligned}$$

Finalmente por diferenciación implicita obtenemos:

$$\begin{aligned}
& Dg(x) + \alpha\nabla^2 f(g(x))Dg(x) = Id \\
\implies \quad & Dg(x) = \left(Id + \alpha\nabla^2 f(g(x))\right)^{-1}
\end{aligned}$$

Luego si $x^* \in \chi^*$ entonces $Dg(x^*) = \left(Id + \alpha\nabla^2 f(x^*)\right)^{-1}$ y tiene autovalores $\left\{\frac{1}{1+\alpha\lambda_i}\right\}$ con λ_i autovalores de $\nabla^2 f(x^*)$. Por lo tanto $x^* \in \mathcal{A}_g^*$ y para $\alpha < \frac{1}{L}$ se tiene que $\det(Dg(x)) \neq 0$. ∎

Corolario 4.3 *Sea g dado por el algoritmo de punto próximo con ecuación 3, bajo 4.1 y $\alpha < \frac{1}{L}$ se tiene que $\mu(W_g) = 0$.*

Demostración Por 4.1 tenemos que vale 3.4 y concluímos que $\mu(W_g) = 0$. ∎

4.3 DESCENSO POR COORDENADAS

Sea $S_1, \ldots, S_b$ una partición disjunta de $\{1, \ldots, d\}$ donde d y b son parámetros del método.

Consideremos el ritmo 1:

Algorithmus 1 : Descenso por coordenadas

Input: $f \in C^1$, $\alpha > 0$, $x_0 \in \chi$
for $k \in \mathbb{N}$ **do**
  for ***block*** $i = 1, \dots, b$ **do**
    for ***index*** $j \in S_i$ **do**
      $y_k^{S_0} = x_k$ e $y_k^{S_i} = \left(x_{k+1}^{S_1}, \dots, x_{k+1}^{S_i}, x_k^{S_{i+1}}, \dots, x_k^{S_b}\right)$
      $x_{k+1}^j \leftarrow x_k^j - \alpha \frac{\partial f}{\partial x_j}\left(y_k^{S_{i-1}}\right)$
    end
  end
end

Luego si definimos $g_i(x) = x - \alpha \sum_{j \in S_i} e_j^T \nabla f(x)$ entonces:

Lema 4.1 *La iteración de Descenso por coordenadas esta dada por:*

$$x_{k+1} = g(x_k) \stackrel{\triangle}{=} g_d \circ g_{d-1} \circ \cdots \circ g_1(x) \tag{4}$$

Lema 4.2 *Si g está dada por 4 entonces si notamos $P_S = \sum_{i \in S} e_i e_i^T$ entonces:*

$$Dg(x_k) = \prod_{i \in \{1,\dots,b\}} \left(Id - \alpha P_{b-i+1} \nabla^2 f(y_k^{S_{b-i}})\right) \tag{5}$$

Demostración Notemos primero que:

$$Dg_i(x) = Id - \alpha P_{S_i} \nabla^2 f(x)$$

Por lo tanto:

$$\begin{aligned} Dg(x_k) = & \quad D\left(g_b \circ \dots g_1\right)(x_k) \\ = & \quad \left(Id - \alpha P_{S_b} \nabla^2 f\right) \left(\underbrace{g_{b-1} \circ \dots g_1(x_k)}_{y_k^{S_{b-1}}}\right) D\left(g_{b-1} \circ \dots g_1\right)(x_k) \\ \vdots & \\ = & \quad \prod_{i \in \{1,\dots,b\}} \left(Id - \alpha P_{b-i+1} \nabla^2 f(y_k^{S_{b-i}})\right) \end{aligned}$$

∎

Observación Sea $f \in C^2$ y notemos $\nabla^2 f|_S$ a la submatriz que resulta de quedarme con filas y columnas indexadas por S. Sea $\max_{i \in \{1,\dots,b\}} \left\|\nabla^2 f(x)|_{S_i}\right\| = L_b$

Proposición 4.3 *Bajo 9 y $\alpha < \frac{1}{L_b}$ se tiene que* $\det(Dg(x)) \neq 0$

Demostración Basta probar que cada término de 5 es invertible, para eso:

$$\begin{aligned}\chi_{Dg_i(x)}(\lambda) = & \quad \det\left(\lambda Id_d - Id_d - \alpha P_{S_{b-i+1}} \nabla^2 f(x)\right) \\ = & \quad (\lambda-1)^{n-|S_i|} \prod_{j\in S_i}\left(\lambda - 1 + \alpha \frac{\partial^2 f}{\partial x_j^2}(x)\right)\end{aligned}$$

Luego si $\alpha < \frac{1}{L_{max}}$ entonces $\lambda - 1 + \alpha \frac{\partial^2 f}{\partial x_j^2}(x) > 0$ para todo $j \in S_i$, $i \in \{1,\dots,b\}$ por lo que todos los autovalores son positivos y $Dg_i(x)$ es invertible para todo i. ∎

Proposición 4.4 *Bajo 9 y $\alpha < \frac{1}{L_{max}}$ se tiene que $\chi^* \subseteq \mathcal{A}_g^*$*

Demostración Sea $x^* \in \chi^*$, $H = \nabla^2 f(x^*)$, $J = Dg(x^*) = \prod_{i\leq b}\left(Id_n - \alpha P_{S_{b-i+1}} H\right)$ e y_0 el autovector correspondiente al menor autovalor de H. Vamos a probar que $\left\|J^t y_0\right\|_2 \geq c(1+\epsilon)^t$ por lo que $\left\|J^t\right\|_2 \geq c(1+\epsilon)^t$, luego por el teorema de Gelfand

Usamos que el radio espectral es el limite de cualquier norma matricial

$$\rho(J) = \lim_{t\to\infty} \left\|J^t\right\|^{1/t} \geq \lim_{t\to\infty} c^{1/t}(1+\epsilon) = 1+\epsilon$$

Y concluímos que $\chi^* \subseteq \mathcal{A}_g^*$.

