MÉTODOS DE PRIMER ORDEN

ANÁLISIS DE CONVERGENCIA Y COMPLEJIDAD

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 [2]

AGRADECIMIENTOS

Agradecimientos para todos

CONTENTS

I Introducción 1

1 INTRODUCCIÓN 3

II Algoritmos de tipo Batch 5

2 CONVERGENCIA PUNTUAL 7
2.1 Intuición 7
2.2 Caso discreto 8
2.3 Acerca de convexidad fuerte y funciones L-Lipshitz 10

3 TEOREMA DE LA VARIEDAD ESTABLE Y LOS PUNTOS FIJOS INESTABLES 15
3.1 Intuición 15
3.2 Resultados previos 18
3.3 Puntos fijos inestables 18

4 CONVERGENCIA CTP A MÍNIMOS : CASO GENERAL 21
4.1 Descenso de Gradiente en Batch 21
4.2 Punto Próximo 22
4.3 Descenso por coordenadas 23

5 RESULTADOS NEGATIVOS 27
5.1 Ejemplos *patológicos* 27

III Algoritmos Estocásticos 35

6 CONVERGENCIA EN L1 37
6.1 Contexto 37
6.2 Algunos lemas fundamentales 38
6.3 Caso Fuertemente Convexo 40
6.4 Caso general 44

7 CONVERGENCIA CTP 49
7.1 Caso débilmente convexo 49
7.2 Caso no convexo 51
7.2.1 Acotación global del algoritmo 52
7.2.2 Convergencia del algoritmo 54

IV Apéndice 57

A APÉNDICE 59
A.1 Proposiciones enunciadas 59
A.2 Demostraciones 59

NEW NAME 61

LIST OF FIGURES

LIST OF TABLES

LISTINGS

ACRÓNIMOS

Part I

Introducción

1

INTRODUCCIÓN

> Mathematics knows no races or geographic boundaries; for mathematics, the cultural world is one country.
>
> "David Hilbert"

De lo dicho en [6] y [7]

Part II

Algoritmos de tipo Batch

En esta parte vamos a analizar los tipos de convergencia de los diferentes algoritmos de primer orden de tipo batch usados en Machine Learning. A su vez vamos a analizar casos donde aunque la convergencia este, no es útil computacionalmente

2 CONVERGENCIA PUNTUAL

> "The book of nature is written in the language of Mathematic"
>
> Galileo

Si existe un algoritmo que todo estudiante o practicante del Machine Learning conoce, es el descenso de gradiente clásico (o *Descenso de gradiente en batch*) [Ver algoritmo 2.1]. Un buen inicio es analizar la convergencia puntual del descenso de gradiente y bajo que condiciones se da.

Algorithmus 2.1 : Descenso de gradiente en batch

1 **Input:** $F \in C^1$, $\alpha_k > 0$, $w_1 \in \mathbb{R}^d$, $X = \{\xi_j\}_{j \leq N}$ muestra
2 **for** $k \in \mathbb{N}$ **do**
3 $\quad w_{k+1} \leftarrow w_k - \alpha_k \sum_{j=1}^{N} \nabla F(\xi_j)$

Asumamos por esta sección la siguiente condición, que llamaremos *convexidad débil*:

Definición 2.0.1 *Decimos que $F \in C^1$ es débilmente convexo si cumple las siguientes dos propiedades:*

- *Existe un único w^* tal que $F_{inf} := F(w^*) \leq F(w)$ para todo $w \in \mathbb{R}^n$.*
- *Para todo $\epsilon > 0$ vale que $\inf_{(w-w^*)^2 > \epsilon} (w - w^*) \nabla F(w) > 0$*

Observación Notemos que existen funciones no convexas tal que cumplen 2.0.1.

2.1 INTUICIÓN

Ganemos intuición acerca del proceso como probar la convergencia del algoritmo 2.1 en el caso continuo. En el caso continuo, tenemos que demostrar que la solución $w(t)$ de la ecuación diferencial 2.1 tiene límite w^* y además que w^* es mínimo de F.

$$\frac{dw}{dt} = -\nabla F(w) \tag{2.1}$$

Para eso, vamos a dividir la demostración en tres pasos:

Usamos funciones de Lyapunov

1. Vamos a definir una *función de Lyapunov*

2. Vamos a verificar computando su derivada temporal que es una función monótona decreciente y acotada, por lo que converge

3. Vamos a probar que converge a 0.

Proposición 2.1.1 (Objetivo débilmente convexo, Versión continua) *Sea $F \in C^1$ que cumple 2.0.1 y supongamos que el algoritmo 2.1 cumple $\alpha_k = \alpha > 0$ para $w(t)$ continua. Luego si notamos al mínimo de F como w^*, vale:*

$$\lim_{t \to \infty} w(t) = w_*$$

Demostración Vayamos con los pasos que definimos:

Paso 1 Definamos la función de Lyapunov:

$$h(t) = (w(t) - w^*)^2 \geq 0$$

Paso 2 Notemos que:

$$\frac{dh}{dt} = 2\left(w(t) - w^*\right)\frac{dw}{dt} = -2\left(w(t) - w^*\right)\nabla F(w) \underbrace{\leq}_{2.0.1} 0 \quad (2.2)$$

Luego como $h(t) \geq 0$ y $\dfrac{dh}{dt} \leq 0$ existe h_{inf} tal que $h(t) \searrow h_{inf}$

Paso 3 Como $h(t) \searrow h_{inf}$ entonces $\dfrac{dh}{dt} \to 0$, supongamos por el absurdo que $h_{inf} > 0$, luego existe $\widetilde{\epsilon} > 0$ y $T \in \mathbb{R}$ tal que para todo $t \geq T$ vale que $h(t) = (w(t) - w^*)^2 > \widetilde{\epsilon}$. Si juntamos entonces 2.2 y 2.0.1 llegamos a un absurdo, concluímos que $h_{inf} = 0$ por lo que:

$$w(t) \to w_*$$

∎

2.2 CASO DISCRETO

Ahora sí, analicemos la convergencia del algoritmo 2.1 para el caso discreto. Para esto enunciamos un lema útil cuya demostración referimos al lector al Apéndice:

Lema 2.2.1 *Sea $\{u_k\}_{k \in \mathbb{N}} \subset \mathbb{R}$ una sucesión tal que $u_k \geq 0$ para todo k. Luego si:*

$$\sum_{k=1}^{\infty} (u_{t+1} - u_t)_+ < \infty$$

Donde $(x)_{\pm} = x * 1_{\{\mathbb{R}_{\pm}\}}$, *entonces:*

$$\sum_{k=1}^{\infty} (u_{t+1} - u_t)_{-} < \infty$$

y (u_k) *converge.*

Es más, si notamos $S_{\infty}^{\pm} = \sum_{k=1}^{\infty} (u_{t+1} - u_t)_{\pm}$ *entonces* $u_{\infty} = \lim_{k \to \infty} u_k = u_0 + S_{\infty}^{+} + S_{\infty}^{-}$

Demostración Ver A

Consideremos ahora el algoritmo 2.1, decimos que los incrementos $\{\alpha_k\}$ cumplen la condición de *Robbins - Monro* (ver [5]) si:

$$\sum_{k=1}^{\infty} \alpha_k = \infty \quad \text{y} \quad \sum_{k=1}^{\infty} \alpha_k^2 < \infty \tag{2.3}$$

Teorema 2.2.2 (Objetivo débilmente convexo, incrementos decrecientes) *Sea* $F \in C^1$, *asumamos* 2.0.1 *y que existen* $A, B \geq 0$ *tal que para todo* $w \in \mathbb{R}^d$ *vale que:*

$$(\nabla F(w))^2 \leq A + B\,(w - w^*)^2 \tag{2.4}$$

Luego si consideramos el algoritmo 2.1 *tal que incrementos* $\{\alpha_k\}$ *cumplen* 2.3 *entonces:*

$$w_k \xrightarrow{k \to \infty} w^* \tag{2.5}$$

Demostración Hagamos los 3 pasos análogos a 2.1.1:

Paso 1 Sea $h_k = (w_k - w^*)$ una sucesión de Lyapunov

Paso 2 Análogo a 2.1.1 notemos que:

$$h_{k+1} - h_k = -2\alpha_k\,(w_k - w^*)\,\nabla F(w_k) + \alpha_k^2(\nabla F(w_k))^2$$

Notemos que a diferencia de antes la naturaleza discreta del algoritmo lleva a un término positivo de ruido en las variaciones. Notemos que si usamos 2.3 y 2.4 entonces:

$$h_{k+1} - \left(1 + \alpha_k^2 B\right) h_k \leq \underbrace{-2\alpha_k\,(w_k - w^*)\,\nabla F(w_k)}_{\leq 0 \text{ por } 2.0.1} + \alpha_k^2 A \leq \alpha_k^2 A$$

Definamos ahora las sucesiones auxiliares:

$$\mu_k = \prod_{j=1}^{k-1} \frac{1}{1 + \alpha_j^2 B} \tag{2.6a}$$

$$h'_k = \mu_k h_k \tag{2.6b}$$

Notemos que $\log(\mu_k) = -\sum_{j=1}^{k-1} \log\left(1 + \underbrace{\alpha_j^2 B}_{\geq 0}\right) \geq -B \sum_{j=1}^{k-1} \alpha_j^2 \geq -B \sum_{j=1}^{\infty} \alpha_j$, por lo que μ_k es una sucesión decreciente acotada inferiormente por $e^{-B \sum_{j=1}^{\infty} \alpha_j}$, luego $\mu_k \searrow \mu_\infty$ >0. Ahora si volvemos a 2.2 tenemos que:

$$h'_{k+1} - h'_k \leq \alpha_k^2 A \mu_k \leq \alpha_k^2 A$$

Como $\sum_{k=1}^{\infty} \alpha_k^2 A < \infty$ entonces $\sum_{k=1}^{\infty} h'_{k+1} - h'_k < \infty$ y por 2.2.1 concluímos que $\{h'_k\}$ converge; como $\underbrace{\mu_k}_{\geq 0} \to \mu_\infty > 0$ entonces $\{h_k\}$ converge.

Paso 3 De 2.2 como ya vimos que $h_{k+1} - \left(1 + \alpha_k^2 B\right) h_k$ es sumable concluímos que:

$$\sum_{k=1}^{\infty} \alpha_k \left(w_k - w^*\right) \nabla F(w_k) < \infty$$

Supongamos que $h_k \to h_{inf} \neq 0$, luego existiría $K \in \mathbb{N}$ y $\widetilde{\epsilon} > 0$ tal que $h_k = (w_k - w^*)^2 > \widetilde{\epsilon}$ para todo $k \geq K$; luego de 2.0.1 concluímos que existe $M > 0$ tal que $M \leq \left(w_k - w^*\right) \nabla F(w_k)$ para todo $k \geq K$. Por 2.3 eso implica que $\sum_{k=1}^{\infty} \alpha_k \left(w_k - w^*\right) \nabla F(w_k) = \infty$, concluímos que $w_k \xrightarrow{k \to \infty} w^*$. ∎

Corolario 2.2.3 *Sea $F \in C^2$, asumamos 2.0.1 y que $\left\|\nabla^2 F\right\|_2^2 \leq L$; si consideramos el algoritmo 2.1 tal que incrementos $\{\alpha_k\}$ cumplen 2.3 entonces:*

$$w_k \xrightarrow{k \to \infty} w^* \tag{2.7}$$

2.3 ACERCA DE CONVEXIDAD FUERTE Y FUNCIONES L-LIPSHITZ

Como ya notamos previamente, la condición de convexidad (en alguna medida) es central para analizar la convergencia de los algoritmos comunes en Machine Learning. Por lo tanto es una buena inversión dedicar esta sección a repasar las equivalencias de dos condiciones que van a aparecer repetidamente: **L-Lipshitz** y **Convexidad fuerte**.

