



MÉTODOS DE PRIMER ORDEN

ANÁLISIS DE CONVERGENCIA

Universidad de Buenos Aires

Facultad de Ciencias Exactas y Naturales  
Departamento de Matemática

Tesis de Licenciatura  
Director de Tesis: Dr. Pablo Amster  
Septiembre 2018 – version 0.2



*Ohana* means family.  
Family means nobody gets left behind, or forgotten.  
— Lilo & Stitch

Dedicado a las 2 chicas que marcan mi Norte, Patri y Sophia

## 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 [7]

## AGRADECIMIENTOS

---

Agradecimientos para todos



## CONTENTS

---

<b>I</b>	<b>Introducción</b>	<b>1</b>
1	INTRODUCCIÓN	3
1.1	Organización de la Tesis	3
2	PRELIMINARES	5
2.1	Espectro	5
2.2	Variedades diferenciables y Teorema de la variedad estable	5
2.2.1	Un repaso por Variedades	5
2.2.2	Teorema de la variedad estable	7
2.3	Preliminares de Procesos Estocásticos	7
2.3.1	Esperanza condicional	8
2.3.2	Martingalas y Cuasi-martingalas	10
<b>II</b>	<b>Algoritmos de tipo Batch</b>	<b>13</b>
3	CONVERGENCIA PUNTUAL	15
3.1	Intuición	15
3.2	Caso discreto	16
3.3	Acerca de convexidad fuerte y funciones L-Lipshitz	18
4	TEOREMA DE LA VARIEDAD ESTABLE Y LOS PUNTOS FIJOS INESTABLES	23
4.1	Intuición	23
4.2	Puntos fijos inestables	26
5	CONVERGENCIA CTP A MÍNIMOS : CASO GENERAL	29
5.1	Descenso de Gradiente en Batch	29
5.2	Punto Próximo	30
5.3	Descenso por coordenadas	31
6	RESULTADOS NEGATIVOS	35
6.1	Ejemplos <i>patológicos</i>	35
<b>III</b>	<b>Algoritmos Estocásticos</b>	<b>43</b>
7	CONVERGENCIA EN L1	45
7.1	Contexto	45
7.2	Algunos lemas fundamentales	46
7.3	Caso Fuertemente Convexo	48

7.4	Caso general	52
8	CONVERGENCIA CTP	57
8.1	Caso débilmente convexo	57
8.2	Caso no convexo	59
8.2.1	Acotación global del algoritmo	60
8.2.2	Convergencia del algoritmo	62

## IV Apéndice 65

A	APÉNDICE	67
A.1	Proposiciones enunciadas	67
A.2	Demostraciones	67
	BIBLIOGRAFÍA	69



## Part I

# Introducción



## INTRODUCCIÓN

---

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

---

"David Hilbert"

### 1.1 ORGANIZACIÓN DE LA TESIS

Esta Tesis esta organizada según la categorización mas estándar de los algoritmos encontrados en Machine Learning.

La primer parte (I) refiere a la motivación tanto matemática como algoritmica y del área para analizar la convergencia de los algoritmos presentados, como a su vez los contenidos preliminares usados a lo largo del documento.

La segunda parte (II) trata exclusivamente los algoritmos determinísticos, comunmente denominados de *tipo batch*. En el Capítulo 3, utilizando la gran referencia [11], analizamos la convergencia puntual del descenso de gradiente con condiciones de convexidad débil y luego la convergencia *lineal* con convexidad mas fuerte. Luego en el capítulo 4 nos basamos en [6] para ver a estos algoritmos como discretizaciones de sistemas dinámicos y con el teorema de la variedad estable concluimos un método práctico para analizar la convergencia *casi todo punto*; esta forma de analizar los algoritmos es aplicada en el capítulo 5 con varios algoritmos estándar en el área. Finalmente en el capítulo 6 nos basamos en [4] para ver que aunque se tiene convergencia *casi todo punto*, la complejidad algoritmica del descenso de gradiente es exponencial.