En pos de eso fijemos $t \geq 1$ una iteración , $y_t = J^t x_0$, $z_1 = y_t$ y definamos $z_{i+1} = (Id - \alpha P_{S_i} H) z_i = z_i - \alpha \sum_{j\in S_i}\left(e_j^T H z_i\right) e_j$. Luego $y_{t+1} = z_{b+1}$, afirmo:

Esta demo es horrenda, hay que pensar una mejor y pionerla en el Anexo

Afirmación 4.5 *Sea $y_t \in Ran(H)$, luego existe $i \in \{1,\dots,b\}$ y $\delta > 0$ tal que $\alpha \sum_{j\in S_i}\left|e_j^T H z_i\right| \geq \delta \left\|z_i\right\|_2$*

Lema 4.6 *Existe $\epsilon > 0$ tal que para todo $t \in \mathbb{N}$:*

$$y_{t+1}^T H y_{t+1} \leq (1+\epsilon) y_t^T H y_t$$

Demostración Manteniendo la notación previa a la afirmación:

$$
\begin{aligned}
z_{i+1}^T H z_{i+1} \leq & \left[z_i^T - \alpha \sum_{j \in S_i} \left(e_j^T H z_i\right) e_j^T\right] H \left[z_i - \alpha \sum_{j \in S_i} \left(e_j^T H z_i\right) e_j\right] \\
= & \; z_i^T H z_i - \alpha \sum_{j \in S_i} \left(z_i^T H e_j\right) \left(e_j^T H z_i\right) - \alpha \sum_{j \in S_i} \left(e_j^T H z_i\right) \left(e_j^T H z_i\right) \\
& + \alpha^2 \left(\sum_{j \in S_i} \left(e_j^T H z_i\right) e_j\right)^T H \left(\sum_{j \in S_i} \left(e_j^T H z_i\right) e_j\right) \\
\left(\|H_{S_i}\|_2 \leq L_b\right) < & \; z_i^T H z_i - 2\alpha \sum_{j \in S_i} \left(e_j^T H z_i\right)^2 + \alpha^2 L_b \left\| \sum_{j \in S_i} \left(e_j^T H z_i\right) e_j \right\|_2^2 \\
= & \; z_i^T H z_i - \alpha \left(2 - \alpha L_b\right) \left\| \sum_{j \in S_i} \left(e_j^T H z_i\right) e_j \right\|_2^2 \\
\left(\alpha L_b < 1\right) < & \; z_i^T H z_i - \alpha \left\| \sum_{j \in S_i} \left(e_j^T H z_i\right) e_j \right\|_2^2
\end{aligned}
$$

Luego juntando todo probamos que $z_i^T H z_i$ es decreciente y cumple la cota:

$$
z_{i+1}^T H z_{i+1} < z_i^T H z_i - \alpha \left\| \sum_{j \in S_i} \left(e_j^T H z_i\right) e_j \right\|_2^2 \tag{6}
$$

Por otro lado sabemos que para todo w vale:

$$
w^T H w \geq \lambda_{min}(H) \|w\|_2^2 \geq -L_b \|w\|_2^2 \tag{7}
$$

Usamos Cauchy Schwartz

Luego si usamos 4.5, 7 y Cauchy-Schwartz existe $i \in \{1, \ldots, b\}$ y $\delta > 0$ tal que:

$$
\begin{aligned}
z_{i+1}^T H z_{i+1} < & \; z_i^T H z_i - \alpha \sum_{j \in S_i} \left(e_j^T H z_i\right)^2 \\
< & \; z_i^T H z_i - \frac{\alpha}{d} \left(\sum_{j \in S_i} \left|e_j^T H z_i\right|\right)^2 \\
< & \; z_i^T H z_i - \frac{\delta^2}{d\alpha} \|z_i\|_2^2 \\
< & \; \left(1 + \frac{\delta^2}{d\alpha L_b}\right) z_i^T H z_i
\end{aligned}
$$

Tomando $\epsilon = \dfrac{\delta^2}{d\alpha L_b}$ probamos que $y_{t+1}^T H y_{t+1} \leq (1+\epsilon) y_t^T H y_t$ para $y_t \in Ran(H)$.

Si $y_t = y_N + y_R$ con $y_N \in Ker(H)$, $y_R \in Ran(H)$ entonces $y_t^T H y_t = y_R^T H y_R$ y $y_{t+1} = J y_t = y_N + J y_R$ por lo que $y_{t+1}^T H y_{t+1} = (J y_R)^T H (J y_R)$. Concluímos:

$$y_{t+1}^T H y_{t+1} = (J y_R)^T H (J y_R) \leq (1+\epsilon) y_R^T H y_R = (1+\epsilon) y_t^T H y_t$$

∎

Volviendo a la demostración general logramos probar que dado y_0 autovector de norma 1 de H con menor autovalor $\lambda < 0$ (pues $x^* \in \chi^*$) vale que:

$$\lambda_{min}(H) \|y_t\|_2^2 \leq y_t^T H y_t \leq (1+\epsilon)^t y_0^T H y_0 \leq (1+\epsilon)^t \lambda$$

Luego:

$$\|y_t\|_2^2 \geq \left(1 + \underbrace{\epsilon}_{<\frac{1}{2}}\right)^{\frac{t}{2}} \frac{\lambda}{\lambda_{min}(H)} \geq \frac{\lambda}{\lambda_{min}(H)} \left(1 + \frac{\epsilon}{4}\right)^t$$

Que era lo que queríamos demostrar con $c = \dfrac{\lambda}{\lambda_{min}(H)}$ y $\tilde{\epsilon} = \frac{\epsilon}{4}$.