Definición 2.3.1 *Sea $f \in C^1$, decimos que es fuertemente convexa o μ-convexa si existe $\mu > 0$ tal que para todos $x, y \in \mathbb{R}^d$ vale:*

$$f(y) \geq f(x) + \nabla F(x)^T (y - x) + \frac{\mu}{2} \|y - x\|_2^2 \tag{2.8}$$

Proposición 2.3.2 *Sea $f \in C^1$ una función $\mu-$convexa, entonces son equivalentes:*

1. $f(y) \geq f(x) + \nabla F(x)^T (y - x) + \frac{\mu}{2} \|y - x\|_2^2$ *para todos $x, y \in \mathbb{R}^d$*
2. $g(x) = f(x) - \frac{\mu}{2} \|x\|_2^2$ *es convexa para todo $x \in \mathbb{R}^d$*
3. $(\nabla f(y) - \nabla f(x))^T (y - x) \leq \mu \|y - x\|_2^2$ *para todos $x, y \in \mathbb{R}^d$*
4. $f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha) f(y) - \dfrac{\alpha (1 - \alpha) \mu}{2} \|y - x\|_2^2$ *para todos $x, y \in \mathbb{R}^d$, $\alpha \in [0, 1]$*

Demostración Ver A

Definición 2.3.3 *Decimos que una función $f \in C^1$ es PL$-$convexa, o cumple la condición de Polyak- Lojasiewicz (ver [**polyak:1963**], [**lojasiewicz:1963**]) si existe $\mu > 0$ tal que para todo $x \in \mathbb{R}^d$ vale:*

$$\frac{1}{2} \|\nabla f(x)\|_2^2 \geq \mu \left(f(x) - f_{inf} \right) \tag{2.9}$$

Proposición 2.3.4 *Sea $f \in C^1$ una función $\mu-$convexa, entonces valen:*

1. *f es PL-convexa*
2. $\|\nabla f(x) - \nabla f(y)\|_2 \geq \mu \|x - y\|_2$
3. $f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{1}{2\mu} \|\nabla f(x) - \nabla f(y)\|_2^2$
4. $(\nabla f(x) - \nabla f(y))^T (x - y) \leq \frac{1}{\mu} \|\nabla f(x) - \nabla f(y)\|_2^2$

Demostración Ver A

De esto podemos deducir un resultado, que aunque no lo usemos *per-se* en esta tesis, es de sumo interes:

Corolario 2.3.5 *Sea $h = f + g$ donde f es fuertemente convexa y g es convexa, entonces h es fuertemente convexa. En particular, si f es convexa entonces el problema regularizado en L2 de minimizar $h = f + \lambda \|x\|_2^2$ es fuertemente convexo.*

Demostración Sean $x, y \in \mathbb{R}^d$ y $\alpha \in [0, 1]$, luego:

$$\begin{aligned}
h(\alpha x + (1 - \alpha)y) &= f(\alpha x + (1 - \alpha)y) + g(\alpha x + (1 - \alpha)y) \\
&\underbrace{\leq}_{2.3.2} \alpha (f + g)(x) + (1 - \alpha)(f + g)(y) - \frac{\mu \alpha (1 - \alpha)}{2} \|x - y\|_2^2 \\
&= \alpha h(x) + (1 - \alpha) h(y) - \frac{\mu \alpha (1 - \alpha)}{2} \|x - y\|_2^2
\end{aligned}$$

∎

Una condición dual a la de convexidad fuerte es la de $L-$ Lipshitz.

Definición 2.3.6 *Sea $f \in C^1$, decimos que es L-Lipshitz si existe $L > 0$ tal que para todos $x, y \in \mathbb{R}^d$ vale:*

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L \|y - x\|_2 \tag{2.10}$$

Proposición 2.3.7 *Sea $f \in C^1$ una función $L-$Lipshitz, entonces para las siguientes propiedades:*

1. $\|\nabla f(x) - \nabla f(y)\|_2 \leq L \|y - x\|_2$
2. $g(x) = \frac{L}{2} x^T x - f(x)$ *es convexa*
3. $f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} \|y - x\|_2^2$
4. $(\nabla f(x) - \nabla f(y))^T (x - y) \leq L \|x - y\|_2^2$
5. $f(\alpha x + (1-\alpha) y) \geq \alpha f(x) + (1-\alpha) f(y) - \dfrac{\alpha(1-\alpha)}{2L} \|y - x\|_2^2$
6. $f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{1}{2L} \|y - x\|_2^2$
7. $(\nabla f(x) - \nabla f(y))^T (x - y) \geq \frac{1}{L} \|\nabla f(x) - \nabla f(y)\|_2^2$
8. $f(\alpha x + (1-\alpha) y) \leq \alpha f(x) + (1-\alpha) f(y) - \dfrac{\alpha(1-\alpha)}{2L} \|\nabla f(y) - \nabla f(x)\|_2^2$

Valen las siguientes cadenas de equivalencias:

$$6 \iff 8 \implies 7 \implies 1 \implies 2 \iff 3 \iff 4 \iff 5$$

Es más, si f además es $\mu-$convexa entonces las 8 propiedades son equivalentes

Demostración Ver A

Con todas estas propiedades, probemos el resultado histórico de **[polyak:1963]**

Teorema 2.3.8 (Convergencia lineal , Objetivos L-Lipshitz y PL-convexos) *Sea $F \in C^1$ tal que existe F_{inf} valor mínimo, F cumple 2.3.6 y 2.3.3; entonces el algortimo 2.1 con incremento fijo $\alpha_k = \frac{1}{L}$ cumple:*

$$f(w_k) - f_{inf} \leq \left(1 - \frac{\mu}{L}\right)^k (f(w_o) - f_{inf}) \tag{2.11}$$

Demostración Notemos que si usamos 2.3.7 entonces tenemos:

$$F(w_{k+1}) - F(w_k) \leq \nabla F(w_k)^T \left(-\frac{1}{L} \nabla F(w_k)\right) + \frac{L}{2} \left\| \frac{\nabla F(w_k)}{L} \right\|_2^2 \leq -\frac{1}{2L} \|\nabla F(w_k)\|_2^2$$

Luego por 2.3.3 tenemos:

$$F(w_{k+1}) - F(w_k) \leq -\frac{\mu}{L}\left(F(w_k) - F_{inf}\right)$$

Luego obtenemos:

$$F(w_{k+1}) - F_{inf} \leq \left(1 - \frac{\mu}{L}\right)\left(F(w_k) - F_{inf}\right) \leq \left(1 - \frac{\mu}{L}\right)^k \left(F(w_0) - F_{inf}\right)$$

$\blacksquare$

3 TEOREMA DE LA VARIEDAD ESTABLE Y LOS PUNTOS FIJOS INESTABLES

> "A Mathematician who is not also something of a poet will never be a complete mathematician"
>
> Karl Weierstrass

3.1 INTUICIÓN

Del capítulo anterior ya sabemos que bajo condiciones de convexidad estándar el algoritmo 2.1 converge puntualmente. Nos surge entonces la pregunta:

Bajo que casos el algoritmo 2.1 converge (en alguna forma) con objetivos no convexos?

En el caso no convexo, como analizamos previamente, existen puntos extremales no óptimos entre los cuales se encuentran los puntos silla, máximos y mínimos locales "grandes" (E. g.Puntos w^* tales que $F_{inf} \ll F(w^*)$) [En la bibliografía a estos puntos se los llama *shallow local minima*]. Los máximos en general no son preocupantes pues la naturaleza misma de los algoritmos de primer orden *escapa* de ellos.

Usemos un caso modelo para ejemplificar porque no es probable que los metodos de primer orden (entre ellos el algoritmo 2.1) convergan a puntos silla. Sea $f : \mathbb{R}^n \to \mathbb{R}^n$ dada por $f(x) = \dfrac{1}{2}x^T Hx$ con $H = \mathbf{diag}\,(\lambda_1, \ldots, \lambda_n)$; supongamos además que $\lambda_1, \ldots, \lambda_k > 0$ y $\lambda_{k+1}, \ldots, \lambda_n < 0$.

Ejemplo Si usamos en la base canónica de $\mathbb{R}^n$, $\mathcal{B} = \{e^1, \ldots, e^n\}$ 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 el algoritmo 2.1 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:

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

Por lo tanto por inducción es fácil probar que:

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

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:

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

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

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

Ejemplo 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:

$$\begin{aligned} \nabla f = & \quad (x, y^3 - y) \\ g = & \quad ((1-\alpha)x, (1+\alpha)y - \alpha y^3) \\ \nabla^2 f = & \quad \begin{pmatrix} 1 & 0 \\ 0 & 3y^2 - 1 \end{pmatrix} \end{aligned} \tag{3.1}$$

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 3.1.1 *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 3.1, 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:

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

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

Usamos 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 $\dfrac{dg_y}{dy}$ 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 $(x,y) \in E_s \iff (x,y) = (t,0)\ t \in \mathbb{R} \iff (x,y) \in W_s$. ∎

3.2 RESULTADOS PREVIOS

Esto quizas deberia ir en prerequisitos cuando lo tengamos

Ahora que vimos un par de ejemplos que nos dan una intuición acerca de la convergencia a puntos silla, usemos las herramientas de los sistemas dinámicos para analizar el caso general.

Por todo este capítulo, $g : \chi \to \chi$ y χ es una $d-$variedad sin borde.

Definición 3.2.1 *Dada una variedad de dimensión d χ 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 $\mu\left(\phi^i\left(E \cap U_i\right)\right) = 0$. En este caso usamos el abuso de notación $\mu(E) = 0$.*

Lema 3.2.2 *Sea $E \subset \chi$ tal que $\mu(E) = 0$; si $\det(Dg(x)) \neq 0$ para todo $x \in \chi$, luego $\mu\left(g^{-1}(E)\right) = 0$*

Demostración Sea $h = g^{-1}$ y (V_i, ψ^i) 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 $(V, \psi) \in \{(V_i, \phi^i)\}$ una carta determinada. Sea $\mathcal{A} := \{(U_i, \varphi^i)\}$ un atlas de χ y notemos $E_i = E \cap U_i$; luego $E = \bigcup_{i\in\mathbb{N}} E_i = \bigcup_{i\in\mathbb{N}} {\varphi^i}^{-1} \circ \varphi^i\left(E_i\right)$ por lo que:

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

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 $i \in \mathbb{N}$. ∎

3.3 PUNTOS FIJOS INESTABLES

Definición 3.3.1 *Sea:*

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

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 prerequisitos

Teorema 3.3.2 *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 para analizar la convergencia de los algoritmos de tipo batch en el caso no convexo:

Teorema 3.3.3 *Sea $g \in C^1(\chi)$ tal que $\det\left(Dg(x)\right) \neq 0$ 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 3.3.3 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 3.3.4 *$x_t \in \bigcap\limits_{k\in\mathbb{N}} g^{-k}(B_{x^*_i})$ para todo $t \geqslant T$*

Si notamos $S_i \triangleq \bigcap\limits_{k\in\mathbb{N}} g^{-k}(B_{x^*_i})$, entonces por 3.3.2 sabemos por un lado que es una subvariedad de W^{cs}_{loc} y por el otro que $\dim(S_i) \leqslant \dim(W^{cs}_{loc}) = \dim(E_s) < d - 1$; por lo que $\mu(S_i) = 0$.