Esto nos motiva pasar a analizar los algoritmos estocásticos en la parte III, donde nos basamos en [2] y [3] para analizar en el capítulo 7 la convergencia en *norma L1* al mínimo (o a un entorno de éste) y finalmente en el capítulo 8 la convergencia *casi todo punto* tanto en los casos convexo como no convexo, ya que el algoritmo induce un proceso estocástico que induce una *cuasi-martingala* convergente.



## PRELIMINARES

---

Mathematics is the most beautiful  
and most powerful creation of the  
human spirit

---

Stefan Banach

## 2.1 ESPECTRO

**Definición 2.1.1** Sea  $f : X \rightarrow Y$ , con  $X, Y$  Banach,  $f \in L(X, Y)$ ; definimos el espectro de  $f$  de la siguiente manera:

$$\sigma(f) = \{\alpha \in \mathbb{C} : f - \alpha \text{ no es inversible}\}$$

Al supremo del espectro le decimos radio espectral (e.g.  $\rho(f) = \sup \{|\alpha| : \alpha \in \sigma(f)\}$ )

**Proposición 2.1.2** Sea  $f \in L(X, Y)$  un operador lineal acotado, entonces:

$$\rho(f) = \lim \|f^n\|^{\frac{1}{n}}$$

**Proposición 2.1.3** Sea  $f \in L(X, Y)$  un operador lineal acotado y  $h \in \text{Hol}(U)$  con  $\sigma(f) \subset U$  una función holomorfa en un entorno del espectro, entonces:

$$\sigma(h(f)) = h(\sigma(f))$$

## 2.2 VARIEDADES DIFERENCIABLES Y TEOREMA DE LA VARIEDAD ESTABLE

Cuando estudiemos los algoritmos de tipo batch es normal analizar al algoritmo como  $x_{k+1} \leftarrow g(x_k)$  para una  $g : X \rightarrow X$  inducida; o sea uno analiza las órbitas bajo la acción de  $g$  en una variedad dada  $X$ . Con esa motivación repasemos los conceptos básicos de sistemas dinámicos.

## 2.2.1 Un repaso por Variedades

**Definición 2.2.1** [Capítulo 1 de [8]] Dado un espacio topológico  $X$  decimos que es una variedad diferenciable de dimensión  $d$  si:

- $X$  es Hausdorff

- $X$  es 2—contable
- Existe un atlas suave para  $X$ , o sea existe un conjunto de pares  $\{(U_i, \phi_i)\}$  tales que:
  1. Para todo  $x \in X$  existe  $(U, \phi)$  con  $x \in U$  y  $\phi : U \rightarrow \phi U$  homeomorfismo
  2. Si existen dos cartas  $(U, \phi), (V, \psi)$  en el entorno de  $x$  con  $U \cap V \neq \emptyset$  entonces  $\phi \circ \psi^{-1} : \psi(U \cap V) \rightarrow \phi(U \cap V)$  es difeomorfismo

**Definición 2.2.2** [Capítulo 6 de [8]] Dada una variedad de dimensión  $d$   $\chi$  y el espacio de medida  $(\mathbb{R}^d, \mathcal{B}, \mu)$ , decimos que  $E \subset \chi$  tiene medida cero si existe un atlas  $\mathcal{A} = \{U_i, \phi^i\}_{i \in \mathbb{N}}$  tal que  $\mu(\phi^i(E \cap U_i)) = 0$ . En este caso usamos el abuso de notación  $\mu(E) = 0$ .

**Definición 2.2.3** [Capítulo 3 de [8]] El diferencial de  $g$  es un operador lineal  $D_g(x) : \mathcal{T}_x \mapsto \mathcal{T}_{g(x)}$ , donde  $\mathcal{T}_x$  es el espacio tangente de  $X$  en el punto  $x$ . Dada una curva  $\gamma$  en  $X$  con  $\gamma(0) = x$  y  $d\gamma(0) = v \in \mathcal{T}_x$ , el operador lineal se define como  $D_g(x)v = \frac{d(g \circ \gamma)}{dt}(0) \in \mathcal{T}_{g(x)}$ . El determinante del operador lineal  $\det(D_g(x))$  es el determinante de la matriz que representa  $D_g(x)$  con respecto a una base arbitraria.

