



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



## 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

---

<b>I</b>	<b>Introducción</b>	<b>1</b>
1	INTRODUCCIÓN	3
<b>II</b>	<b>Algoritmos de tipo Batch</b>	<b>5</b>
2	INTUICIÓN	7
3	TEOREMA DE LA VARIEDAD ESTABLE Y LOS PUNTOS FI- JOS INESTABLES	11
3.1	Resultados previos	11
3.2	Puntos fijos inestables	11
4	APLICACIONES	15
4.1	Gradient Descent	15
4.2	Punto Próximo	15
4.3	Descenso por coordenadas	16
<b>III</b>	<b>Algoritmos Estocásticos</b>	<b>21</b>
<b>IV</b>	<b>Apéndice</b>	<b>23</b>
A	APÉNDICE	25
	NEW NAME	27

## LIST OF FIGURES

---

## LIST OF TABLES

---

## LISTINGS

---

## ACRÓNIMOS

---



## Part I

# Introducción



## INTRODUCCIÓN

---

De lo dicho en [2] y [3]



## 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



## 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  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  dada por  $f(x) = \frac{1}{2}x^T H x$  con  $H = \mathbf{diag}(\lambda_1, \dots, \lambda_n)$ ; supongamos además que  $\lambda_1, \dots, \lambda_k > 0$  y  $\lambda_{k+1}, \dots, \lambda_n < 0$ .

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

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

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:

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

Por lo tanto por inducción es fácil probar que:

$$x_{t+1} = (1 - \alpha \lambda_i)^t \langle x_0, e^i \rangle e^i$$

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 concluimos que:

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

Finalmente, si  $k < n$  entonces concluimos 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:

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

De lo que vemos que los puntos críticos son:

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

Y del criterio del Hessiano concluimos 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 2.0.1** Para  $f(x, y) = \frac{1}{2}x^2 + \frac{1}{4}y^4 - \frac{1}{2}y^2$  resulta que  $E_s = \langle t * e^1 / t \in \mathbb{R} \rangle := W_s$

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:

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

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

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

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



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

Si analizamos simplemente los signos de  $g$  y  $\frac{dg_y}{dy}$  en los otros intervalos podemos concluir 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  $(x, y) \in E_s \iff (x, y) = (t, 0) \ t \in \mathbb{R} \iff (x, y) \in W_s$ . ■



## TEOREMA DE LA VARIEDAD ESTABLE Y LOS PUNTOS FIJOS INESTABLES

### 3.1 RESULTADOS PREVIOS

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

*Esto quizás debería ir en prerequisites cuando lo tengamos*

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

**Lema 3.1.1** Sea  $E \subset \chi$  tal que  $\mu(E) = 0$ ; si  $\det(Dg(x)) \neq 0$  para todo  $x \in \chi$ , luego  $\mu(g^{-1}(E)) = 0$

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

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

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

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

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

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

### 3.2 PUNTOS FIJOS INESTABLES

**Definición** Sea:

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

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 prerequisites*

**Teorema 3.2.1** Sea  $x^*$  un punto fijo de  $g \in C^r(\chi)$  un difeomorfismo local. Supongamos que  $E = E_s \oplus E_u$  donde

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

Entonces existe  $W_{loc}^{cs} \hookrightarrow \chi$  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  $g(W_{loc}^{cs}) \cap B \subseteq W_{loc}^{cs}$  y  $\bigcap_{k \in \mathbb{N}} g^{-k}(B) \subseteq W_{loc}^{cs}$

Con todos estos resultados demostremos el teorema principal:

**Teorema 3.2.2** 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 3.2.2 existe  $B_{x^*}$  un entorno abierto; es más,  $\bigcup_{x^* \in \mathcal{A}_g^*} B_{x^*}$  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:

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

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

Sea  $x_0 \in W$ , luego como  $x_k \rightarrow x^* \in \mathcal{A}_g^*$  entonces existe  $T \in \mathbb{N}$  tal que para todo  $t \geq T$ ,  $x_t \in \bigcup_{i \in \mathbb{N}} B_{x_i^*}$  por lo que  $x_t \in B_{x_i^*}$  para algún