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

Usamos que una subvariedad de dimension menor tiene medida 0

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

$$\begin{aligned} \mu(W) \leqslant \quad & \mu\left(\bigcup_{i \in \mathbb{N}} \bigcup_{k \in \mathbb{N}} g^{-k}(S_i) \right) \\ \leqslant \quad & \sum_{i \in \mathbb{N}} \sum_{k \in \mathbb{N}} \mu\left(g^{-k}(S_i) \right) \\ \overset{3.2.2}{=} \quad & 0 \end{aligned}$$

∎

Para finalizar veamos un caso simple que nos encontraremos seguido:

Corolario 3.3.5 *Bajo las mismas hipótesis que en 3.3.3 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 $\mu(W_g) \leqslant \mu(W) \overset{3.3.3}{=} 0$. ∎

4

CONVERGENCIA CTP A MÍNIMOS : CASO GENERAL

4.1 DESCENSO DE GRADIENTE EN BATCH

> "The difference between mathematicians and physicists is that after physicists prove a big result they think it is fantastic but after mathematicians prove a big result they think it is trivial."
>
> Richard Feynman

Como una aplicación del teorema en 3.3.3 demostremos que el *descenso de gradiente en batch* tiene probabilidad cero de converger a puntos silla. Consideremos el algoritmo 2.1 con incrementos constantes $\alpha_k = \alpha$:

$$x_{k+1} = g(x_k) \stackrel{\triangle}{=} x_k - \alpha \nabla f(x_k) \tag{4.1}$$

Hipótesis 4.1.1 *Asumamos que $f \in \mathcal{C}^2$ y $\left\|\nabla^2 f(x)\right\|_2 \leqslant L$*

Proposición 4.1.2 *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 $\{1 - \alpha\lambda_i \ : \ \lambda_i \in \{\mu \ : \ \nabla^2 f(x^*)v = \mu v \quad \text{para algún } v \neq 0\}\}$. 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 4.1.3 *Bajo 4.1.1 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:

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

Luego por 4.1.1 tenemos que $\alpha < \frac{1}{|\lambda_i|}$ y entonces $1 - \alpha\lambda_i > 0$ para todo $i \in \{1, \ldots, d\}$; concluímos que $\det(Dg(x)) > 0$. ∎

Corolario 4.1.4 *Sea g dada por el algoritmo 2.1, bajo 4.1.1 y $\alpha < \frac{1}{L}$ se tiene que $\mu\left(W_g\right) = 0$.*

Demostración Por 4.1.2 y 4.1.3 tenemos que vale 3.3.5 y concluímos que $\mu\left(W_g\right) = 0$. ∎

4.2 PUNTO PRÓXIMO

El algoritmo de punto próximo esta dado por la iteración:

$$x_{k+1} = g(x_k) \triangleq \arg\min_{z\in\chi} f(z) + \frac{1}{2\alpha}\|x_k - z\|_2^2 \tag{4.2}$$

Proposición 4.2.1 *Bajo 4.1.1 y $\alpha < \frac{1}{L}$ entonces vale:*

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

Probamos esto? Me parece un poco claro

Demostración Veamos primero el siguiente lema:

Lema 4.2.2 *Bajo 4.1.1, $\alpha < \frac{1}{L}$ 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.

$$\begin{array}{ccccc} x_{k+1} = & g(x_k) = & \arg\min_{z\in\chi} f(z) + \frac{1}{2\alpha}\|x_k - z\|_2^2 \\ \downarrow & \downarrow & \downarrow \\ x = & g(x) = & \arg\min_{z\in\chi} f(z) + \frac{1}{2\alpha}\|x - z\|_2^2 \end{array}$$

$$\Longleftrightarrow \nabla_z\left(f(z) + \frac{1}{2\alpha}\|x-z\|^2\right)(g(x)) = 0$$

$$\Longleftrightarrow \nabla f(g(x)) - \frac{1}{\alpha}(x - g(x)) = 0$$

$$\Longleftrightarrow g(x) + \alpha\nabla f(g(x)) = x$$

Finalmente por diferenciación implicita obtenemos:

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

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 $\det\left(Dg(x)\right) \neq 0$. ∎

Corolario 4.2.3 *Sea g dado por el algoritmo de punto próximo con ecuación 4.2, bajo 4.1.1 y $\alpha < \frac{1}{L}$ se tiene que $\mu\left(W_g\right) = 0$.*

Demostración Por 4.2.1 tenemos que vale 3.3.5 y concluímos que $\mu\left(W_g\right) = 0$. ∎

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 algoritmo 4.1:

Algorithmus 4.1 : Descenso por coordenadas

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

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

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

$$x_{k+1} = g(x_k) \triangleq g_d \circ g_{d-1} \circ \cdots \circ g_1(x) \tag{4.3}$$

Lema 4.3.2 *Si g está dada por 4.3 entonces si notamos $P_S = \sum\limits_{i \in S} e_i e_i^T$ entonces:*

$$Dg(x_k) = \prod_{i \in \{1, \ldots, b\}} \left(Id - \alpha P_{b-i+1} \nabla^2 f(y_k^{S_{b-i}})\right) \tag{4.4}$$

Demostración Notemos primero que:

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

Por lo tanto:

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

∎

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 $\max\limits_{i \in \{1, \ldots, b\}} \left\|\nabla^2 f(x)|_{S_i}\right\| = L_b$

Proposición 4.3.3 *Bajo 6 y $\alpha < \frac{1}{L_b}$ se tiene que* $\det(Dg(x)) \neq 0$

Demostración Basta probar que cada término de 4.4 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 $j \in S_i$, $i \in \{1, \ldots, b\}$ por lo que todos los autovalores son positivos y $Dg_i(x)$ es invertible para todo i. ∎

Proposición 4.3.4 *Bajo 6 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 4.3.5 *Sea $y_t \in Ran(H)$, luego existe $i \in \{1, \ldots, b\}$ 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 4.3.6 *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:

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

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

$$
z_{i+1}^T H z_{i+1} < z_i^T H z_i - \alpha \left\| \sum_{j \in S_i} \left(e_j^T H z_i \right) e_j \right\|_2^2 \tag{4.5}
$$

Por otro lado sabemos que para todo w vale:

$$
w^T H w \geq \lambda_{min}(H) \|w\|_2^2 \geq -L_b \|w\|_2^2 \tag{4.6}
$$

Usamos Cauchy Schwartz

Luego si usamos 4.3.5, 4.6 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 4.3.7 *Sea g dado por el algoritmo de descenso por coordenadas con ecuación 4.3, bajo 6 y $\alpha < \frac{1}{L_b}$ se tiene que $\mu\left(W_g\right) = 0$.*

Demostración Por 4.3.3 y 4.3.4 tenemos que vale 3.3.5 y concluímos que $\mu\left(W_g\right) = 0$. ∎

5 RESULTADOS NEGATIVOS

Ya vimos de 4.1.4, que el descenso de gradiente, con cualquier inicialización aleatoria razonable, siempre escapará de los puntos de silla estrictos *eventualmente*, pero sin ninguna garantía sobre el número de pasos requeridos. Esto motiva a la siguiente pregunta:

¿El descenso de gradiente inicializado aleatoriamente generalmente escapa de los puntos de silla en tiempo polinomial?

5.1 EJEMPLOS *patológicos*

Inicialización uniforme en una banda exponencialmente chica Consideremos $f \in C^2(\mathbb{R}^2)$ con un punto silla estricto en $(0,0)$. Supongamos que a orden chico en $U = [-1,1]^2$ un entorno del punto silla f es localmente de la forma $f(x_1, x_2) = x_1^2 - x_2^2$, luego si utilizamos el algoritmo 2.1 con $\alpha_k = \alpha = \frac{1}{4}$ nos queda:

$$\left(x_1^{k+1}, x_2^{k+1}\right) = \left(x_1^k, x_2^k\right) - \frac{1}{4}\left(2x_1^k, -2x_2^k\right) = \left(\frac{x_1^k}{2}, \frac{3x_2^k}{2}\right)$$

Luego si tomamos $\epsilon > 0$ y $w_0 = (x_1^0, x_2^0)$ uniformemente en $w_0 \in \widetilde{U} = [-1,1] \times \left[-\frac{3}{2}^{-e^{\frac{1}{\epsilon}}}, \frac{3}{2}^{-e^{\frac{1}{\epsilon}}}\right]$ entonces el algoritmo 2.1 necesita $k \geq e^{\frac{1}{\epsilon}}$ pasos para que $w_k \notin U$. Concluímos que el algoritmo es exponencial en converger a cualquier mínimo si $w_0 \in \widetilde{U}$. ∎

Inicialización exponencialmente lejana Consideremos nuevamente $f \in C^2(\mathbb{R}^2)$ dada por:

$$f(x_1, x_2) = \begin{cases} x_1^2 - x_2^2 & \text{si } x_1 \in (-1,1) \\ -4x_1 + x_2^2 & \text{si } x_1 < -2 \\ h(x_1, x_2) & \text{sino} \end{cases}$$

Con h una función suave tal que $f \in C^2$ y x_2 no crezca demasiado en el intervalo donde es h (Una forma de definir esto es con splines cúbicos).

Luego si para el algoritmo 2.1 tomamos $\alpha_k = \alpha = \frac{1}{4}$ tendríamos la siguiente dinámica:

$$\begin{aligned} \left(x_1^{k+1}, x_2^{k+1}\right) &= \begin{cases} \left(x_1^k, x_2^k\right) - \frac{1}{4}\left(2x_1^k, -2x_2^k\right) & \text{si } x_1 \in (-1,1) \\ \left(x_1^k, x_2^k\right) - \frac{1}{4}\left(-4, 2x_2^k\right) & \text{si } x_1 < -2 \end{cases} \\ &= \begin{cases} \left(\frac{x_1^k}{2}, \frac{3x_2^k}{2}\right) & \text{si } x_1 \in (-1,1) \\ \left(x_1^k + 1, \frac{x_2^k}{2}\right) & \text{si } x_1 < -2 \end{cases} \end{aligned}$$

Luego si tomamos $R > 0$ grande y $w_0 = (x_1^0, x_2^0)$ uniformemente en $w_0 \in \widetilde{U} = [-R-1, -R+1] \times [-1, 1]$ entonces notando t como la primera vez que $x_1 > -1$ tenemos que $t \approx R$, con lo que $x_2^t = x_2^0 \left(\frac{1}{2}\right)^R$. Por ende, el algoritmo nuevamente necesita $R \approx e^{\frac{1}{\epsilon}}$ iteraciones para poder salir de $U = [-1, 1]^2$; concluímos que el algoritmo es exponencial en converger a cualquier mínimo si $w_0 \in \widetilde{U}$. ∎

Notacion Ω

Teorema 5.1.1 (Convergencia exponencial, Inicialización uniforme en el cubo) *Consideremos el algoritmo 2.1 con w_0 elegido uniformemente en $[-1, 1]^d$; luego existe $f : \mathbb{R}^d \mapsto \mathbb{R}$ $B-$acotada, $l-$Lipshitz, $\mu-$Lipshitz en el Hessiano con $B, l, \mu \in poly(d)$ tal que si $\alpha_k = \alpha \leq \frac{1}{l}$ entonces w_k va a estar a $\Omega(1)$ de cualquier mínimo para todo $k \leq e^{\Omega(d)}$*

Antes de pasa a la prueba veamos un ejemplo modelo para generar intuición de la demostración:

Escapar de dos puntos silla consecutivos Sean $L > \gamma > 0$ y $f \in [0, 3] \times [0, 3]$ dada por:

$$f(x_1, x_2) = \begin{cases} -\gamma x_1^2 + L x_2^2 & \text{si } (x_1, x_2) \in [0,1] \times [0,1] \\ L\left(x_1 - 2\right)^2 - \gamma x_2^2 & \text{si } (x_1, x_2) \in [1,3] \times [0,1] \\ L\left(x_1 - 2\right)^2 + L\left(x_2 - 2\right)^2 & \text{si } (x_1, x_2) \in [1,3] \times [1,3] \end{cases} \tag{5.1}$$

Notemos que f tiene dos puntos silla estrictos en $(0, 0)$ y $(2, 0)$, mientras que tiene un óptimo en $(2, 2)$. Sean $U = [0, 1]^2$, $V = [1, 3] \times [0, 1]$ y $W = [1, 3]^2$ entornos respectivos de los tres puntos críticos, supongamos que $w_0 = \left(x_1^0, x_2^0\right) \in U$ y definamos:

$$k_1 = \inf_{x_1^k \geq 1} k = \min_{x_1^k \geq 1} k$$
$$k_2 = \inf_{x_2^k \geq 1} k = \min_{x_2^k \geq 1} k$$

Notemos que como la dirección de escape en $(0, 0)$ es por x_1 y *luego* por x_2 (por el cambio de comportamiento de f) podemos concluir que k_1, k_2 estan bien definidos y que $k_2 \geq k_1 \geq 0$; la observación clave va a ser que $k_2 = C k_1$ con $C > 1$, es decir que el tiempo en pasar el siguiente punto silla es exponencialmente mayor que los anteriores. En pos de esto, veamos como va a ser la iteración del algoritmo 2.1:

$$\left(x_1^{k+1}, x_2^{k+1}\right) = \begin{cases} \left(x_1^k, x_2^k\right) - \alpha\left(-2\gamma x_1^k, 2L x_2^k\right) & \text{si } x_1 \leq 1 \\ \left(x_1^k, x_2^k\right) - \alpha\left(2L\left(x_1^k - 2\right), -2\gamma x_2^k\right) & \text{si } x_1 \geq 1\,,\ x_2 \leq 1 \\ \left(x_1^k, x_2^k\right) - \alpha\left(2L\left(x_1^k - 2\right), 2L\left(x_2^k - 2\right)\right) & \text{si } x_1 \geq 1\,,\ x_2 \geq 1 \end{cases}$$

$$= \begin{cases} \left(\left(1 + 2\alpha\gamma\right) x_1^k, \left(1 - \alpha 2L\right) x_2^k\right) & \text{si } x_1 \leq 1 \\ \left(\left(1 - 2L\alpha\right) x_1^k + 4L\alpha, \left(1 + 2\alpha\gamma\right) x_2^k\right) & \text{si } x_1 \geq 1\,,\ x_2 \leq 1 \\ \left(\left(1 - 2L\alpha\right) x_1^k + 4L\alpha, \left(1 - 2L\alpha\right) x_2^k + 4L\alpha\right) & \text{si } x_1 \geq 1\,,\ x_2 \geq 1 \end{cases}$$

Luego evaluando en k_1 y k_2:

$$
\begin{aligned}
x_1^{k_1} &= (1+2\alpha\gamma)^{k_1} x_1^0 \\
x_2^{k_1} &= (1-2\alpha L)^{k_1} x_1^0
\end{aligned}
$$

$$
\begin{aligned}
x_1^{k_2} &= (1-2L\alpha)^{k_2-k_1}(1+2\alpha\gamma)^{k_1} x_1^0 + K \geq 1 \quad K \text{ constante} \\
x_2^{k_2} &= (1+2\alpha\gamma)^{k_2-k_1}(1-2\alpha L)^{k_1} x_2^0 \geq 1
\end{aligned}
$$

Concluímos que:

$$k_2 \geq \frac{2\alpha(L+\gamma)k_1 - \log\left(x_2^0\right)}{2\alpha\gamma} \geq \frac{L+\gamma}{\gamma}k_1 \tag{5.3}$$

Esta f que presentamos tiene varios problemas:

1. f no es continua, y mucho menos C^2
2. f no podemos asegurar que sea $l-$Lipshitz o $\mu-$Lipshitz en el hessiano
3. Los puntos críticos estan en el borde del dominio, lo que no es ideal
4. f no está definida en todo $\mathbb{R}^d$
5. Estrictamente f es aún resuelto en tiempo polinomial

La clave va a ser usar splines para resolver los primeros puntos, espejar f para hacer los puntos extremales interiores, asignar d puntos críticos similares para generar el tiempo exponencial en d y extender esa función $\tilde{f}$ a $\mathbb{R}^d$ con el Teorema de extensión de Whitney. Aunque la demostración es larga y tediosa, la idea clave es la vista aquí.