∎

Corolario 4.7 *Sea g dado por el algoritmo de descenso por coordenadas con ecuación 4, bajo 9 y $\alpha < \frac{1}{L_b}$ se tiene que $\mu\left(W_g\right) = 0$.*

Demostración Por 4.3 y 4.4 tenemos que vale 3.4 y concluímos que $\mu\left(W_g\right) = 0$. ∎

Part III

Algoritmos Estocásticos

5

CONTEXTO

En esta parte vamos a analizar la convergencia de algoritmos estocásticos para optimizar una $F : \mathbb{R}^d \mapsto \mathbb{R}$ que puede representar tanto el costo esperado como el empírico. Recordemos que F lo asumimos parametrizado por$w \in \mathbb{R}^d$ e imaginamos a los datos (x, y) como extraídos de una variable aleatoria ξ, cuya distribución desconocida es P, luego F se representa como:

$$F(w) = \begin{cases} R(w) = & \mathbb{E}\left[f(w, \xi)\right] \\ & \text{o} \\ R_n(w) = & \frac{1}{n}\sum_{i=1}^{n} f_i(w) \end{cases} \tag{8}$$

Sea el algoritmo estocástico 2

Algorithmus 2 : Descenso Estocastico (DE)

Input: $w_1 \in \mathbb{R}^d$ el inicio de la iteración, $\{\xi_k\}$ iid
for $k \in \mathbb{N}$ **do**
    Generar una muestra de la variable aleatoria ξ_k
    Calcular el vector estocástico $g(w_k, \xi_k)$
    Elegir $\alpha_k > 0$
    $w_{k+1} \leftarrow w_k - \alpha_k g(w_k, \xi_k)$
end

Notemos que representa en forma general los algoritmos estocásticos mas comunes. En particular, una muestra de ξ_k puede ser un único par (x_i, y_i) como en el *Descenso por gradiente estocástico* o una muestra $S_n = \{(x_i, y_i)\}_{i \leq n}$ como en *Mini-Batch Descenso por gradiente estocástico* ; a su vez, $g(w_k, \xi_k)$ puede ser varias estimaciones del gradiente como por ejemplo:

$$g(w_k, \xi_k) = \begin{cases} \nabla f(w_k, \xi_k) \\ \frac{1}{n_k}\sum_{i=1}^{n_k} \nabla f(w_k, \xi_{k,i}) \\ H_k \frac{1}{n_k}\sum_{i=1}^{n_k} \nabla f(w_k, \xi_{k,i}) \end{cases} \tag{9}$$

Donde H_k es una matriz simétrica definida positiva como en los métodos de Newton-Gauss.

Para iniciar el analisis de la convergencia, lo mínimo que necesitamos es que el gradiente se mantenga controlado, por lo tanto asumamos:

Hipótesis 5.1 (*F* es *l*-Lipshitz) *La función a optimizar $F \in C^1(\mathbb{R}^d)$ y existe $L > 0$ tal que para todos $w, z \in \mathbb{R}^d$:*

$$\|\nabla F(w) - \nabla F(z)\|_2 \leq L \|w - z\|_2$$

Observación Sea F bajo 5.1, luego para todos $w, z \in \mathbb{R}^d$ vale:

$$F(w) \leq F(z) + \nabla F(z)^T (w - z) + \frac{1}{2} L \|w - z\|_2^2$$

Demostración Notemos que:

$$\begin{aligned}
F(w) = & \quad F(z) + \int_0^1 \frac{\partial F\left(z + t(w - z)\right)}{\partial t} dt \\
= & \quad F(z) + \int_0^1 \nabla F\left(z + t(w - z)\right)^T (w - z) dt \\
= & \quad F(z) + \nabla F(z)^T (w - z) + \int_0^1 \left[\nabla F\left(z + t(w - z)\right) - \nabla F(z)\right]^T (w - z) dt \\
\leq & \quad F(z) + \nabla F(z)^T (w - z) + \int_0^1 L \|t(w - z)\|_2 \|w - z\|_2 dt \\
= & \quad F(z) + \nabla F(z)^T (w - z) + \frac{1}{2} L \|w - z\|_2^2 . \qquad \blacksquare
\end{aligned}$$

Definamos ahora $\mathbb{E}_{\xi_k}[.] := \mathbb{E}_{P_k}[.|w_k]$ la esperanza condicional bajo la distribución de ξ_k dado w_k.

Lema 5.2 *Bajo 5.1 las iteraciones de 2 satisfacen que para todo $k \in N$:*

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leq -\alpha_k \nabla F(w_k)^T \mathbb{E}_{\xi_k}[g(w_k, \xi_k)] + \frac{1}{2} \alpha_k^2 \mathbb{E}_{\xi_k}\left[\|g(w_k, \xi_k)\|_2^2\right] \tag{10}$$

Demostración Notemos que por 5.1 vale que:

$$\begin{aligned}
F(w_{k+1}) - F(w_k) \leq & \quad \nabla F(w_k)^T (wk + 1 - w_k) + \frac{1}{2} L \|wk + 1 - w_k\|_2^2 \\
\leq & \quad -\alpha_k \nabla F(w_k)^T g(w_k, \xi_k) + \frac{1}{2} \alpha_k^2 L \|g(w_k, \xi_k)\|_2^2
\end{aligned}$$

Aca usamos propiedades basicas de la esperanza condicional

Luego tomando esperanza de ambos lados y recordando que si X, Y son independientes entonces $\mathbb{E}_{X,Y}[Y|X] = \mathbb{E}[Y]$:

$$\begin{aligned}
\mathbb{E}_{\xi_k}[F(w_{k+1}) - F(w_k)] \leq & \quad -\alpha_k \mathbb{E}_{\xi_k}\left[\nabla F(w_k)^T g(w_k, \xi_k)\right] + \frac{1}{2} \alpha_k^2 L \mathbb{E}_{\xi_k}\left[\|g(w_k, \xi_k)\|_2^2\right] \\
\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leq & \quad -\alpha_k \nabla F(w_k)^T \mathbb{E}_{\xi_k}[g(w_k, \xi_k)] + \frac{1}{2} \alpha_k^2 L \mathbb{E}_{\xi_k}\left[\|g(w_k, \xi_k)\|_2^2\right]
\end{aligned}$$