**Proposición 2.2.4** Sea  $X$  una variedad de dimensión  $d$ , luego para todo  $x \in X$  vale que  $\mathcal{T}_x$  es un espacio vectorial de dimensión  $d$ .

**Teorema 2.2.5** Sea  $F : X \mapsto N$  una función diferenciable tal que  $dF_x$  es un isomorfismo lineal, luego existe  $x \in U \subset X$  abierto tal que  $F|_U$  es un difeomorfismo e. g.  $F, F^{-1} \in C^\infty(U)$

**Proposición 2.2.6** Sea  $f : \mathbb{R}^d \rightarrow \mathbb{R}^d$  una función localmente Lipshitz, luego si  $\mu(E) = 0$  vale que  $\mu(f(E)) = 0$

**Lema 2.2.7** Sea  $E \subset \chi$  tal que  $\mu(E) = 0$ ; si  $\det(Dg(x)) \neq 0$  para todo  $x \in \chi$ , luego  $\mu(g^{-1}(E)) = 0$

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

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

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

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

Por hipótesis  $\varphi^i(E_i)$  es de medida cero, luego como  $g$  es difeomorfismo local por 2.2.5 entonces  $\psi \circ h \circ \varphi^{i-1} \in C^1$ . Como si  $f \in C^1(\mathbb{R}^d)$  entonces es localmente Lipschitz concluimos por 2.2.6 que  $\mu(\psi \circ h \circ \varphi^{i-1}(\varphi^i(E_i))) = 0$  para todo  $i \in \mathbb{N}$ . ■

Concluimos con un resultado natural, pero no por eso menos crucial a la hora de ver la probabilidad de un conjunto dado en  $X$ .

**Proposición 2.2.8** Sea  $N \hookrightarrow M$  una subvariedad de dimensión  $n < m$ , luego para todo  $U \subset N$  abierto relativo vale que  $\mu(U) = 0$

### 2.2.2 Teorema de la variedad estable

**Definición 2.2.9** Sea  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  tal que  $f \in C^2$ , luego:

- Un punto  $x^*$  es crítico de  $f$  si  $\nabla f(x^*) = 0$
- Un mínimo local si  $x^*$  es crítico y  $\lambda_{\min}(\nabla^2 f(x^*)) > 0$
- Un punto silla estricto de  $f$  si es crítico y  $\lambda_{\min}(\nabla^2 f(x^*)) < 0$

Notaremos  $\chi^*$  al conjunto de puntos silla estrictos de  $f$ .

**Teorema 2.2.10** Sea  $x^*$  un punto fijo de  $g \in C^r(\chi)$  un difeomorfismo local. Supongamos que  $E = E_s \oplus E_u$  donde

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

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

**Demostración** Ver Teorema III.1 de [16]

## 2.3 PRELIMINARES DE PROCESOS ESTOCÁSTICOS

Debido a la naturaleza de los algoritmos estocásticos, tiene sentido repasar los conceptos básicos que utilizaremos en el estudio de ellos.

## 2.3.1 Esperanza condicional

Dado un espacio de probabilidad  $(\Omega, \mathcal{F}, P)$  definimos una *variable aleatoria* como una función  $X : \Omega \mapsto \mathbb{R}$  tal que  $X^{-1}(B) \in \mathcal{F}$  para todo  $B \in \mathcal{B}$  boreliano.

Por otro lado, dado un conjunto  $\Omega$  y una familia  $(X_\gamma)_{\gamma \in C}$  tal que  $X_\gamma : \Omega \mapsto \mathbb{R}$  definimos la *sigma algebra generada por las  $X_\gamma$*   $\mathcal{F} = \sigma(X_\gamma)$  como la menor sigma álgebra (en el sentido de la inclusión) tal que todas las  $X_\gamma$  son  $\mathcal{F}$  medibles.