Demostración Vayamos de a pasos

Paso 1 - Definiciones

Fijemos 4 constantes: $L = e, \gamma = 1, \tau = e, \eta$ a definir proximamente; inspirados en el ejemplo anterior vamos a construir una f definida en un cerrado D_0 tal que tenga $d-1$ puntos silla estrictos y la complejidas del algoritmo 2.1 sea exponencial. Sea D_0 dado por:

$$
\begin{aligned}
D_0 = & \bigcup_{i=1}^{d+1} \left\{x \in \mathbb{R}^d \ : \ 6\tau \geq x_1, \ldots, x_{i-1} \geq 2\tau; 2\tau \geq x_i \geq 0; \tau \geq x_{i+1}, \ldots, x_d \geq 0\right\} \\
:= & \bigcup_{i=1}^{d+1} D_i
\end{aligned} \tag{5.4}
$$

Y partamos $D_i = D_{i,1} \cup D_{i,2}$ donde $D_{i,1} = \{x \in D_i \ : \ 0 \leq x_i \leq \tau\}$ y $D_{i,2} = \{x \in D_i, \ : \ \tau \leq x_i \leq 2\tau\}$.

Para un dado $1 \leq i \leq d-1$ definamos:

$$f|_{D_i}(x) = \begin{cases} \sum_{j=1}^{i-1} L\left(x_j - 4\tau\right)^2 - \gamma x_i^2 + \sum_{j=i+1}^{d} L x_j^2 - (i-1)\,\eta \\ \triangleq f_{i,1}(x) \text{ si } x \in D_{i,1} \\ \\ \sum_{j=1}^{i-1} L\left(x_j - 4\tau\right)^2 + g(x_i, x_{i+1}) + \sum_{j=i+2}^{d} L x_j^2 - (i-1)\,\eta \\ \triangleq f_{i,2}(x) \text{ si } x \in D_{i,2} \end{cases} \tag{5.5}$$

Donde nuevamente η esta pendiente de definición y $g : \mathbb{R}^2 \to \mathbb{R}$ también la definiremos proximamente para que f resulte C^2, $B-$acotada, $l-$Lipshitz, $\mu-$Lipshitz en el Hessiano con $B, l, \mu \in \text{poly}(d)$.

Para $i = d$ definamos:

$$f|_{D_d}(x) = \begin{cases} \sum_{j=1}^{d-1} L\left(x_j - 4\tau\right)^2 - \gamma x_d^2 - (d-1)\,\eta \\ \triangleq f_{d,1}(x) \text{ si } x \in D_{d,1} \\ \\ \sum_{j=1}^{d-1} L\left(x_j - 4\tau\right)^2 + g_1(x_d) - (d-1)\,\eta \\ \triangleq f_{d,2}(x) \text{ si } x \in D_{d,2} \end{cases} \tag{5.6}$$

Donde como antes, g_1 lo definiremos proximamente. Finalmente si $i = d+1$ entonces $6\tau \geq x_i \geq 2_tau$ para todo $1 \leq i \leq d$ y definimos:

$$f|_{D_{d+1}} = \sum_{j=1}^{d} L\left(x_j - 4\tau\right)^2 - d\eta \triangleq f_{d+1,1} \tag{5.7}$$

Lema 5.1.2 *Sea $g(x_i, x_{i+1}) = g_1(x_i) + g_2(x_i)x_{i+1}^2$, existen g_1, g_2 polinomios y $\eta = -g_1(2\tau) + 4L\tau^2$ tal que para todo $1 \leq i \leq d$ si $x_i = \tau$ vale:*

$$\begin{aligned} f_{i,2}(x) &= f_{i,1}(x) \\ \nabla f_{i,2}(x) &= \nabla f_{i,1}(x) \\ \nabla^2 f_{i,2}(x) &= \nabla^2 f_{i,1}(x) \end{aligned}$$

Y si $x_i = 2\tau$ entonces:

$$\begin{aligned} f_{i,2}(x) &= f_{i+1,1}(x) \\ \nabla f_{i,2}(x) &= \nabla f_{i+1,1}(x) \\ \nabla^2 f_{i,2}(x) &= \nabla^2 f_{i+1,1}(x) \end{aligned}$$

Es más, si $x \in D_{i,2} \cap D_{i+1,1}$ entonces:

$$
\begin{aligned}
-4L\tau \leq & \quad \frac{\partial g}{\partial x_i}(x_i, x_{i+1}) \leq \quad -2\gamma\tau \\
-2\gamma x_{i+1} \leq & \quad \frac{\partial g}{\partial x_{i+1}}(x_i, x_{i+1})
\end{aligned}
$$

Y finalmente si $x \in D_{i,2}$ entonces:

$$-4L\tau \leq \frac{\partial g_1}{\partial x_i}(x_i) \leq -2\gamma\tau$$

Demostración Ver A

Observación Del lema anterior podemos ver que $\deg(g_1), \deg(g) \leq 5$ por lo que estan acotados. Concluímos que ambas son $B-$acotadas y $\mu-$Lipshitz con $B, \mu \in \text{poly}(L)$

Observación Notemos que $\|g_1\|, \|g\| > \gamma\tau > 0$ por lo que ninguna de las dos aporta puntos críticos en D_0.

Observación Notemos que f queda C^2 y que sus $d+1$ puntos críticos son $z_i = \left(4\tau, \dots, \underbrace{4\tau}_{i}, 0, \dots, 0\right)$ donde todos son puntos silla estrictos menos $z_d = (4\tau, \dots, 4\tau)$ que es mínimo.

Paso 2- Cota superior a T_k^τ

Supongamos ahora que $\tau > e$, $\alpha \leq \frac{1}{2L}$ y tomemos $w_0 \in [-1,1]^d \cap D_0$, veamos que para todo $T \leq \left(\frac{L+\gamma}{\gamma}\right)^{d-1}$ vale que $x_d^T \leq 2\tau \notin D_{d+1}$.

Sea $T_0 = 0$ y definamos $T_k = \min\limits_{x_k^t \geq 2\tau}$ el tiempo de escape de $D_{k,2}$; notemos que como $x^0 \in D_{1,1}$ vale que $T_k \geq 0$ para todo k y esta bien definido. Definamos además T_k^τ como la cantidad de iteraciones que x^k esta en $D_{k,2}$ antes de escapar; como del lema $\frac{\partial g}{\partial x_k}(x_k, x_{k+1}) \leq -2\gamma\tau$ tenemos que $\left|x^k - x^{k+1}\right| \geq 2\alpha\gamma\tau$ por lo que:

$$T_k^\tau \leq \frac{\tau}{2\alpha\gamma\tau} = \frac{1}{2\gamma\alpha} \qquad \forall k \in \{1, \dots, d+1\}$$

Paso 3 - Cota inferior para T_1:

Notemos que T_1 es el mínimo valor tal que $x_1^{T_1} \geq 2\tau$ y entonces vale que $x_1^{T_1 - T_1^\tau} \geq \tau$, como del algoritmo 2.1 sabemos que en $D_{1,2}$ vale la relación:

$$x_1^t = (1 + 2\alpha\gamma)^t \, x_1^0$$

Tenemos que:

$$\begin{aligned} x_1^0 (1+2\alpha\gamma)^{T_1 - T_1^\tau} &\geq \tau \\ \Longrightarrow \qquad T_1 - T_1^\tau &\geq \frac{1}{2\alpha\gamma} \underbrace{\log\left(\frac{\tau}{x_1^0}\right)}_{\geq 1} \geq T_1^\tau \end{aligned}$$

Paso 4 - El algoritmo 2.1 se queda confinado a D_0:

Si $x^t \in D_{k,1}$ luego las iteraciones del algortimo son:

$$x_j^{t+1} = \begin{cases} (1-\alpha L)\, x_j^t - 4\alpha L\tau \in [2\tau, 6\tau] & 1 \leq j \leq k-1 \\ (1+2\alpha\gamma)\, x_j^t \tau \in [0, 2\tau] & j = k \\ (1-2\alpha L)\, x_j^t \in [0,\tau] & j \geq k+1 \end{cases}$$

Mientras que si $x^t \in D_{k,2}$ entonces:

$$x_j^{t+1} = \begin{cases} (1-\alpha L)\, x_j^t - 4\alpha L\tau \in [2\tau, 6\tau] & 1 \leq j \leq k-1 \\ x_j^t - \alpha \dfrac{\partial g}{\partial x_k}(x_k, x_{k+1}) \leq x_j^t + 2\alpha\gamma\tau \in [0, 6\tau] & j = k \\ (1-2\alpha L)\, x_j^t \in [0,\tau] & j \geq k+2 \end{cases}$$

Separemos el caso $j = k+1$, donde el lema 5.1.2 nos dice que:

$$\frac{\partial f}{\partial x_{k+1}}(x) \geq -2\gamma x_{k+1}$$

Luego para $t = T_k - T_k^\tau + 1, \ldots, T_k$ vale:

$$x_{k+1}^t \leq x_{k+1}^0 (1-2\alpha L)^{T_k - T_k^\tau} (1+2\alpha\gamma)^{t - \left(T_k - T_k^\tau\right)} \leq \tau$$

Y concluímos que $x^t \in D_0$

Paso 5 - Relación entre T_{k+1} y T_k:

Por un lado, por la definición de T_k y T_k^τ:

$$x_{k+1}^{T_k} \leq x_{k+1}^0 (1-2\alpha L)^{T_k - T_k^\tau} (1+2\gamma\alpha)^{T_k^\tau}$$

Por el otro, usando el mismo arguemento que cuando acotamos por debajo a T_1:

$$\begin{aligned} & x_{k+1}^{T_{k+1} - T_{k+1}^\tau} && \geq \tau \\ \Rightarrow\, & x_{k+1}^{T_k} (1+2\alpha\gamma)^{T_{k+1} - T_{k+1}^\tau - T_k} && \geq \tau \\ \Rightarrow\, & x_{k+1}^0 (1-2\alpha L)^{T_k - T_k^\tau} (1+2\gamma\alpha)^{T_k^\tau} (1+2\alpha\gamma)^{T_{k+1} - T_{k+1}^\tau - T_k} && \geq \tau \end{aligned}$$

Luego como $\alpha < \frac{1}{2L}$:

$$2\alpha\gamma\left(T_{k+1} - T_{k+1}^{\tau} - (T_k - T_k^{\tau})\right) \geq \underbrace{\log\left(\frac{\tau}{x_{k+1}^{0}}\right)}_{\geq 1} + 2\alpha L\left(T_k - T_k^{\tau}\right)$$

$$\Rightarrow \quad T_{k+1} - T_{k+1}^{\tau} \geq \frac{L+\gamma}{\gamma}\left(T_k - T_k^{\tau}\right)$$