$\blacksquare$

Observación Notemos que si $g(w_k, \xi_k)$ es un estimador insesgado de $\nabla F(w_k)$ entonces de 5.2:

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leq -\alpha_k \|\nabla F(w_k)\|^2 + \frac{1}{2}\alpha_k^2 \mathbb{E}_{\xi_k}\left[\|g(w_k, \xi_k)\|_2^2\right] \tag{11}$$

Luego entonces para controlar la convergencia de 2 también hay que poner suposiciones sobre el segundo momento de g, luego si definimos:

$$\mathbb{V}_{\xi_k}[g(w_k, \xi_k)] := \mathbb{E}_{\xi_k}\left[\|g(w_k, \xi_k)\|_2^2\right] - \|\mathbb{E}_{\xi_k}[g(w_k, \xi_k)]\|_2^2 \tag{12}$$

Asumamos:

Hipótesis 5.3 (Acotaciones al primer y segundo momento de g) *Supongamos que dada F función objetivo y g la estimación del gradiente en 2 vale:*

1. *Existe $U \subset \mathbb{R}^d$ tal que $\{w_k\} \subset U$ y que existe F_{inf} tal que $F|_U \geq F_{inf}$*
2. *Existen $\mu_G \geq \mu \geq 0$ tal que para todo $k \in N$ valen:*

$$\nabla F(w_k)^T \mathbb{E}_{\xi_k}[g(w_k, \xi_k)] \geq \mu \|\nabla F(w_k)\|_2^2 \tag{13a}$$

Y

$$\|\mathbb{E}_{\xi_k}[g(w_k, \xi_k)]\|_2 \leq \mu_G \|\nabla F(w_k)\|_2 \tag{13b}$$

3. *Existen $M, M_V \geq 0$ tal que para todo $k \in \mathbb{N}$:*

$$\mathbb{V}_{\xi_k}[g(w_k, \xi_k)] \leq M + M_V \|\nabla F(w_k)\|_2^2 \tag{14}$$

Observación Notemos que si g es un estimador insesgado de ∇F entonces 13a y 13b valen con $\mu_G = \mu = 1$. Dejamos de ejercicio al lector notar que si H_k es simétrica positiva definida tal que H_k es independiente de ξ_k entonces tanto 13a como 13b valen.

Observación Bajo 5.3 y por 12 tenemos que:

$$\begin{aligned}\mathbb{E}_{\xi_k}\left[\|g(w_k, \xi_k)\|_2^2\right] &\leq \|\mathbb{E}_{\xi_k}[g(w_k, \xi_k)]\|_2^2 + M + M_V \|\nabla F(w_k)\|_2^2 \\ &\leq M + M_G \|\nabla F(w_k)\|_2^2\end{aligned}$$

$M_G := M_V + \mu_G^2 \geq \mu^2 \geq 0$

Lema 5.4 *Bajo 5.3 y 5.1 las iteraciones de 2 satisfacen para todo $k \in \mathbb{N}$:*

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leq -\mu\alpha_k \|\nabla F(w_k)\|_2^2 + \frac{1}{2}\alpha_k^2 L \mathbb{E}_{\xi_k}\left[\|g(w_k, \xi_k)\|_2^2\right] \tag{15a}$$

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leq -\left(\mu - \frac{1}{2}\alpha_k L M_G\right)\alpha_k \|\nabla F(w_k)\|_2^2 + \frac{1}{2}\alpha_k^2 L M \tag{15b}$$

Demostración Por 5.2 y 13a vale que:

$$\begin{aligned}
\mathbb{E}_{\xi_k}\left[F(w_{k+1})\right] - F(w_k) \leq & \quad -\alpha_k \nabla F(w_k)^T \mathbb{E}_{\xi_k}\left[g(w_k, \xi_k)\right] + \frac{1}{2} L \alpha_k^2 \mathbb{E}_{\xi_k}\left[\|g(w_k, \xi_k)\|_2^2\right] \\
\mathbb{E}_{\xi_k}\left[F(w_{k+1})\right] - F(w_k) \leq & \quad -\mu\alpha_k \|\nabla F(w_k)\|_2^2 + \frac{1}{2}\alpha_k^2 L \mathbb{E}_{\xi_k}\left[\|g(w_k, \xi_k)\|_2^2\right]
\end{aligned}$$

Que es 15a; luego por 7 obtenemos 15b. ∎

Corolario 5.5 *Bajo 5.3 y 5.1 las iteraciones de 2 satisfacen para todo $k \in \mathbb{N}$ que $\{w_k\}$ es una cadena de Markov de primer orden.*

6

CONVERGENCIA EN L1

6.1 CASO FUERTEMENTE CONVEXO

Consideremos primero los casos de convexidad donde sabemos que el mínimo existe y es único, por lo tanto asumamos por ahora:

Hipótesis 6.1 (Convexidad fuerte) *Supongamos que la función objetivo* $F : \mathbb{R}^d \mapsto \mathbb{R}$ *cumple que existe* $c > 0$ *tal que para todos* $z, w \in \mathbb{R}^d$*:*

$$F(z) \geq F(w) + \nabla F(w)^T (z - w) + \frac{1}{2} c \|z - w\|_2^2 \tag{16}$$

Luego existe un único $w_* \in \mathbb{R}^d$ *tal que* $F_{inf} = F(w_*) \leq F(w)$ *para todo* $w \in \mathbb{R}^d$

Notemos que de 6.1 y 7 vale que $c \leq L$

Lema 6.2 *Supongamos que F cumple 6.1, luego para todo* $w \in \mathbb{R}^d$ *vale que:*

$$2c \left(F(w) - F_{inf}\right) \leq \|\nabla F(w)\|_2^2 \tag{17}$$

Demostración Dado $w \in \mathbb{R}^d$ sea:

$$q(z) = F(w) + \nabla F(w)^T (z - w) + \frac{1}{2} c \|z - w\|_2^2$$

Se puede verificar que $z_* := w - \frac{1}{c} \nabla F(w)$ cumple que $q(z_*) = F(w) - \frac{1}{2c} \|\nabla F(w)\|_2^2 \leq q(z)$ para todo $z \in \mathbb{R}^d$; luego por 6.2 se tiene:

$$F_{inf} \geq F(w) + \nabla F(w)^T (w_* - w) + \frac{1}{2} c \|w_* - w\|_2^2 \geq F(w) - \frac{1}{2c} \|\nabla F(w)\|_2^2$$

∎

Ya estamos en condiciones de demostrar nuestro primer resultado de convergencia para 2 con $\alpha_k = \alpha$, pero notemos que *a priori* lo mas que podemos asumir es quedar en un entorno de F_{inf} ya que de 15b se ve que el segundo término es constante.