Recordemos además:

**Teorema 2.3.1 (Teorema de la convergencia monótona)** Sea  $(f_n)$  una sucesión positiva de elementos medibles en  $(\Omega, \Sigma, \mu)$  un espacio de medida tal que  $f_n \nearrow f$ ; luego:

$$\int_{\Omega} f_n d\mu \nearrow \int_{\Omega} f d\mu$$

**Teorema 2.3.2 (Teorema de la convergencia dominada)** Sea  $(f_n)$  una sucesión de elementos medibles en  $(\Omega, \Sigma, \mu)$  un espacio de medida tal que  $f_n \nearrow f$  ctp; si existe  $g \in L^1$  tal que  $|f_n| \leq g$  entonces:

$$\int_{\Omega} |f_n - f| d\mu \rightarrow 0$$

**Observación** Una observación clave en el análisis de los algoritmos de tipo estocástico es que si llamamos  $W := \{w_k\}$  a las iteraciones del algoritmo, entonces notemos que podemos ver a  $w_k$  como una variable aleatoria. En efecto,  $w_k := w_{k-1} - \alpha_k g(w_{k-1}, \cdot) : \Omega \rightarrow \mathbb{R}$  y por hipótesis es  $\mathcal{F}$  medible; es más, podemos ver a  $W$  como un proceso estocástico discreto.

**Proposición 2.3.3** Sea  $(\Omega, \mathcal{F}, P)$  un espacio de probabilidad y  $X$  una variable aleatoria tal que  $X \in L^1(\Omega)$  ( $E.g.E[|X|] < \infty$ ). Si  $\mathcal{G}$  es una sub- $\sigma$ -álgebra de  $\mathcal{F}$  entonces existe  $Y$  una variable aleatoria tal que:

1.  $Y$  es  $\mathcal{G}$  medible
2.  $Y \in L^1(\Omega)$
3. Para todo  $G \in \mathcal{G}$  vale:

$$\int_G Y dP = \int_G X dP$$

Es más, si  $\tilde{Y}$  es otra variable aleatoria que cumple las propiedades, entonces  $Y = \tilde{Y}$  ctp.



**Definición 2.3.4** Dados  $X, \mathcal{G}$  como en la proposición, a la variable aleatoria cuya existencia se prueba en 2.3.3 se le llama una versión de la esperanza condicional de  $X$  dado  $\mathcal{G}$  y se lo nota  $\mathbb{E}[X|\mathcal{G}]$ .

A su vez, dada  $Z$  otra variable aleatoria definimos la esperanza condicional de  $X$  dado  $Z$  como  $\mathbb{E}[X|Z] := \mathbb{E}[X|\sigma(Z)]$

**Demostración** Demostremos la existencia y unicidad:

Unicidad ctp Sea  $X \in L^1$  e  $Y, \tilde{Y}$  dos versiones de  $\mathbb{E}[X|\mathcal{G}]$  tal que no son iguales ctp, luego como  $\left\{Y - \tilde{Y} > \frac{1}{n}\right\} \nearrow \left\{Y > \tilde{Y}\right\}$  existe  $N$  tal que:

$$P\left(Y - \tilde{Y} > \frac{1}{N}\right) > 0$$

Luego como  $Y, \tilde{Y}$  son  $\mathcal{G}$  medibles  $\left\{Y - \tilde{Y} > \frac{1}{N}\right\} \in \mathcal{G}$  y entonces:

$$0 \underbrace{=} \int_{Y, \tilde{Y} = \mathbb{E}[X|\mathcal{G}]\left\{Y - \tilde{Y} > \frac{1}{N}\right\}} Y - \tilde{Y} \geq \frac{1}{N} P\left(\left\{Y - \tilde{Y} > \frac{1}{N}\right\}\right) > 0$$

Luego  $Y = \tilde{Y}$  ctp.