Inductivamente:

$$T_d \geq T_d - T_d^{\tau} \geq \left(\frac{L+\gamma}{\gamma}\right)^{d-1}\left(T_1 - T_1^{\tau}\right) \geq \frac{1}{2\alpha\gamma}\left(\frac{L+\gamma}{\gamma}\right)^{d-1} \geq \left(\frac{L+\gamma}{\gamma}\right)^{d-1} \tag{5.8}$$

Paso 6 - Extender D_0 para que los puntos extremales sean interiores

Ya probamos que si $x^0 \in [-1,1]^d \cap D_0$ entonces el algoritmo 2.1 necesita tiempo exponencial para converger al mínimo, ataquemos el caso $x^0 \in [-1,1]^d \cap D_0^c$

Para $a = 0, \ldots, 2^d - 1$ sea a_2 la representación binaria de a y notemos $a_2(0)$ los índices donde a_2 tiene 0 y análogo con $a_2(1)$. Definamos:

$$D_a = \bigcup_{i=1}^{d} \Big\{x \in \mathbb{R}^d \; : \; x_i \geq 0 \text{ si } i \in a_2(0), \; x_i \leq 0 \text{ sino}, \\ 6\tau \geq |x_1|, \ldots, |x_{i-1}| \geq 2\tau, \; |x_i| \leq 2\tau, \; |x_{i+1}|, \ldots, |x_d| \leq \tau\Big\}$$

$$D = \bigcup_{a=0}^{2^d-1} D_a$$

Notemos que D es cerrado y que $[-1,1]^d \subset D$. Ahora definamos la función f; sea $i = 0, \ldots, d$ y definamos los subdominios:

$$\widetilde{D}_{i,1} = \left\{x \in \mathbb{R}^d \; : \; 6\tau \geq |x_1|, \ldots, |x_{i-1}| \geq 2\tau, \; |x_i| \leq \tau, \; |x_{i+1}|, \ldots, |x_d| \leq \tau\right\}$$
$$\widetilde{D}_{i,2} = \left\{x \in \mathbb{R}^d \; : \; 6\tau \geq |x_1|, \ldots, |x_{i-1}| \geq 2\tau, \; \tau \leq |x_i| \leq 2\tau, \; |x_{i+1}|, \ldots, |x_d| \leq \tau\right\}$$
$$\widetilde{D}_{d+1} = \left\{x \in \mathbb{R}^d \; : \; 6\tau \geq |x_1|, \ldots, |x_d| \geq 2\tau\right\}$$

Luego definimos:

$$f(x)=\begin{cases}
\sum\limits_{j\leq i-1, j\in a_2(0)} L\left(x_j-4\tau\right)^2+\sum\limits_{j\leq i-1, j\in a_2(1)} L\left(x_j+4\tau\right)^2 & \\
\qquad -\gamma x_i^2+\sum\limits_{j\geq i+1} Lx_j^2-(i-1)\,\eta & \text{si } x\in D_{i,1},\ i<d \\
\sum\limits_{j\leq i-1, j\in a_2(0)} L\left(x_j-4\tau\right)^2+\sum\limits_{j\leq i-1, j\in a_2(1)} L\left(x_j+4\tau\right)^2 & \\
\qquad +G(x_i,x_{i+1}|)+\sum\limits_{j\geq i+2} Lx_j^2-(i-1)\,\eta & \text{si } x\in D_{i,2},\ i<d \\
\sum\limits_{j\leq d-1, j\in a_2(0)} L\left(x_j-4\tau\right)^2+\sum\limits_{j\leq d-1, j\in a_2(1)} L\left(x_j+4\tau\right)^2 & \\
\qquad -\gamma x_d^2-(d-1)\,\eta & \text{si } x\in D_{d,1} \\
\sum\limits_{j\leq d-1, j\in a_2(0)} L\left(x_j-4\tau\right)^2+\sum\limits_{j\leq d-1, j\in a_2(1)} L\left(x_j+4\tau\right)^2 & \\
\qquad +G_1(x_d)-(d-1)\,\eta & \text{si } x\in D_{d,2} \\
\sum\limits_{j\leq d, j\in a_2(0)} L\left(x_j-4\tau\right)^2+\sum\limits_{j\leq d, j\in a_2(1)} L\left(x_j+4\tau\right)^2 & \\
\qquad -d\eta & \text{si } x\in D_{d+1}
\end{cases}$$

Donde:

$$G(x_i,x_{i+1})=\begin{cases} g(x_i,x_{i+1}) & \text{si } i\in a_2(0) \\ g(-x_i,x_{i+1}) & \text{si } i\in a_2(1)\end{cases}$$

$$G_1(x_i)=\begin{cases} g_1(x_i) & \text{si } i\in a_2(0) \\ g_1(-x_i) & \text{si } i\in a_2(1)\end{cases}$$

Notemos que por simetría de la definición, si espejamos la demostración del punto anterior es claro que si $\tau\geq e$ y $x^0\in[-1,1]^d$ entonces el algoritmo 2.1 con $\alpha<\frac{1}{2L}$ cumple $x_d^T\leq 2\tau$ para todo $T\leq\left(\frac{L+\gamma}{\gamma}\right)$ y por lo tanto necesita $e^{\Omega(d)}$ operaciones para llegar al único mínimo $(4\tau,\ldots,4\tau)$.

Paso 7 - Extender de D a $\mathbb{R}^d$

Por A.1.2 sabemos que existe $F:\mathbb{R}^d\to\mathbb{R}$ que extiende a f y que $\|F\|_\infty,\|F\|_{C^m}\leq\mathcal{O}(\text{poly}(d))$; y aunque F puede admitir nuevos puntos críticos, del paso 4 y 6 sabemos que si $x^0\in[-1,1]^d$ entonces $\{x^k\}\subset D$. ∎

Part III

Algoritmos Estocásticos

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

6 CONVERGENCIA EN L1

> "The Axiom of Choice is obviously true, the well-ordering principle obviously false, and who can tell about Zorn's lemma?"
>
> Jerry Bona

6.1 CONTEXTO

En esta parte vamos a analizar la convergencia en L1 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:

$$F(w) = \begin{cases} R(w) = & \mathbb{E}\left[f(w,\xi)\right] \\ \quad\text{o} & \\ R_n(w) = & \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} f_i(w) \end{cases} \tag{6.1}$$

Sea el algoritmo estocástico 6.1

Algorithmus 6.1 : Descenso Estocastico (DE)

```
Input: w_1 ∈ R^d el inicio de la iteración, {ξ_k} iid
for k ∈ N do
    Generar una muestra de la variable aleatoria ξ_k
    Calcular el vector estocástico g(w_k, ξ_k)
    Elegir α_k > 0
    w_{k+1} ← w_k − α_k g(w_k, ξ_k)
```

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:

$$g(w_k, \xi_k) = \begin{cases} \nabla f(w_k, \xi_k) \\ \dfrac{1}{n_k}\sum_{i=1}^{n_k} \nabla f(w_k, \xi_{k,i}) \\ H_k \dfrac{1}{n_k}\sum_{i=1}^{n_k} \nabla f(w_k, \xi_{k,i}) \end{cases} \tag{6.2}$$

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 recordemos la condición 2.3.6:

Hipótesis 6.1.1 (*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 6.1.1, 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 Ver 2.3.7

6.2 ALGUNOS LEMAS FUNDAMENTALES

Con el contexto claro, veamos algunos lemas que van a ser clave en la demostración de la convergencia L1 del algoritmo 6.1.

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 6.2.1 *Bajo 6.1.1 las iteraciones de 6.1 satisfacen que para todo* $k \in N$*:*

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leq -\alpha_k \nabla F(w_k)^T \mathbb{E}_{\xi_k}[g(w_k, \xi_k)] + \frac{1}{2}\alpha_k^2 \mathbb{E}_{\xi_k}\left[\|g(w_k, \xi_k)\|_2^2\right] \tag{6.3}$$

Demostración Notemos que por 6.1.1 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 $\mathbb{E}_{X,Y}[Y|X] = \mathbb{E}[Y]$:

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

∎

Observación Notemos que si $g(w_k, \xi_k)$ es un estimador insesgado de $\nabla F(w_k)$ entonces de 6.2.1:

$$
\mathbb{E}_{\xi_k}\left[F(w_{k+1})\right] - F(w_k) \leq -\alpha_k \left\|\nabla F(w_k)\right\|^2 + \frac{1}{2}\alpha_k^2 \mathbb{E}_{\xi_k}\left[\|g(w_k, \xi_k)\|_2^2\right] \tag{6.4}
$$

Luego entonces para controlar la convergencia de 6.1 también hay que poner suposiciones sobre el segundo momento de g, luego si definimos:

$$
\mathbb{V}_{\xi_k}\left[g(w_k, \xi_k)\right] := \mathbb{E}_{\xi_k}\left[\|g(w_k, \xi_k)\|_2^2\right] - \left\|\mathbb{E}_{\xi_k}\left[g(w_k, \xi_k)\right]\right\|_2^2 \tag{6.5}
$$

Asumamos:

Hipótesis 6.2.2 (Acotaciones al primer y segundo momento de g) *Supongamos que dada F función objetivo y g la estimación del gradiente en 6.1 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:*

$$
\nabla F(w_k)^T \mathbb{E}_{\xi_k}\left[g(w_k, \xi_k)\right] \geq \mu \left\|\nabla F(w_k)\right\|_2^2 \tag{6.6a}
$$

Y

$$
\left\|\mathbb{E}_{\xi_k}\left[g(w_k, \xi_k)\right]\right\|_2 \leq \mu_G \left\|\nabla F(w_k)\right\|_2 \tag{6.6b}
$$

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

$$
\mathbb{V}_{\xi_k}\left[g(w_k, \xi_k)\right] \leq M + M_V \left\|\nabla F(w_k)\right\|_2^2 \tag{6.7}
$$

Observación Notemos que si g es un estimador insesgado de ∇F entonces 6.6a y 6.6b 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 6.6a como 6.6b valen.

Observación Bajo 6.2.2 y por 6.5 tenemos que:

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

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

Lema 6.2.3 *Bajo 6.2.2 y 6.1.1 las iteraciones de 6.1 satisfacen para todo $k \in \mathbb{N}$:*

$$\mathbb{E}_{\xi_k}\left[F(w_{k+1})\right] - F(w_k) \leq -\mu\alpha_k \left\|\nabla F(w_k)\right\|_2^2 + \frac{1}{2}\alpha_k^2 L \mathbb{E}_{\xi_k}\left[\left\|g(w_k,\xi_k)\right\|_2^2\right] \tag{6.8a}$$

$$\mathbb{E}_{\xi_k}\left[F(w_{k+1})\right] - F(w_k) \leq -\left(\mu - \frac{1}{2}\alpha_k L M_G\right)\alpha_k \left\|\nabla F(w_k)\right\|_2^2 + \frac{1}{2}\alpha_k^2 L M \tag{6.8b}$$

Demostración Por 6.2.1 y 6.6a vale que:

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

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

Que es 6.8a; luego por 6.2 obtenemos 6.8b. ∎

Corolario 6.2.4 *Bajo 6.2.2 y 6.1.1 las iteraciones de 6.1 satisfacen para todo $k \in \mathbb{N}$ que $\{w_k\}$ es una cadena de Markov de primer orden.*

6.3 CASO FUERTEMENTE CONVEXO

Consideremos primero los casos de convexidad donde sabemos que el mínimo existe y es único, por lo tanto asumamos por ahora 2.3.1:

Hipótesis 6.3.1 (Convexidad fuerte) *Supongamos que la función objetivo $F : \mathbb{R}^d \mapsto \mathbb{R}$ es fuertemente convexa, es decir que cumple 2.3.1.*

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

Recordemos que de 2.3.1 y 6 vale que $c \leq L$

Lema 6.3.2 (Fuertemente convexa es PL-convexa) *Supongamos que F cumple 2.3.1, luego para todo $w \in \mathbb{R}^d$ vale que:*

$$2c\left(F(w) - F_{inf}\right) \leq \left\|\nabla F(w)\right\|_2^2 \tag{6.9}$$

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

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

Se puede verificar que $z_* := w - \frac{1}{c}\nabla F(w)$ cumple que $q(z_*) = F(w) - \frac{1}{2c}\left\|\nabla F(w)\right\|_2^2 \leq q(z)$ para todo $z \in \mathbb{R}^d$; luego por 6.3.2 se tiene:

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

∎

Ya estamos en condiciones de demostrar nuestro primer resultado de convergencia para 6.1 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 6.8b 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 6.3.3 (Objetivo fuertemente convexo, Incremento constante) *Supongamos 6.1.1, 6.2.2 y 2.3.1; además supongamos que dado 6.1* $\alpha_k = \alpha > 0$ *constante tal que:*

$$0 < \alpha \leq \frac{\mu}{LM_G} \tag{6.10}$$

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

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

Demostración Usando 6.2.3 con 6.10 y 6.3.2 tenemos para todo $k \in \mathbb{N}$ que:

$$\begin{aligned}
\mathbb{E}_{\xi_k}\left[F(w_{k+1}) - F(w_k)\right] \quad &\underbrace{\leq}_{6.2.3} \quad -\left(\mu - \tfrac{1}{2}\alpha LM_G\right)\alpha\left\|\nabla F(w_k)\right\|_2^2 + \tfrac{1}{2}\alpha^2 LM \\
&\underbrace{\leq}_{6.10} \quad -\tfrac{1}{2}\alpha\mu\left\|\nabla F(w_k)\right\|_2^2 + \tfrac{1}{2}\alpha^2 LM \\
&\underbrace{\leq}_{6.3.2} \quad -\alpha\mu c\left(F(w_k) - F_{inf}\right) + \tfrac{1}{2}\alpha^2 LM
\end{aligned}$$

Luego si restamos F_{inf} y tomamos esperanza total (definida en 6.3:

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

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 4.1.4.