$x_i^* \in \mathcal{A}_g^*$  y  $t \geq T$ . Afirмо que:

**Lema 3.2.3**  $x_t \in \bigcap_{k \in \mathbb{N}} g^{-k}(B_{x_i^*})$  para todo  $t \geq T$

*Pablo: Hace falta demostrar esto??*

Si notamos  $S_i \triangleq \bigcap_{k \in \mathbb{N}} g^{-k}(B_{x_i^*})$ , entonces por 3.2.1 sabemos por un lado que es una subvariedad de  $W_{loc}^{cs}$  y por el otro que  $\dim(S_i) \leq \dim(W_{loc}^{cs}) = \dim(E_s) < d - 1$ <sup>1</sup>; por lo que  $\mu(S_i) = 0$ .

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  $W \subseteq \bigcup_{i \in \mathbb{N}} \bigcup_{k \in \mathbb{N}} g^{-k}(S_i)$ . Concluimos:

$$\begin{aligned} \mu(W) &\leq \mu\left(\bigcup_{i \in \mathbb{N}} \bigcup_{k \in \mathbb{N}} g^{-k}(S_i)\right) \\ &\leq \sum_{i \in \mathbb{N}} \sum_{k \in \mathbb{N}} \mu(g^{-k}(S_i)) \\ &\stackrel{3.1.1}{=} 0 \end{aligned}$$

■

Para finalizar veamos un caso simple que nos encontraremos seguido:

**Corolario 3.2.4** *Bajo las mismas hipótesis que en 3.2.2 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  $\mu(W_g) \leq \mu(W) \stackrel{3.2.2}{=} 0$ . ■

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

*Usamos que una subvariedad de dimension menor tiene medida 0*

<sup>1</sup> Por que???



## APLICACIONES

### 4.1 GRADIENT DESCENT

Como una aplicación del teorema en 3.2.2 demostremos que *gradient descent* tiene probabilidad cero de converger a puntos silla. Consideremos *gradient descent* con *learning rate*  $\alpha$ :

$$x_{k+1} = g(x_k) \triangleq x_k - \alpha \nabla f(x_k) \quad (2)$$

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

**Proposición 4.1.1** *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  $\{1 - \alpha \lambda_i : \lambda_i \in \{\mu : \nabla^2 f(x^*)v = \mu v \text{ para algún } v \neq 0\}\}$ . Como  $x^* \in \chi^*$  existe  $\lambda_{j^*} < 0$  por lo que  $1 - \alpha \lambda_{j^*} > 1$ ; concluimos que  $x^* \in \mathcal{A}_g^*$ . ■

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

**Proposición 4.1.2** *Bajo 4.1 y  $\alpha < \frac{1}{L}$  entonces  $\det(Dg(x)) \neq 0$ .*

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

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

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

**Corolario 4.1.3** *Gradient descent converge a mínimos Sea  $g$  dada por Gradient descent en 2, bajo 4.1 y  $\alpha < \frac{1}{L}$  se tiene que  $\mu(W_g) = 0$ .*

**Demostración** Por 4.1.1 y 4.1.2 tenemos que vale 3.2.4 y concluimos que  $\mu(W_g) = 0$ . ■

### 4.2 PUNTO PRÓXIMO

El algoritmo de punto próximo esta dado por la iteración:

$$x_{k+1} = g(x_k) \triangleq \arg \min_{z \in \mathcal{X}} f(z) + \frac{1}{2\alpha} \|x_k - z\|_2^2 \quad (3)$$

**Proposición 4.2.1** *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 4.2.2** 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 4.2.2 podemos tomar límite, i. e.