Existencia en  $L^2$  Sea  $\mathcal{K} = L^2(\Omega, \mathcal{G}, P)$ , sabemos que  $\mathcal{K}$  es completo en  $L^2$  y como  $L^2$  es Hilbert existe  $Y \in \mathcal{K}$  tal que:

$$\mathbb{E}\left[(X - Y)^2\right] = \inf \left\{ \mathbb{E}\left[(X - W)^2\right] : W \in \mathcal{K} \right\}$$

$$\langle X - Y, Z \rangle = 0 \quad Z \in \mathcal{K}$$

Luego si  $G \in \mathcal{G}$  entonces  $Z = 1_G \in \mathcal{G}$  por lo que:

$$\langle X - Y, 1_G \rangle = 0 \implies \int_G X dP = \int_G Y dP$$

Concluimos que  $Y = \mathbb{E}[X|\mathcal{G}]$

Existencia en  $L^1$  Notemos que basta verlo para  $X \geq 0$ , luego existen  $X_n \geq 0$  acotadas tal que  $X_n \nearrow X$ ; como cada  $X_n \in L^2$  si definimos  $Y = \limsup \mathbb{E}[X_n|\mathcal{G}]$  entonces 2.3.1 demuestra lo que necesitábamos. ■

**Teorema 2.3.5 (Propiedades de la esperanza condicional)** Sea  $(\Omega, \mathcal{F}, P)$  un espacio de medida,  $X \in L^1$  y  $\mathcal{G}, \mathcal{H}$  sub- $\sigma$ -álgebras de  $\mathcal{F}$ , luego:

$$1. \mathbb{E}[\mathbb{E}[X|\mathcal{G}]] = \mathbb{E}[X]$$

2. Si  $X$  es  $\mathcal{G}$  medible entonces  $X = \mathbb{E}[X|\mathcal{G}]$  ctp
3.  $\mathbb{E}[aX + bY|\mathcal{G}] = a\mathbb{E}[X|\mathcal{G}] + b\mathbb{E}[Y|\mathcal{G}]$
4. Si  $X \geq 0$  ctp, entonces  $\mathbb{E}[X|\mathcal{G}] \geq 0$  ctp
5. Si  $0 \leq X_n \nearrow X$  entonces  $\mathbb{E}[X_n|\mathcal{G}] \nearrow \mathbb{E}[X|\mathcal{G}]$
6. Si  $0 \leq X_n$  ctp entonces  $\mathbb{E}[\liminf X_n|\mathcal{G}] \leq \liminf \mathbb{E}[X_n|\mathcal{G}]$
7. Si  $|X_n| \leq V$  con  $V \in L^1$  entonces si  $X_n \rightarrow X$  ctp vale que  $\mathbb{E}[X_n|\mathcal{G}] \rightarrow \mathbb{E}[X|\mathcal{G}]$  ctp.
8. Si  $\mathcal{H}$  es una sub- $\sigma$ -álgebra de  $\mathcal{G}$  entonces:

$$\mathbb{E}[X|\mathcal{G}|\mathcal{H}] := \mathbb{E}[\mathbb{E}[X|\mathcal{G}]|\mathcal{H}] = \mathbb{E}[X|\mathcal{H}]$$

9. Si  $Z$  es  $\mathcal{G}$  medible y acotada entonces  $\mathbb{E}[ZX|\mathcal{G}] = Z\mathbb{E}[X|\mathcal{G}]$  ctp
10. Si  $\mathcal{H}$  es independiente de  $\sigma(\sigma(X), \mathcal{G})$  entonces:

$$\mathbb{E}[X|\sigma(\mathcal{G}, \mathcal{H})] = \mathbb{E}[X|\mathcal{G}] \text{ ctp}$$

En particular, si  $X$  es independiente de  $\mathcal{G}$  vale que  $\mathbb{E}[X|\mathcal{G}] = \mathbb{E}[X]$  ctp.

**Demostración** En general con cuentas bastante similares a las vistas en cualquier curso de Análisis Real teniendo cuidado de las proyecciones. Para una mejor referencia ver [17].