Observación Notemos además que si $M = 0$ (o sea el algortimo 6.1 no tiene ruido) entonces la convergencia es lineal, recuperando el resultado de 2.3.8.

Observación Notemos finalmente que hay un compromiso entre el primer y segundo término de 6.3.3 donde a un α más cercano a $\dfrac{\mu}{LM_G}$ 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 $\mathbb{E}\left[F(w_{k_1}) - F_{inf}\right] \leq \dfrac{\alpha_1 LM}{2c\mu}$. Luego se tomaba $\alpha_2 = \frac{\alpha_1}{2}$ y se seguía inductivamente.

Teorema 6.3.4 (Objetivo fuertemente convexo, Incremento decreciente) *Supongamos 6.1.1, 6.2.2 y 2.3.1; además supongamos que dado 6.1 α_k cumple:*

$$\alpha_k = \frac{\beta}{\gamma + k} \quad \text{para algún } \beta > \frac{1}{c\mu} \text{ y } \gamma > 0 \text{ tal que } \alpha_1 \leq \frac{\mu}{LM_G} \tag{6.11}$$

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

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

Donde:

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

Demostración Notemos primero que por 6.11 para todo $k \in \mathbb{N}$ vale:

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

Luego por 6.2.3 y 6.3.2 uno tiene para todo $k \in \mathbb{N}$:

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

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 6.3.5 (Objetivo Fuertemente Convexo, Reducción del Ruido) *Supongamos que valen 6.1.1, 6.2.2 y 2.3.1 pero reforcemos 6.7 a la existencia de una constante $M \geq 0$ y $\xi \in (0, 1)$ tal que para todo $k \in \mathbb{N}$:*

$$\mathbb{V}_{\xi_k}\left[g(w_k, \xi_k)\right] \leq M\xi^{k-1} \tag{6.12}$$

Supongamos además que 6.1 tiene $\alpha_k = \alpha$ para todo $k \in \mathbb{N}$ satisfaciendo:

$$0 < \alpha \leq \min\left\{\frac{\mu}{L\mu_G^2}, \frac{1}{\mu}\right\} \tag{6.13}$$

Luego vale:

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

Donde:

$$\omega := \max\left\{\frac{\alpha LM}{c\mu}, F(w_1) - F_{inf}\right\} \tag{6.14a}$$

$$\rho := \max\left\{1 - \frac{\alpha c\mu}{2}, \xi\right\} < 1 \tag{6.14b}$$

Demostración Por 6.8a vale que:

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

Luego si juntamos 6.5, 6.6b, 6.13 y 6.12 entonces:

$$\begin{aligned}
\mathbb{E}_{\xi_k}\left[F(w_{k+1})\right] - F(w_k) \leq & \quad -\mu\alpha \left\|\nabla F(w_k)\right\|_2^2 + \frac{1}{2}\alpha^2 L\left(\mu_G^2 \left\|\nabla F(w_k)\right\|_2^2 + M\xi^{k-1}\right) \\
\leq & \quad -\left(\mu - \frac{1}{2}\alpha L \mu_G^2\right)\alpha \left\|\nabla F(w_k)\right\|_2^2 + \frac{1}{2}\alpha^2 L M \xi^{k-1} \\
\leq & \quad -\frac{1}{2}\mu\alpha \left\|\nabla F(w_k)\right\|_2^2 + \frac{1}{2}\alpha^2 L M \xi^{k-1} \\
\leq & \quad -c\mu\alpha\left(F(w_k) - F_{inf}\right) + \frac{1}{2}\alpha^2 L M \xi^{k-1}
\end{aligned}$$

Por lo tanto:

Por que vale k=1??

$$\mathbb{E}\left[F(w_{k+1}) - F_{inf}\right] \leq (1 - c\mu\alpha)\,\mathbb{E}\left[F(w_k) - F_{inf}\right] + \frac{1}{2}\alpha^2 L M \xi^{k-1} \tag{6.15}$$

Probemos ahora la identidad por inducción. Para esto, notemos que el caso $k = 1$ vale por la definición de ω; luego si asumimos que vale para algún $k \geq 1$ entonces de 6.15, 6.14a y 6.14b:

$$\begin{aligned}
\mathbb{E}\left[F(w_{k+1}) - F_{inf}\right] \leq & \quad (1 - c\mu\alpha)\,\omega\rho^{k-1} + \frac{1}{2}\alpha^2 L M \xi^{k-1} \\
= & \quad \omega\rho^{k-1}\left(1 - c\mu\alpha + \frac{\alpha^2 L M}{2\omega}\left(\frac{\xi}{\rho}\right)^{k-1}\right) \\
\leq & \quad \omega\rho^{k-1}\left(1 - c\mu\alpha + \frac{\alpha^2 L M}{2\omega}\right) \\
\leq & \quad \omega\rho^{k-1}\left(1 - c\mu\alpha + \frac{\alpha c\mu}{2}\right) \\
= & \quad \omega\rho^{k-1}\left(1 - \frac{\alpha c\mu}{2}\right) \\
\leq & \quad \omega\rho^k
\end{aligned}$$

∎

6.4 CASO GENERAL

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

Teorema 6.4.1 (Objetivo no convexo, Incrementos fijos) *Asumiendo 6.1.1 y 6.2.2 y suponiendo que en 6.1 tenemos $\alpha_k = \alpha$ tal que:*

$$0 < \alpha \leq \frac{\mu}{LM_G} \tag{6.16}$$

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

$$\mathbb{E}\left[\sum_{k=1}^{K} \|\nabla F(w)_k)\|_2^2\right] \leq \frac{K\alpha LM}{\mu} + \frac{2(F(w_1) - F_{inf})}{\mu\alpha} \tag{6.17a}$$

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K} \|\nabla F(w)_k)\|_2^2\right] \leq \frac{\alpha LM}{\mu} + \frac{2(F(w_1) - F_{inf})}{K\mu\alpha} \xrightarrow{K \to \infty} \frac{\alpha LM}{\mu} \tag{6.17b}$$

Demostración Recordemos 6.8a y si tomamos esperanza total e imponemos 6.16 tenemos:

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

Luego como por 6.2.2 tenemos que $F_{inf} \leq \mathbb{E}[F(w_k)]$ para todo $k \in \mathbb{N}$ vale:

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

∎

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 [3].

En cambio, cuando $M \neq 0$ aunque no podemos acotar $\|\nabla F(w_k)\|_2^2$ *per-se*, podemos decir de 6.17b 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 (6.3.3) 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* 2.3:

Teorema 6.4.2 (Objetivo no convexo, Incrementos decrecientes) *Asumiendo 6.1.1 y 6.2.2, suponiendo además que en 6.1 los $\{\alpha_k\}$ satisfacen 2.3; si notamos $A_K := \sum_{k=1}^{K} \alpha_k$ vale para todo $k \in \mathbb{N}$::*

$$\lim_{K\to\infty} \mathbb{E}\left[\sum_{k=1}^{K} \alpha_k \|\nabla F(w)_k)\|_2^2\right] < \infty \tag{6.18a}$$

$$\mathbb{E}\left[\frac{1}{A_K}\sum_{k=1}^{K}\alpha_k \|\nabla F(w)_k)\|_2^2\right] \xrightarrow{K\to\infty} 0 \tag{6.18b}$$

Demostración Como $\alpha_k \to 0$ por 2.3 entonces podemos asumir sin pérdida de generalidad que $\alpha_k LM_G \leq \mu$ para todo $k \in \mathbb{N}$, luego:

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

Luego como por 6.2.2 tenemos que $F_{inf} \leq \mathbb{E}\left[F(w_k)\right]$ para todo $k \in \mathbb{N}$ vale:

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

Luego:

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

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 6.18a esta probado. Finalmente como por 2.3 tenemos que $A_K \to \infty$ se tiene 6.18b. ∎

Corolario 6.4.3 *Asumiendo 6.1.1 y 6.2.2, suponiendo además que en 6.1 los $\{\alpha_k\}$ satisfacen 2.3 entonces :*

$$\liminf_{k\to\infty}\mathbb{E}\left[\|\nabla F(w_k)\|_2^2\right] = 0 \tag{6.19}$$

Corolario 6.4.4 *Bajo las mismas hipótesis de 6.4.2 sea $k(K) \in \{1, \ldots, K\}$ un índice aleatorio elegido con probabilidades respectivas $\{\alpha_k\}_{k=1}^{K}$; luego $\|\nabla F(w_k)\|_2 \to 0$ en probabilidad.*

Usamos la desigualdad de Markov

Demostración Sea $\epsilon > 0$, luego de 6.18*a* y la desigualdad de Markov:

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

Porque aca es la esperanza de la esperanza condicional?

∎

.

Teorema 6.4.5 (Objetivo no convexo regular, Incrementos decrecientes) *Bajo las mismas hipótesis de 6.4.2 si además pedimos que $F \in C^2$ y que $w \mapsto \|\nabla F(w)\|_2^2$ sea $l-$ Lipshitz entonces:*

$$\lim_{k\to\infty}\mathbb{E}\left[\|\nabla F(w_k)\|_2^2\right] = 0 \tag{6.20}$$

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:

$$\begin{aligned} G(w_{k+1}) - G(w_k) \underbrace{\leq}_{6} & \quad \nabla G(w_k)^T(w_{k+1} - w_k) + \frac{1}{2}L_G \|w_k - w_{k+1}\|_2^2 \\ \leq & \quad -\alpha_k \nabla G(w_k)^T g(w_k, \xi_k) + \frac{1}{2}\alpha_k L_G \|g(w_k, \xi_k)\|_2^2 \end{aligned}$$

Si tomamos esperanza condicional a ξ_k y usamos 6.1.1, 6.2.2 entonces:

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

Luego obtenemos tomando esperanza total:

$$\mathbb{E}[G(w_{k+1})] - \mathbb{E}[G(w_k)] \leq 2\alpha_k L\mu_G \mathbb{E}\left[\|\nabla F(w_k)\|_2^2\right] + \frac{1}{2}\alpha_k^2 L_G \left(M + M_V \mathbb{E}\left[\|\nabla F(w_k)\|_2^2\right]\right) \tag{6.21}$$

Notemos que existe $K \in \mathbb{N}$ tal que $\alpha_k^2 \leq \alpha_k$ y luego por 6.4.2 el lado derecho cumple:

$$\lim_{N\to\infty} 2L\mu_G \underbrace{\sum_{k=K}^{K+N} \mathbb{E}\left[\alpha_k \|\nabla F(w_k)\|_2^2\right]}_{6.18a} + \frac{1}{2}L_G \left(M \underbrace{\sum_{k=K}^{K+N} \alpha_k^2}_{2.3} + M_V \underbrace{\sum_{k=K}^{K+N} \mathbb{E}\left[\alpha_k^2 \|\nabla F(w_k)\|_2^2\right]}_{6.18a} \right) = 0$$

Sean:

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

Luego como en 6.21 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 6.4.3 tenemos $\mathbb{E}\left[\|\nabla F(w_k)\|_2^2\right] = \mathbb{E}[G(w_k)] \to 0$.