Dado w_k que depende de $\xi_1, \ldots, \xi_{k-1}$ definamos:

$$\mathbb{E}\left[F(w_k)\right] = \mathbb{E}_{\xi_1} \mathbb{E}_{\xi_2} \ldots \mathbb{E}_{\xi_{k-1}} \left[F(w_k)\right]$$

Teorema 6.3 (Objetivo fuertemente convexo, Incremento constante) *Supongamos 5.1, 5.3 y 6.1; además supongamos que dado 2* $\alpha_k = \alpha > 0$ *constante tal que:*

$$0 < \alpha \leq \frac{\mu}{LM_G} \tag{18}$$

Luego para todo $k \in \mathbb{N}$ vale que:

$$\begin{array}{rcl} \mathbb{E}\left[F(w_k) - F_{inf}\right] & \leq & \frac{\alpha LM}{2c\mu} + (1 - \alpha c\mu)^{k-1}\left(F(w_1) - F_{inf} - \frac{\alpha LM}{2c\mu}\right) \\ & \xrightarrow{k \to \infty} & \frac{\alpha LM}{2c\mu} \end{array}$$

Demostración Usando 5.4 con 18 y 6.2 tenemos para todo $k \in \mathbb{N}$ que:

$$\begin{array}{rcl} \mathbb{E}_{\xi_k}\left[F(w_{k+1}) - F(w_k)\right] & \underbrace{\leq}_{5.4} & -\left(\mu - \frac{1}{2}\alpha LM_G\right)\alpha \left\|\nabla F(w_k)\right\|_2^2 + \frac{1}{2}\alpha^2 LM \\ & \underbrace{\leq}_{18} & -\frac{1}{2}\alpha\mu \left\|\nabla F(w_k)\right\|_2^2 + \frac{1}{2}\alpha^2 LM \\ & \underbrace{\leq}_{6.2} & -\alpha\mu c\left(F(w_k) - F_{inf}\right) + \frac{1}{2}\alpha^2 LM \end{array}$$

Luego si restamos F_{inf} y tomamos esperanza total (definida en 6.1:

$$\begin{array}{rcl} \mathbb{E}_{\xi_k}\left[F(w_{k+1}) - F(w_k)\right] & \leq & -\alpha\mu c\left(F(w_k) - F_{inf}\right) + \frac{1}{2}\alpha^2 LM \\ \implies \mathbb{E}\left[F(w_{k+1}) - F_{inf}\right] & \leq & (1 - \alpha c\mu)\,\mathbb{E}\left[F(w_k) - F_{inf}\right] + \frac{1}{2}\alpha^2 LM \\ \implies \mathbb{E}\left[F(w_{k+1}) - F_{inf}\right] - \frac{\alpha LM}{2c\mu} & \leq & (1 - \alpha c\mu)\,\mathbb{E}\left[F(w_k) - F_{inf}\right] + \frac{1}{2}\alpha^2 LM - \frac{\alpha LM}{2c\mu} \\ & = & (1 - \alpha c\mu)\left(\mathbb{E}\left[F(w_k) - F_{inf}\right] - \frac{\alpha LM}{2c\mu}\right) \end{array}$$

Por otro lado notemos que:

$$0 < \alpha c\mu \leq \frac{c\mu^2}{LM_G} \leq \frac{c\mu^2}{L\mu^2} = \frac{c}{L} \leq 1$$

Luego deducimos inductivamente que:

$$\mathbb{E}\left[F(w_{k+1}) - F_{inf}\right] - \frac{\alpha LM}{2c\mu} \leq (1 - \alpha c\mu)^k\left(F(w_1) - F_{inf} - \frac{\alpha LM}{2c\mu}\right)$$

∎

Observación Notemos que si g es un estimador insesgado de ∇F entonces $\mu = M_G = 1$ por lo que $\alpha \in [0, \frac{1}{L})$ que es la condición que pedimos en 4.3.

Observación Notemos además que si $M = 0$ (o sea el algortimo 2 no tiene ruido) entonces la convergencia es lineal, recuperando el resultado de [2].

Observación Notemos finalmente que hay un compromiso entre el primer y segundo término de 6.3 donde a un α más cercano a $\frac{\mu}{LM_G}$ acelera la convergencia del primer término, pero a costa de un entorno final de mayor volúmen.

Luego esto llevo a varios investigadores a tomar un enfoque artesanal donde se tomaba un $\alpha_k = \alpha_1$ para $k \leq k_1$ donde k_1 es tal que $\mathbb{E}\left[F(w_{k_1}) - F_{inf}\right] \leq \frac{\alpha_1 LM}{2c\mu}$. Luego se tomaba $\alpha_2 = \frac{\alpha_1}{2}$ y se seguía inductivamente.