### 2.3.2 Martingalas y Cuasi-martingalas

**Definición 2.3.6** Dada una sucesión creciente (en el sentido de la inclusión) de  $\sigma$  álgebras  $\mathcal{F}_n \subset \mathcal{F}$  decimos que  $\{\mathcal{F}_n\}$  es una filtración y que el espacio  $(\Omega, \mathcal{F}, \{F_n\}, P)$  es un espacio filtrado.

Un proceso  $X = (X_n)$  decimos que es adaptado si  $X_n$  es  $\mathcal{F}_n$  medible para todo  $n$  en un espacio filtrado.

A su vez, dado un proceso  $(X_n)$ , éste induce una filtración (llamada natural) en un espacio de probabilidad  $(\Omega, \mathcal{F}, P)$  dada por  $\mathcal{F}_n = \sigma(X_1, \dots, X_n)$

**Definición 2.3.7** Dado un espacio filtrado, decimos que un proceso  $X = (X_n)$  es una martingala relativa a la filtración  $\{\mathcal{F}_n\}$  si:

- $X$  es adaptado
- $X_n \in L^1$
- $\mathbb{E}[X_n|\mathcal{F}_{n-1}] = X_{n-1}$  ctp

A su vez decimos que es una (sub)supermartingala si vale la condición  $\mathbb{E}[X_n|\mathcal{F}_{n-1}] (\leq) \geq X_{n-1}$

**Definición 2.3.8** Dado un espacio filtrado y  $X$  un proceso integrable y adaptado, decimos que es una cuasi-martingala si para todo  $n \in \mathbb{N}$ :

$$\mathbb{V}_n(X) = \sup_{J \subset \{1, \dots, n\}} \left\{ \mathbb{E} \left[ \sum_{\substack{i_k \in J \\ 1 \leq k \leq |J|}} |\mathbb{E}[X_{i_{k+1}} - X_{i_k} | \mathcal{F}_{i_k}]| \right] \right\} < \infty \quad (2.1)$$

**Observación** El concepto de cuasi-martingalas es una generalización natural de las martingalas, submartingalas y supermartingalas. Fueron introducidos por primera vez por Fisk [5] para extender la descomposición de Doob-Meyer a una clase más grande de procesos. La forma en que las cuasimartingalas se relacionan con sub y súper martingalas es muy similar a cómo las funciones de variación finita se relacionan con funciones crecientes y decrecientes. En particular, mediante la descomposición de Jordan, cualquier función de variación finita en un intervalo se descompone como la suma de una función creciente y una función decreciente. De manera similar, un proceso estocástico es una cuasimartingala si y solo si puede escribirse como la suma de una submartingala y una supermartingala. Este importante resultado fue mostrado primero por Rao [13], y significó el inicio de la extensión de gran parte de la teoría de submartingalas a cuasimartingalas.

**Proposición 2.3.9** Toda martingala, submartingala o supermartingala es una cuasi-martingala

**Demostración** En efecto, reemplazando  $X$  por  $-X$  podemos suponer que  $\mathbb{E}[X_{i_{k+1}} - X_{i_k} | \mathcal{F}_{i_k}] \geq 0$ , luego por 2.3.5 resulta que  $\mathbb{V}_n(X) = |\mathbb{E}[X_n - X_0]| < \infty$ . ■

Ahora estamos en condiciones de enunciar el resultado principal de esta sección: **El teorema de convergencia de cuasi-martingalas**. Este resultado es crucial en el análisis de convergencia de algoritmos estocásticos porque veremos más adelante que el proceso estocástico  $\{w_k\}$  inducido por el algoritmo induce una cuasi-martingala  $\{w'_k\}$  que será convergente; de lo cual deduciremos la convergencia de  $\{w_k\}$ .