∎

7 CONVERGENCIA CTP

> At a purely formal level, one could call probability theory the study of measure spaces with total measure one, but that would be like calling number theory the study of strings of digits which terminate.
>
> — Terence Tao

Ahora que ya analizamos la convergencia en L1 de 6.1 vimos que hay una distinción entre el caso convexo y no convexo; donde en el caso convexo usualmente podemos asegurar convergencia a una cercanía del mínimo mientras que en el no convexo solo asegurabamos la convergencia a un punto crítico. Nuevamente en el estudio de la convergencia *casi todo punto* vamos a separar en esos dos casos y va a volver a aparecer 2.3. Por simpleza en los cálculos vamos a asumir en este capítulo que g es un estimador insesgado de $\nabla F(w_k)$.

7.1 CASO DÉBILMENTE CONVEXO

Hipótesis 7.1.1 (Acotaciones al segundo momento de g) *Supongamos que dada F función objetivo cumple 2.0.1 y g la estimación insesgada del gradiente en 6.1 vale que existen $A, B \geq 0$ tales que:*

$$\mathbb{E}\left[g(w_k, \xi_k)^2\right] \leq A + B\left(w_k - w^*\right)^2$$

Observación Notemos que si F cumple 6.1.1 y g cumple 6.2.2 entonces automáticamente cumplen 7.1.1.

Teorema 7.1.2 (Objetivo débilmente convexo, incrementos decrecientes) *Supongamos 2.0.1, 7.1.1; además supongamos que dado 6.1 α_k cumple 2.3:*

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

$$w_k \xrightarrow[ctp]{k \to \infty} w^* \tag{7.1a}$$

$$\left(w_k - w^*\right) \nabla F(w_k) \xrightarrow[ctp]{k \to \infty} 0 \tag{7.1b}$$

Demostración Vayamos de a pasos como cuando demostramos 2.2.2.

Paso 1 Definamos el proceso estocástico de Lyapunov $h_k := \left(w_k - w^*\right)^2$

Paso 2 Análogamente a casos anteriores notemos que:

$$h_{k+1} - h_k = -2\alpha_k (w_k - w^*) g(w_k, \xi_k) + \alpha_k^2 g(w_k, \xi_k)^2 \qquad (7.2)$$

Usamos filtraciones y el teorema de convergencia de cuasi-martingalas

Como definimos bien la filtracion? Es saber toda esa informacion a tiempo k, como lo notamos?

Definamos ahora la filtración $\mathcal{P}_k = \{\xi_0, \ldots, \xi_{k-1}, w_0, \ldots, w_k, \alpha_0, \ldots, \alpha_k\}$ que determina toda la información a tiempo k **antes** de tomar la muestra ξ_k; luego si tomamos la esperanza condicional a esta filtración:

$$\begin{aligned}
\mathbb{E}\left[h_{k+1} - h_k | \mathcal{P}_k\right] &= -2\alpha_k \mathbb{E}\left[\underbrace{(w_k - w^*)}_{\mathcal{P}_k \text{ medible}} g(w_k, \xi_k) | \mathcal{P}_k\right] + \alpha_k^2 \mathbb{E}\left[g(w_k, \xi_k)^2 | \mathcal{P}_k\right] \\
&= -2\alpha_k (w_k - w^*) \mathbb{E}\left[g(w_k, \xi_k) | \mathcal{P}_k\right] + \alpha_k^2 \mathbb{E}\left[g(w_k, \xi_k)^2 | \mathcal{P}_k\right]
\end{aligned}$$

Pero como ξ_k es **independiente** de $\mathcal{P}_k$ entonces si recordamos que g es insesgado tenemos:

$$\mathbb{E}\left[h_{k+1} - h_k | \mathcal{P}_k\right] = -2\alpha_k (w_k - w^*) \nabla F(w_k) + \alpha_k^2 \mathbb{E}\left[g(w_k, \xi_k)^2 | \mathcal{P}_k\right]$$

Ahora si incorporamos 7.1.1 obtenemos:

$$\mathbb{E}\left[h_{k+1} - \left(1 + \alpha_k^2 B\right) h_k | \mathcal{P}_k\right] \leq -2\alpha_k (w_k - w^*) \nabla F(w_k) + \alpha_k^2 A \qquad (7.3)$$

Definamos ahora las sucesiones auxiliares:

$$\mu_k = \prod_{j=1}^{k-1} \frac{1}{1 + \alpha_j^2 B} \qquad (7.4a)$$

$$h'_k = \mu_k h_k \qquad (7.4b)$$

Por lo que replicando las operaciones en 2.2.2 llegamos a:

$$\mathbb{E}\left[h'_{k+1} - h'_k | \mathcal{P}_k\right] \leq \alpha_k^2 \mu_k A$$

Definamos ahora el *proceso de variaciones positivas asociadas a un proceso* $\{u_k\}$:

$$\delta_k^u := \begin{cases} 1 & \text{si } \mathbb{E}\left[u_{k+1} - u_k | \mathcal{P}_k\right] > 0 \\ 0 & \text{si no} \end{cases} \qquad (7.5)$$

Luego notemos que:

$$\mathbb{E}\left[\delta_k^{h'}\left(h'_{k+1}-h'_k\right)\right] \underbrace{=}_{\text{por definición de } \delta_k^{h'}} \mathbb{E}\left[\delta_k^{h'}\mathbb{E}\left[h'_{k+1}-h'_k|\mathcal{P}_k\right]\right] \leq \alpha_k^2\mu_k A$$

Por el mismo motivo que en 2.2.2 concluímos que:

$$h'_k \geq 0 \tag{7.6a}$$

$$\sum_{k=1}^{\infty}\mathbb{E}\left[\delta_k^{h'}\left(h'_{k+1}-h'_k\right)\right] < \infty \tag{7.6b}$$

Por el teorema de convergencia de cuasi-martingalas concluímos que h'_k converge *ctp*; como $\underbrace{\mu_k}_{\geq 0} \to \mu_\infty > 0$ entonces $\{h_k\}$ converge *ctp*.

Paso 3 Como h_k converge ctp, de 7.3 concluímos que:

$$\sum_{k=1}^{\infty}\alpha_k\left(w_k-w^*\right)\nabla F(w_k) < \infty \qquad ctp$$

Supongamos que $\mathbb{P}\left(\left\{\lim_k h_k > 0\right\}\right) > \tilde{\epsilon}$, luego $\mathbb{P}\left(\left\{\alpha_k\left(w_k-w^*\right)\nabla F(w_k) > C\alpha_k\right\}\right) > \tilde{\epsilon}$ lo que implicaría por 2.3 que $\mathbb{P}\left(\left\{\sum_{k=1}^{\infty}\alpha_k\left(w_k-w^*\right)\nabla F(w_k) = \infty\right\}\right) > \tilde{\epsilon}$; concluímos entonces que:

$$w_k \xrightarrow[ctp]{k\to\infty} w^* \tag{7.7a}$$

$$\left(w_k-w^*\right)\nabla F(w_k) \xrightarrow[ctp]{k\to\infty} 0 \tag{7.7b}$$

∎

7.2 caso no convexo

Vamos a tomar las siguientes hipótesis para probar la convergencia ctp a un punto extremal.

Hipótesis 7.2.1 (Hipótesis caso no convexo) *Sea F una función de costo objetivo y supongamos que el algoritmo 6.1 cumple que g es un estimador insesgado, ie:* $\mathbb{E}\left[g(w_k,\xi_k)\right] = \nabla F(w_k)$*, luego tomemos las siguientes hipótesis:*

1. $F \in C^3$
2. *Existe* $w^* \in \chi$ *tal que* $F_{inf} = F(w^*) \leq F(w)$*, aunque notemos que no necesariamente es único*

3. *$F(w) \geq 0$ para todo $w \in \chi$ (Notemos que como hay un mínimo global, podemos redefinir $\widetilde{F} = F - F_{inf} \geq 0$)*

4. *Sean $\{\alpha_k\}$ los incrementos del algoritmo 6.1, entonces estos cumplen 2.3*

5. *Para $j = 2, 3, 4$ existen $A_j, B_j \geq 0$ tal que:*

$$\mathbb{E}\left[\|g(w_k, \xi_k)\|_2^j\right] \leq A_j + B_j \|w\|_2^j \tag{7.8}$$

6. *Existe $D > 0$ tal que:*

$$\inf_{(w)^2 > D} w \nabla F(w) > 0 \tag{7.9}$$

7.2.1 *Acotación global del algoritmo*

Usemos 7.9 para probar que existe un entorno $w_1 \in U \subset \chi$ tal que si $\{w_k\}$ son las iteraciones del algoritmo 6.1 con F, g cumpliendo 7.2.1, entonces $\{w_k\} \subset U$.

Teorema 7.2.2 (Acotación global del algoritmo estocástico insesgado) *Sea F función de costo objetivo y g un estimador insesgado de ∇F tal que ambos cumplen 7.2.1, luego existe un entorno $w_1 \in U \subset \chi$ tal que si $\{w_k\}$ son las iteraciones del algoritmo 6.1 entonces $\{w_k\} \subset U$.*

Demostración Nuevamente probemos esto en tres pasos:

Paso 1 Sea D el parámetro de horizonte dado por 7.9 y definamos $\phi : \chi \mapsto \mathbb{R}$ dada por:

$$\phi(x) = \begin{cases} 0 & \text{si } x < D \\ (x - D)^2 & \text{si } x \geq D \end{cases}$$

Luego, sea:

$$f_k = \phi\left(w_k^2\right)$$

Paso 2 Notemos que por la definición vale:

$$\phi(y) - \phi(x) \leq (y - x)\,\phi\prime(x) + (y - x)^2$$

Y la igualdad se da si y sólo si $x, y \geq D$; con esto en forma análoga a antes calculemos las variaciones del proceso de Lyapunov $\{f_k\}$:

$$\begin{aligned} f_{k+1} - f_k \leq & \left(-2\alpha_k w_k g(w_k, \xi_k) + \alpha_k^2 g(w_k, \xi_k)^2\right) \phi'(w_k^2) \\ & +4\alpha_k^2 \left(w_k g(w_k, \xi_k)\right)^2 - 4\alpha_k^3 w_k g(w_k, \xi_k)^3 \\ & +4\alpha_k^4 g(w_k, \xi_k)^4 \end{aligned} \quad (7.10)$$

Usamos Cauchy Schwartz

Por Cauchy-Schwartz sabemos que $w_k g(w_k, \xi_k) = \langle w_k, g(w_k, \xi_k)\rangle \leq \|w_k\|_2 \|g(w_k, \xi_k)\|_2$ y que $g(w_k, \xi_k)^2 = \langle g(w_k, \xi_k), g(w_k, \xi_k)\rangle \leq \|g(w_k, \xi_k)\|_2^2$, si sumamos esto a la anterior ecuación tenemos:

$$\begin{aligned} f_{k+1} - f_k \leq & -2\alpha_k w_k g(w_k, \xi_k)\phi'(w_k^2) + \alpha_k^2 \phi'(w_k^2) \|g(w_k, \xi_k)\|_2^2 \\ & +4\alpha_k^2 \|w_k\|_2^2 \|g(w_k, \xi_k)\|_2^2 - 4\alpha_k^3 \|w_k\|_2 \|g(w_k, \xi_k)\|_2^3 \\ & +4\alpha_k^4 \|g(w_k, \xi_k)\|_2^4 \end{aligned}$$

Lo que implica si tomamos esperanza condicional a la filtración $\{\mathcal{P}_k\}$, recordando que $\alpha_k, w_k, \phi'(w_k^2)$ son $\mathcal{P}_k$ medibles y $g(w_k, \xi_k)$ es independiente de $\mathcal{P}_k$:

$$\begin{aligned} \mathbb{E}\left[f_{k+1} - f_k | \mathcal{P}_k\right] \leq & -2\alpha_k w_k \nabla F(w_k)\phi'(w_k^2) + \alpha_k^2 \phi'(w_k^2)\mathbb{E}\left[\|g(w_k, \xi_k)\|_2^2\right] \\ & +4\alpha_k^2 \|w_k\|_2^2 \mathbb{E}\left[\|g(w_k, \xi_k)\|_2^2\right] - 4\alpha_k^3 \|w_k\|_2 \mathbb{E}\left[\|g(w_k, \xi_k)\|_2^3\right] \\ & +4\alpha_k^4 \mathbb{E}\left[\|g(w_k, \xi_k)\|_2^4\right] \end{aligned}$$

Porque pablo? Aca esto no me cierra que tantas acotaciones hace!