Teorema 6.4 (Objetivo fuertemente convexo, Incremento decreciente) *Supongamos 5.1, 5.3 y 6.1; además supongamos que dado 2 α_k cunmple:*

$$\alpha_k = \frac{\beta}{\gamma + k} \quad \text{para algún } \beta > \frac{1}{c\mu} \text{ y } \gamma > 0 \text{ tal que } \alpha_1 \leq \frac{\mu}{LM_G} \tag{19}$$

Luego para todo $k \in \mathbb{N}$ vale que:

$$\mathbb{E}\left[F(w_k) - F_{inf}\right] \leq \frac{\eta}{\gamma + k}$$

Donde:

$$\eta := \max\left\{\frac{\beta^2 LM}{2\left(\beta c\mu - 1\right)}, (\gamma + 1)\left(F(w_1) - F_{inf}\right)\right\}$$

Demostración Notemos primero que por 19 para todo $k \in \mathbb{N}$ vale:

$$\alpha_k LM_G \leq \alpha_1 LM_G \leq \mu$$

Luego por 5.4 y 6.2 uno tiene para todo $k \in \mathbb{N}$:

$$\begin{aligned}
\mathbb{E}_{\xi_k}\left[F(w_{k+1})\right] - F(w_k) \leq & -\left(\mu - \frac{1}{2}\alpha_k LM_G\right)\alpha_k \left\|\nabla F(w_k)\right\|_2^2 + \frac{1}{2}\alpha_k^2 LM \\
\leq & -\frac{1}{2}\mu\alpha_k \left\|\nabla F(w_k)\right\|_2^2 + \frac{1}{2}\alpha_k^2 LM \\
\leq & -\alpha_k c\mu\left(F(w_k) - F(w_*)\right) + \frac{1}{2}\alpha_k^2 LM
\end{aligned}$$

Luego restando F_{inf}, tomando esperanza y reordenando vale:

$$\mathbb{E}\left[F(w_{k+1}) - F_{inf}\right] \leq (1 - \alpha_k c\mu)\,\mathbb{E}\left[F(w_k) - F_{inf}\right] + \frac{1}{2}\alpha_k^2 LM$$

Probemos ahora el resultado por inducción. Por la definición de η tenemos que $k = 1$ vale, luego si asumimos que vale el resultado para algún $k \geq 1$ entonces:

Porque vale k=1?

$$\begin{aligned}
\mathbb{E}\left[F(w_{k+1}) - F_{inf}\right] &\leq \left(1 - \frac{\beta c \mu}{\gamma + k}\right) \frac{\eta}{\gamma + k} + \frac{\beta^2 LM}{2(\gamma + k)^2} \\
&= \left(\frac{(\gamma + k) - \beta c \mu}{(\gamma + k)^2}\right) \eta + \frac{\beta^2 LM}{2(\gamma + k)^2} \\
&= \left(\frac{(\gamma + k) - 1}{(\gamma + k)^2}\right) \eta \underbrace{- \left(\frac{\beta c \mu - 1}{(\gamma + k)^2}\right) \eta + \frac{\beta^2 LM}{2(\gamma + k)^2}}_{\leq 0 \text{ Por definición de } \eta} \\
&\underbrace{\leq}_{(\gamma+k)^2 \geq (\gamma+k+1)(\gamma+k-1)} \frac{\eta}{\gamma + k + 1}
\end{aligned}$$

∎

Notemos entonces que en el caso fuertemente convexo con incrementos fijos tenemos convergencia en un entorno del mínimo mientras que si reducimos los incrementos tenemos convergencia en L1, cabría preguntarse (inspirados en la observación del caso α fijo con $M = 0$) si con el ruido existente pero controlado podemos mantener la convergencia en L1.

Teorema 6.5 (Objetivo Fuertemente Convexo, Reducción del Ruido) *Supongamos que valen 5.1, 5.3 y 6.1 pero reforcemos 14 a la existencia de una constante $M \geq 0$ y $\xi \in (0, 1)$ tal que para todo $k \in \mathbb{N}$:*

$$\mathbb{V}_{\xi_k}\left[g(w_k, \xi_k)\right] \leq M \xi^{k-1} \tag{20}$$

Supongamos además que 2 tiene $\alpha_k = \alpha$ para todo $k \in \mathbb{N}$ satisfaciendo:

$$0 < \alpha \leq \min\left\{\frac{\mu}{L\mu_G^2}, \frac{1}{\mu}\right\} \tag{21}$$

Luego vale:

$$\mathbb{E}\left[F(w_k) - F_{inf}\right] \leq \omega \rho^{k-1}$$

Donde:

$$\omega := \max\left\{\frac{\alpha LM}{c\mu}, F(w_1) - F_{inf}\right\} \tag{22a}$$

$$\rho := \max\left\{1 - \frac{\alpha c \mu}{2}, \xi\right\} < 1 \tag{22b}$$

6.2 CASO GENERAL

Manteniendo las mismas hipótesis y notaciones veamos el caso general, nuevamente separando entre incrementos constantes o decrecientes.

Teorema 6.1 (Objetivo no convexo, Incrementos fijos) *Asumiendo 5.1 y 5.3 y suponiendo que en 2 tenemos $\alpha_k = \alpha$ tal que:*

$$0 < \alpha \leq \frac{\mu}{LM_G} \tag{23}$$

Entonces vale para todo $k \in \mathbb{N}$:

$$\mathbb{E}\left[\sum_{k=1}^{K} \|\nabla F(w)_k\|_2^2\right] \leq \frac{K\alpha LM}{\mu} + \frac{2(F(w_1) - F_{inf})}{\mu\alpha} \tag{24a}$$

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K} \|\nabla F(w)_k\|_2^2\right] \leq \frac{\alpha LM}{\mu} + \frac{2(F(w_1) - F_{inf})}{K\mu\alpha} \xrightarrow{K \to \infty} \frac{\alpha LM}{\mu} \tag{24b}$$