**Definición 2.3.10** Dado un proceso estocástico  $\{u_k\}$  adaptado a un espacio filtrado  $(\Omega, \mathcal{F}, \{\mathcal{P}_k\}, P)$  definimos el proceso de variaciones positivas asociadas a  $\{u_k\}$  como :

$$\delta_k^u := \begin{cases} 1 & \text{si } \mathbb{E}[u_{k+1} - u_k | \mathcal{P}_k] > 0 \\ 0 & \text{si no} \end{cases} \quad (2.2)$$

**Teorema 2.3.11 (Teorema de convergencia de cuasi-martingalas)** Dado un proceso estocástico  $\{u_k\}$  adaptado a un espacio filtrado  $(\Omega, \mathcal{F}, \{\mathcal{P}_k\}, P)$  tal que:

- $u_k \geq 0$  ctp
- $\sum_{k=1}^{\infty} \mathbb{E} [\delta_k^u (u_{k+1} - u_k)] < \infty$

Entonces  $\{u_k\}$  es una cuasi-martingala tal que  $u_k \rightarrow u_{\infty} \geq 0$  ctp.

**Demostración** Una buena revisión de la demostración se encuentra en el capítulo 9 de [10]

## 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



## 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 3.1]. Un buen inicio es analizar la convergencia puntual del descenso de gradiente y bajo que condiciones se da.

**Algoritmo 3.1 :** Descenso de gradiente en batch

<p><b>1 Input:</b> <math>F \in C^1</math>, <math>\alpha_k &gt; 0</math>, <math>w_1 \in \mathbb{R}^d</math>, <math>X = \{\xi_j\}_{j \leq N}</math> muestra</p> <p><b>2 for</b> <math>k \in \mathbb{N}</math> <b>do</b></p> <p><b>3</b>     <math>w_{k+1} \leftarrow w_k - \alpha_k \sum_{j=1}^N \nabla F(\xi_j)</math></p>
---

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

**Definición 3.0.1** 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 3.0.1.

### 3.1 INTUICIÓN

Ganemos intuición acerca del proceso como probar la convergencia del algoritmo 3.1 en el caso continuo. En el caso continuo, tenemos que demostrar que la solución  $w(t)$  de la ecuación diferencial 3.1 tiene límite  $w^*$  y además que  $w^*$  es mínimo de  $F$ .

$$\frac{dw}{dt} = -\nabla F(w) \quad (3.1)$$

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

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 3.1.1 (Objetivo débilmente convexo, Versión continua)**

Sea  $F \in C^1$  que cumple 3.0.1 y supongamos que el algoritmo 3.1 cumple  $\alpha_k = \alpha > 0$  para  $w(t)$  continua. Luego si notamos al mínimo de  $F$  como  $w^*$ , vale:

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

**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:

$$\frac{dh}{dt} = 2(w(t) - w^*) \frac{dw}{dt} = -2(w(t) - w^*) \nabla F(w) \stackrel{\text{3.0.1}}{\leq} 0 \quad (3.2)$$

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

Paso 3 Como  $h(t) \searrow h_{inf}$  entonces  $\frac{dh}{dt} \rightarrow 0$ , supongamos por el absurdo que  $h_{inf} > 0$ , luego existe  $\tilde{\epsilon} > 0$  y  $T \in \mathbb{R}$  tal que para todo  $t \geq T$  vale que  $h(t) = (w(t) - w^*)^2 > \tilde{\epsilon}$ . Si juntamos entonces 3.2 y 3.0.1 llegamos a un absurdo, concluimos que  $h_{inf} = 0$  por lo que:

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

■

### 3.2 CASO DISCRETO

Ahora sí, analicemos la convergencia del algoritmo 3.1 para el caso discreto. Para esto enunciamos un lema útil cuya demostración referimos al lector al Apéndice:

**Lema 3.2.1** Sea  $\{u_k\}_{k \in \mathbb{N}} \subset \mathbb{R}$  una sucesión tal que  $u_k \geq 0$  para todo  $k$ . Luego si:

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



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  $u_\infty = \lim_{k \rightarrow \infty} u_k = u_0 + S_\infty^+ + S_\infty^-$