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

Finalmente por diferenciación implícita obtenemos:

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

Luego si  $x^* \in \chi^*$  entonces  $Dg(x^*) = (Id + \alpha \nabla^2 f(x^*))^{-1}$  y tiene autovalores  $\left\{ \frac{1}{1 + \alpha \lambda_i} \right\}$  con  $\lambda_i$  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 4.2.3** 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 4.2.1 tenemos que vale 3.2.4 y concluimos que  $\mu(W_g) = 0$ . ■

#### 4.3 DESCENSO POR COORDENADAS

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



Consideremos el algoritmo 1:

<b>Algorithmus 1</b> : Descenso por coordenadas	
1	<b>Input:</b> $f \in C^1$ , $\alpha > 0$ , $x_0 \in \chi$
2	<b>for</b> $k \in \mathbb{N}$ <b>do</b>
3	<b>for block</b> $i = 1, \dots, b$ <b>do</b>
4	<b>for index</b> $j \in S_i$ <b>do</b>
5	$y_k^{S_0} = x_k$ e $y_k^{S_i} = (x_{k+1}^{S_1}, \dots, x_{k+1}^{S_i}, x_k^{S_{i+1}}, \dots, x_k^{S_b})$
6	$x_{k+1}^j \leftarrow x_k^j - \alpha \frac{\partial f}{\partial x_j} (y_k^{S_{i-1}})$
7	<b>end</b>
8	<b>end</b>
9	<b>end</b>

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

**Lema 4.3.1** La iteración de Descenso por coordenadas esta dada por:

$$x_{k+1} = g(x_k) \triangleq g_d \circ g_{d-1} \circ \dots \circ g_1(x) \quad (4)$$

**Lema 4.3.2** Si  $g$  está dada por 4 entonces si notamos  $P_S = \sum_{i \in S} e_i e_i^T$  entonces:

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

**Demostración** Notemos primero que:

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

Por lo tanto:

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

■

**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  $\max_{i \in \{1, \dots, b\}} \|\nabla^2 f(x)|_{S_i}\| = L_b$

**Proposición 4.3.3** 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:

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

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

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

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

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

Y concluimos 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  $z_{i+1} = (Id - \alpha P_{S_i} H) z_i = z_i - \alpha \sum_{j \in S_i} (e_j^T H z_i) e_j$ . Luego  $y_{t+1} = z_{b+1}$ , afirmo:

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

**Afirmación 4.3.5** Sea  $y_t \in \text{Ran}(H)$ , luego existe  $i \in \{1, \dots, b\}$  y  $\delta > 0$  tal que  $\alpha \sum_{j \in S_i} |e_j^T H z_i| \geq \delta \|z_i\|_2$

**Lema 4.3.6** 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:

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

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

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

Por otro lado sabemos que para todo  $w$  vale:

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

Luego si usamos 4.3.5, 7 y Cauchy-Schwartz existe  $i \in \{1, \dots, b\}$  y  $\delta > 0$  tal que:

Usamos Cauchy  
Schwartz

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

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

Si  $y_t = y_N + y_R$  con  $y_N \in \text{Ker}(H)$ ,  $y_R \in \text{Ran}(H)$  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  $y_{t+1}^T H y_{t+1} = (J y_R)^T H (J y_R)$ .  
Concluimos:

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

■

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  $c = \frac{\lambda}{\lambda_{\min}(H)}$  y  $\tilde{\epsilon} = \frac{\epsilon}{4}$ .

■

**Corolario 4.3.7** Sea  $g$  dado por el algoritmo de descenso por coordenadas con ecuación 4, bajo 9 y  $\alpha < \frac{1}{L_b}$  se tiene que  $\mu(W_g) = 0$ .

**Demostración** Por 4.3.3 y 4.3.4 tenemos que vale 3.2.4 y concluimos que  $\mu(W_g) = 0$ . ■

Part III

# Algoritmos Estocásticos



Part IV

Apéndice







## APÉNDICE

---



- [1] Donald E. Knuth. «Computer Programming as an Art.» In: *Communications of the ACM* 17.12 (1974), pp. 667–673.
- [2] Krizhevsky et al. «Imagenet classification with deep convolutional neural networks.» In: (2012).
- [3] Lee et al. *Gradient descent only converges to minimizers*. Conference on learning theory, 2016, pp. 1246–1257.