Demostración Recordemos 15a y si tomamos esperanza total e imponemos 23 tenemos:

$$\begin{aligned}\mathbb{E}\left[F(w_{k+1})\right] - \mathbb{E}\left[F(w_k)\right] &\leq -\left(\mu - \tfrac{1}{2}\alpha LM_G\right)\alpha\mathbb{E}\left[\|\nabla F(w_k)\|_2^2\right] + \tfrac{1}{2}\alpha^2 LM \\ &\leq -\tfrac{1}{2}\alpha\mu\mathbb{E}\left[\|\nabla F(w_k)\|_2^2\right] + \tfrac{1}{2}\alpha^2 LM\end{aligned}$$

Luego como por 5.3 tenemos que $F_{inf} \leq \mathbb{E}\left[F(w_k)\right]$ para todo $k \in \mathbb{N}$ vale:

$$F_{inf} - F(w_1) \leq \mathbb{E}\left[F(w_{K+1})\right] - F(w_1) \leq -\frac{1}{2}\alpha\mu\sum_{1}^{K}\mathbb{E}\left[\|\nabla F(w_k)\|_2^2\right] + \frac{1}{2}K\alpha^2 LM$$

∎

Observación Notemos que si $M = 0$ (no hay ruido o crece comparable a $\|\nabla F(w_k)\|_2^2$) entonces obtenemos que $\sum_{k=1}^{\infty} \|\nabla F(w)_k\|_2^2 < \infty$ por lo que $\left\{\|\nabla F(w_k)\|_2^2\right\}_{k\in\mathbb{N}} \xrightarrow{k \to \infty} 0$, que es el resultado obtenido en [2].

En cambio, cuando $M \neq 0$ aunque no podemos acotar $\|\nabla F(w_k)\|_2^2$ *per-se*, podemos decir de 24b que en esperanza el valor del gradiente es cada vez menor en un entorno de radio $\frac{\alpha LM}{\mu}$. Luego recuperamos la intuición del caso convexo (6.3) donde a menor incremento el entorno es menor (el algortimo es más preciso) pero la cantidad de iteraciones es mayor.

Para el de incrementos decrecientes, asumamos que $\{\alpha_k\}$ cumple la condición de *Robbins - Monro* (ver [3]):

$$\sum_{k=1}^{\infty} \alpha_k = \infty \quad \text{y} \quad \sum_{k=1}^{\infty} \alpha_k^2 < \infty \tag{25}$$

Teorema 6.2 (Objetivo no convexo, Incrementos decrecientes) *Asumiendo 5.1 y 5.3, suponiendo además que en 2 los $\{\alpha_k\}$ satisfacen 25; si notamos $A_K := \sum_{k=1}^{K} \alpha_k$ vale para todo $k \in \mathbb{N}$::*

$$\lim_{K\to\infty} \mathbb{E}\left[\sum_{k=1}^{K} \alpha_k \left\|\nabla F(w)_k)\right\|_2^2\right] < \infty \tag{26a}$$

$$\mathbb{E}\left[\frac{1}{A_K}\sum_{k=1}^{K} \alpha_k \left\|\nabla F(w)_k)\right\|_2^2\right] \xrightarrow{K\to\infty} 0 \tag{26b}$$

Demostración Como $\alpha_k \to 0$ por 25 entonces podemos asumir sin pérdida de generalidad que $\alpha_k L M_G \leq \mu$ para todo $k \in \mathbb{N}$, luego:

$$\begin{aligned}\mathbb{E}\left[F(w_{k+1})\right] - \mathbb{E}\left[F(w_k)\right] &\leq -\left(\mu - \tfrac{1}{2}\alpha_k L M_G\right)\alpha_k \mathbb{E}\left[\left\|\nabla F(w_k)\right\|_2^2\right] + \tfrac{1}{2}\alpha_k^2 L M \\ &\leq -\tfrac{1}{2}\alpha_k \mu \mathbb{E}\left[\left\|\nabla F(w_k)\right\|_2^2\right] + \tfrac{1}{2}\alpha_k^2 L M\end{aligned}$$

Luego como por 5.3 tenemos que $F_{inf} \leq \mathbb{E}\left[F(w_k)\right]$ para todo $k \in \mathbb{N}$ vale:

$$F_{inf} - \mathbb{E}\left[F(w_1)\right] \leq \mathbb{E}\left[F(w_{K+1})\right] - \mathbb{E}\left[F(w_1)\right] \leq -\frac{1}{2}\mu \sum_{1}^{K} \alpha_k \mathbb{E}\left[\left\|\nabla F(w_k)\right\|_2^2\right] + \frac{1}{2} L M \sum_{k=1}^{K} \alpha_k^2$$

Luego:

$$\sum_{1}^{K} \alpha_k \mathbb{E}\left[\left\|\nabla F(w_k)\right\|_2^2\right] \leq \frac{2(\mathbb{E}\left[F(w_1) - F_{inf}\right])}{\mu} + \frac{LM}{\mu} \underbrace{\sum_{k=1}^{K} \alpha_k^2}_{\xrightarrow{K\to\infty} C < \infty}$$

Porque si $\lim \mathbb{E}\left[X_k\right] < \infty$ entonces $\lim \mathbb{E}\left[\frac{X_k}{A_k}\right] = 0$ si $A_k \to \infty$ con A_k escalar ?

Por lo que 26a esta probado. Finalmente como por 25 tenemos que $A_K \to \infty$ se tiene 26b. ∎

Corolario 6.3 *Asumiendo 5.1 y 5.3, suponiendo además que en 2 los $\{\alpha_k\}$ satisfacen 25 entonces :*

$$\liminf_{k\to\infty} \mathbb{E}\left[\left\|\nabla F(w_k)\right\|_2^2\right] = 0 \tag{27}$$

Corolario 6.4 *Bajo las mismas hipótesis de 6.2 sea $k(K) \in \{1, \ldots, K\}$ un índice aleatorio elegido con probabilidades respectivas $\{\alpha_k\}_{k=1}^{K}$; luego $\left\|\nabla F(w_k)\right\|_2 \to 0$ en probabilidad.*

Usamos la desigualdad de Markov

Demostración Sea $\epsilon > 0$, luego de 26a y la desigualdad de Markov:

$$\mathbb{P}\left[\left\|\nabla F(w_k)\right\|_2 \geq \epsilon\right] = \mathbb{P}\left[\left\|\nabla F(w_k)\right\|_2^2 \geq \epsilon^2\right] \leq \epsilon^{-2}\mathbb{E}\left[\mathbb{E}_{\xi_k}\left[\left\|\nabla F(w_k)\right\|_2^2\right]\right] \to 0$$

Porque aca es la esperanza de la esperanza condicional?