Si ahora incluímos 7.8 entonces esto implica que existen $A, B \geq 0$ tal que:

$$\mathbb{E}\left[f_{k+1} - f_k | \mathcal{P}_k\right] \leq -2\alpha_k w_k \nabla F(w_k)\phi'(w_k^2) + \alpha_k^2 \left(A + B f_k\right)$$

Notemos ahora que si $w_k^2 < D$ entonces $\phi'(w_k^2) = 0$, y si $w_k^2 \geq D$ entonces $-2\alpha_k w_k \nabla F(w_k)\phi'(w_k^2) < 0$ por 7.9 por lo que deducimos:

$$\mathbb{E}\left[f_{k+1} - f_k | \mathcal{P}_k\right] \leq \alpha_k^2 \left(A + B f_k\right) \quad (7.11)$$

Dado que hicimos lo mismo 3 veces, valdría que sea un lema?

Ahora siguiendo los mismo pasos que al demostrar 7.1.2 definiendo μ_k, f_k' y usando el teorema de convergencia de cuasi-martingalas concluímos que $\{f_k\}$ converge ctp.

Paso 3 Supongamos que $f_{inf} > 0$, entonces existe $T \in \mathbb{N}$ tal que $w_k^2, w_{k+1}^2 > D$ para todo $k \geq T$ por lo que 7.10 es una igualdad y deducimos que:

$$\sum_{k=1}^{\infty} \alpha_k w_k \nabla F(w_k)\phi'(w_k^2) < \infty \quad ctp \quad (7.12)$$

Pero por otro lado como $f_{inf} > 0$, $\sum_{k=1}^{\infty} \alpha_k = \infty$, $0 < \lim \phi'(w_k^2) < \infty$ existe $\widetilde{\epsilon} > 0$ y $M > 0$ tal que:

$$\begin{aligned} \widetilde{\epsilon} &< \mathbb{P}\left(\left\{\sum_{k=1}^{\infty} \alpha_k w_k \nabla F(w_k)\phi'(w_k^2) = \infty\right\}\right) \\ &= \mathbb{P}\left(\left\{\sum_{k=1}^{\infty} \alpha_k w_k \nabla F(w_k)\phi'(w_k^2) > M \liminf \phi'(w_k^2) \sum_{k=1}^{\infty} \alpha_k\right\}\right) \end{aligned}$$

Concluímos que $f_{inf} = 0$ y entonces existe $K \in \mathbb{N}$ tal que $\{w_k\}_{k \geq K} \subset \{x \in \chi \ : \ \|x\| < D\}$. ∎

7.2.2 *Convergencia del algoritmo*

Teorema 7.2.3 (Convergencia a puntos extremales, Caso insesgado) *Sea F función de costo objetivo y g un estimador insesgado de ∇F tal que ambos cumplen 7.2.1, si $\{w_k\}$ son las iteraciones del algoritmo 6.1 entonces valen:*

$$F(w_k) \xrightarrow[ctp]{k \to \infty} F_\infty \tag{7.13a}$$

$$\nabla F(w_k) \xrightarrow[ctp]{k \to \infty} 0 \tag{7.13b}$$

Demostración Vayamos como siempre de a pasos definiendo el proceso de Lyapunov correspondiente:

Paso 1 Definamos $h_k = F(w_k) \geq 0$ por hipótesis

Paso 2 Por 7.2.2 si suponemos que $\dim(\chi) < \infty$ entonces existe $K_1 > 0$ tal que $\nabla^2 F(w_k) \leq K_1$, luego si desarrollamos en Taylor en w_k tenemos:

$$\begin{aligned} h_{k+1} - h_k = & \quad \nabla F(w_k)\,(w_{k+1} - w_k) + \frac{1}{2}\nabla^2 F(w_k)\,(w_{k+1} - w_k) \\ = & \; -\alpha_k \nabla F(w_k) g(w_k, \xi_k)^2 + \frac{1}{2}\nabla^2 F(w_k)\alpha_k^2 g(w_k, \xi_k)^2 \\ \leq & \qquad -2\alpha_k \nabla F(w_k) g(w_k, \xi_k)^2 + K_1 \alpha_k^2 g(w_k, \xi_k)^2 \end{aligned}$$

Tomando esperanza condicional respecto a $\mathcal{P}_k$:

$$\mathbb{E}\left[h_{k+1} - h_k | \mathcal{P}_k\right] \leq \underbrace{-2\alpha_k \left(\nabla F(w_k)\right)^2}_{\leq 0} + \alpha_k^2 K_1 \underbrace{\mathbb{E}\left[g(w_k, \xi_k)\right]}_{\leq K_2 \text{ por } 7.2.2} \tag{7.14}$$

Lo que implica:

$$\mathbb{E}\left[h_{k+1} - h_k | \mathcal{P}_k\right] \leq \alpha_k^2 K_1 K_2 \tag{7.15}$$

Luego como tenemos de esto que:

$$\begin{aligned} \sum_{k=1}^{\infty} \mathbb{E}\left[\delta_k^h \left(h_{k+1} - h_k\right)\right] &= \sum_{k=1}^{\infty} \mathbb{E}\left[\delta_k^h \mathbb{E}\left[\left(h_{k+1} - h_k\right) | \mathcal{P}_k\right]\right] \\ &\leq K_1 K_2 \sum_{k=1}^{\infty} \alpha_k^2 < \infty \end{aligned}$$

Por lo que por el teorema de convergencia de cuasi martingalas obtenemos:

$$F(w_k) \xrightarrow[\text{ctp}]{k \to \infty} F_\infty \tag{7.16}$$

Paso 3 Si retomamos 7.14, reordenamos, sumamos hasta k y tomamos esperanza tenemos:

$$\begin{aligned} 2 \sum_{k=1}^{K} \alpha_k \left(\nabla F(w_k)\right)^2 &\leq \sum_{k=1}^{K} \mathbb{E}\left[h_{k+1} - h_k\right] + K_2 K_1 \sum_{k=1}^{K} \alpha_k^2 \\ &= \mathbb{E}\left[\sum_{k=1}^{K} h_{k+1} - h_k\right] + K_2 K_1 \sum_{k=1}^{K} \alpha_k^2 \\ &= \mathbb{E}\left[h_{k+1}\right] + K_2 K_1 \sum_{k=1}^{K} \alpha_k^2 \\ &\xrightarrow[\text{CDL}]{\text{ctp}} F_\infty + K_2 K_1 \sum_{k=1}^{\infty} \alpha_k^2 \end{aligned} \tag{7.17}$$

Usamos el teorema de convergencia dominada en esperanza condicional

Sea ahora $g_k = \left(\nabla F(w_k)\right)^2$ y volvamos a expandir Taylor en w_k:

$$\begin{aligned} g_{k+1} - g_k &= \nabla g(w_k)\left(w_{k+1} - w_k\right) + \frac{1}{2} \nabla^2 g(w_k)\left(w_{k+1} - w_k\right)^2 \\ &\underbrace{\leq}_{\text{por } 7.2.2} -2\alpha_k \nabla F(w_k) K_4 g(w_k, \xi_k) + K_3 \alpha_k^2 g(w_k, \xi_k)^2 \end{aligned}$$

Tomando esperanza condicional respecto a $\mathcal{P}_k$:

$$\mathbb{E}\left[g_{k+1} - g_k | \mathcal{P}_k\right] \leq 2\alpha_k K_4 \left(\nabla F(w_k)\right)^2 + K_2 K_3 \alpha_k^2 \tag{7.18}$$

Por lo tanto si sumamos las variaciones asociadas al proceso estocástico $\{g_k\}$:

$$\begin{aligned}\sum_{k=1}^{\infty}\mathbb{E}\left[\delta_k^g\left(g_{k+1}-g_k\right)\right] &= \sum_{k=1}^{\infty}\mathbb{E}\left[\delta_k^g\mathbb{E}\left[(g_{k+1}-g_k)|\mathcal{P}_k\right]\right]\\ &\leq K_4\underbrace{\sum_{k=1}^{\infty}\alpha_k\left(\nabla F(w_k)\right)^2}_{7.17}+K_2K_3\underbrace{\sum_{k=1}^{\infty}\alpha_k^2}_{2.3}\\ &< \infty\end{aligned} \tag{7.19}$$

Nuevamente por el teorema de convergencia de cuasi martingalas concluímos que $\{g_k\}$ converge *ctp*, y por 7.17 este límite es 0; luego:

$$g_k \xrightarrow[\text{ctp}]{k\to\infty} 0 \tag{7.20a}$$

$$\nabla F(w_k) \xrightarrow[\text{ctp}]{k\to\infty} 0 \tag{7.20b}$$

∎

Part IV

Apéndice

APÉNDICE A

A.1 PROPOSICIONES ENUNCIADAS

Teorema A.1.1 *Dados $y_0 < y_1$, valores $f(y_0), f(y_1)$ y sus derivadas $f'(y_0), f'(y_1)$ con $f'(y_0) < 0$ el polinomio cúbico interpolante de Hermite se define por:*

$$p(y) = c_0 + c_1\delta_y + c_2\delta_y^2 + c_3\delta_y^3 \tag{A.1}$$

Donde:

$$\begin{aligned}
y &\in [y_0, y_1] \\
c_0 &= f(y_0) \\
c_1 &= f'(y_0) \\
c_2 &= \frac{3S - f'(y_0) - 2f'(y_0)}{y_1 - y_0} \\
c_3 &= -\frac{2S - f'(y_1) - f'(y_0)}{(y_1 - y_0)^2} \\
\delta_y &= y - y_0 \\
S &= \frac{f(y_1) - f(y_0)}{y_1 - y_0}
\end{aligned}$$

Y $p(y)$ satisface $p(y_0) = f(y_0)$, $p(y_1) = f(y_1)$, $p'(y_0) = f'(y_0)$ y $p'(y_1) = f'(y_1)$; además si $f(y_1) < f(y_0) < 0$ y:

$$f'(y_1) \geq \frac{3\left(f(y_1) - f(y_0)\right)}{y_1 - y_0}$$

Entonces para $y \in [y_0, y_1]$ vale que $p(y) \in [f(y_1), f(y_0)]$

Demostración Ver [4]

Teorema A.1.2 *Sea $E \subset \mathbb{R}^d$ y dotemos a $C^m(E)$ de la norma $\|f\|_{C^m} = \sup\{\|\partial^\alpha f|_E\|_\infty : |\alpha| < m\}$, si E es cerrado en $\mathbb{R}^d$ entonces existe $T \in L\left(C^m(E), C^m\left(\mathbb{R}^d\right)\right)$ tal que $T(f)|_E = f$ y $T(f) \in C^\infty(E^c)$. Es más, $\|T\| \leq C(m)d^{\frac{5m}{2}}$.*

Demostración Ver [1]

A.2 DEMOSTRACIONES

Demostración [De 2.3.2] TODO

Demostración [De 2.3.4] TODO

Demostración [De 2.3.7] TODO

Demostración [De 2.2.1] Definamos:

$$S_t^+ := \sum_{k=1}^{t-1} (u_{k+1} - u_k)_+ \tag{A.2a}$$

$$S_t^- := \sum_{k=1}^{t-1} (u_{k+1} - u_k)_- \tag{A.2b}$$

Donde recordemos que $(x)_\pm = x1_{\{\mathbb{R}_\pm\}}$. Como sabemos que $(u_{k+1} - u_k)_+ \geq 0$ para todo $k \in \mathbb{N}$ entonces $S_t^+ \nearrow S_\infty^+$; asimismo, $(u_{k+1} - u_k)_- \leq 0$ para todo $k \in \mathbb{N}$ entonces $S_t^- \leq 0$. Por lo tanto:

$$0 \leq u_k = u_0 + S_k^+ + S_k^- \leq u_0 + S_\infty^+ \tag{A.3a}$$

$$-u_0 - S_\infty^+ \leq S_k^- \leq 0 \tag{A.3b}$$

Luego como $S_{k+1}^- \leq S_k^-$ concluímos que $S_k^- \searrow S_\infty^-$. Por lo tanto como S_k^+, S_k^- convergen entonces $u_k = u_0 + S_k^+ + S_k^-$ converge. ∎

[1] Cheng Alan. «The Whitney extension theorem in high dimensions.» In: (2015).

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

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

[4] Alan S Edelman Randall L Dougherty and James M Hyman. *Nonnegativity, monotonicity, or convexity preserving cubic and quintic Hermite interpolation*. Vol. 52(186). Mathematics of Computation, 1989, pp. 471–494.

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

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

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