∎

.

Teorema 6.5 (Objetivo no convexo regular, Incrementos decrecientes) *Bajo las mismas hipótesis de 6.2 si además pedimos que $F \in C^2$ y que $w \mapsto \|\nabla F(w)\|_2^2$ sea $l-$ Lipshitz entonces:*

$$\lim_{k\to\infty} \mathbb{E}\left[\|\nabla F(w_k)\|_2^2\right] = 0 \tag{28}$$

Demostración Sea $G(w) := \|\nabla F(w)\|_2^2$ y sea L_G la constante de Lipshitz de $\nabla G(w) = 2\nabla^2 F(w)\nabla F(w)$, luego:

$$\begin{aligned} G(w_{k+1}) - G(w_k) \underbrace{\leq}_{7} & \quad \nabla G(w_k)^T(w_{k+1} - w_k) + \frac{1}{2}L_G \|w_k - w_{k+1}\|_2^2 \\ \leq & \quad -\alpha_k \nabla G(w_k)^T g(w_k, \xi_k) + \frac{1}{2}\alpha_k L_G \|g(w_k, \xi_k)\|_2^2 \end{aligned}$$

Si tomamos esperanza condicional a ξ_k y usamos 5.1, 5.3 entonces:

$$\begin{aligned} \mathbb{E}_{\xi_k}\left[G(w_{k+1}) - G(w_k)\right] \leq & \quad -2\alpha_k \nabla F(w_k)^T \nabla^2 F(w_k)^T \mathbb{E}_{\xi_k}\left[g(w_k, \xi_k)\right] + \\ & \quad \frac{1}{2}\alpha_k^2 L_G \mathbb{E}_{\xi_k}\left[\|g(w_k, \xi_k)\|_2^2\right] \\ \leq & \quad 2\alpha_k \|\nabla F(w_k)\|_2 \left\|\nabla^2 F(w_k)\right\|_2 \|\mathbb{E}_{\xi_k}\left[g(w_k, \xi_k)\right]\|_2 + \\ & \quad \frac{1}{2}\alpha_k^2 L_G \mathbb{E}_{\xi_k}\left[\|g(w_k, \xi_k)\|_2^2\right] \\ \leq & \quad 2\alpha_k L \mu_G \|\nabla F(w_k)\|_2^2 + \\ & \quad \frac{1}{2}\alpha_k^2 L_G \left(M + M_V \|\nabla F(w_k)\|_2^2\right) \end{aligned}$$

Luego obtenemos tomando esperanza total:

$$\mathbb{E}\left[G(w_{k+1})\right] - \mathbb{E}\left[G(w_k)\right] \leq 2\alpha_k L \mu_G \mathbb{E}\left[\|\nabla F(w_k)\|_2^2\right] + \frac{1}{2}\alpha_k^2 L_G \left(M + M_V \mathbb{E}\left[\|\nabla F(w_k)\|_2^2\right]\right) \tag{29}$$

Notemos que existe $K \in \mathbb{N}$ tal que $\alpha_k^2 \leq \alpha_k$ y luego por 6.2 el lado derecho cumple:

$$\lim_{N\to\infty} 2L\mu_G \underbrace{\sum_{k=K}^{K+N} \mathbb{E}\left[\alpha_k \|\nabla F(w_k)\|_2^2\right]}_{26a} + \frac{1}{2}L_G \left(M \underbrace{\sum_{k=K}^{K+N} \alpha_k^2}_{25} + M_V \underbrace{\sum_{k=K}^{K+N} \mathbb{E}\left[\alpha_k^2 \|\nabla F(w_k)\|_2^2\right]}_{26a} \right) = 0$$

Sean:

$$\begin{aligned} S_K^+ = & \quad \sum_{k=1}^{K} \max\left(0, \mathbb{E}\left[G(w_{k+1})\right] - \mathbb{E}\left[G(w_k)\right]\right) \\ S_K^- = & \quad \sum_{k=1}^{K} \max\left(0, \mathbb{E}\left[G(w_k)\right] - \mathbb{E}\left[G(w_{k+1})\right]\right) \end{aligned}$$

Luego como en 29 el lado derecho es positivo y su suma es convergente tenemos que $\{S_K^+\}$ es monótona, acotada superiormente y por ende convergente. Además como $0 \leq \mathbb{E}[G(w_k)] = \mathbb{E}[G(w_0)] + S_k^+ - S_k^-$ tenemos que $\{S_K^-\}$ también es monótona y acotada superiormente, por lo que es convergente; concluímos que $\mathbb{E}[G(w_k)]$ debe ser convergente, y por 6.3 tenemos $\mathbb{E}\left[\|\nabla F(w_k)\|_2^2\right] = \mathbb{E}[G(w_k)] \to 0$.

∎

Part IV

Apéndice

A

APÉNDICE

[1] Donald E. Knuth. «Computer Programming as an Art.» In: *Communications of the ACM* 17.12 (1974), pp. 667–673.

[2] Yurii Nesterov. *Introductory Lectures on Convex Optimization*. Vol. 87. Springer Science & Business Media, 2004.

[3] H. Robbins and S. Monro. *A Stochastic Approximation Model*. Vol. 22(3). The Annal of Mathematical Statistics, 1951, pp. 400–407.

[4] Krizhevsky et al. «Imagenet classification with deep convolutional neural networks.» In: (2012).

[5] Lee et al. *Gradient descent only converges to minimizers*. Conference on learning theory, 2016, pp. 1246–1257.