**Demostración** Ver [A](#)

Consideremos ahora el algoritmo [3.1](#), decimos que los incrementos  $\{\alpha_k\}$  cumplen la condición de *Robbins - Monro* (ver [\[15\]](#)) si:

$$\sum_{k=1}^{\infty} \alpha_k = \infty \quad \text{y} \quad \sum_{k=1}^{\infty} \alpha_k^2 < \infty \quad (3.3)$$

**Teorema 3.2.2 (Objetivo débilmente convexo, incrementos decrecientes)**

Sea  $F \in C^1$ , asumamos [3.0.1](#) y que existen  $A, B \geq 0$  tal que para todo  $w \in \mathbb{R}^d$  vale que:

$$(\nabla F(w))^2 \leq A + B(w - w^*)^2 \quad (3.4)$$

Luego si consideramos el algoritmo [3.1](#) tal que incrementos  $\{\alpha_k\}$  cumplen [3.3](#) entonces:

$$w_k \xrightarrow{k \rightarrow \infty} w^* \quad (3.5)$$

**Demostración** Hagamos los 3 pasos análogos a [3.1.1](#):

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

Paso 2 Análogo a [3.1.1](#) 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 [3.3](#) y [3.4](#) entonces:

$$h_{k+1} - (1 + \alpha_k^2 B) h_k \leq \underbrace{-2\alpha_k (w_k - w^*) \nabla F(w_k)}_{\leq 0 \text{ por } 3.0.1} + \alpha_k^2 A \leq \alpha_k^2 A$$

Definamos ahora las sucesiones auxiliares:

$$\mu_k = \prod_{j=1}^{k-1} \frac{1}{1 + \alpha_j^2 B} \quad (3.6a)$$

$$h'_k = \mu_k h_k \quad (3.6b)$$

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  $\mu_k$  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 3.2 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 3.2.1 concluimos que  $\{h'_k\}$  converge; como  $\underbrace{\mu_k}_{\geq 0} \rightarrow \mu_{\infty} > 0$  entonces  $\{h_k\}$  converge.

Paso 3 De 3.2 como ya vimos que  $h_{k+1} - (1 + \alpha_k^2 B) h_k$  es sumable concluimos que:

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

Supongamos que  $h_k \rightarrow h_{inf} \neq 0$ , luego existiría  $K \in \mathbb{N}$  y  $\tilde{\epsilon} > 0$  tal que  $h_k = (w_k - w^*)^2 > \tilde{\epsilon}$  para todo  $k \geq K$ ; luego de 3.0.1 concluimos que existe  $M > 0$  tal que  $M \leq (w_k - w^*) \nabla F(w_k)$  para todo  $k \geq K$ . Por 3.3 eso implica que  $\sum_{k=1}^{\infty} \alpha_k (w_k - w^*) \nabla F(w_k) = \infty$ , concluimos que  $w_k \xrightarrow{k \rightarrow \infty} w^*$ . ■

**Corolario 3.2.3** Sea  $F \in C^2$ , asumamos 3.0.1 y que  $\|\nabla^2 F\|_2^2 \leq L$ ; si consideramos el algoritmo 3.1 tal que incrementos  $\{\alpha_k\}$  cumplen 3.3 entonces:

$$w_k \xrightarrow{k \rightarrow \infty} w^* \quad (3.7)$$

### 3.3 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 3.3.1** Sea  $f \in C^1$ , decimos que es fuertemente convexa o  $\mu$ -convexa si existe  $\mu > 0$  tal que para todos  $x, y \in \mathbb{R}^d$  vale:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2} \|y - x\|_2^2 \quad (3.8)$$

**Proposición 3.3.2** 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.  $f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y) - \frac{\alpha(1 - \alpha)\mu}{2} \|y - x\|_2^2$  para todos  $x, y \in \mathbb{R}^d, \alpha \in [0, 1]$

**Demostración** Ver [A](#)

**Definición 3.3.3** Decimos que una función  $f \in C^1$  es PL-convexa, o cumple la condición de Polyak-Lojasiewicz (ver [\[12\]](#), [\[9\]](#)) si existe  $\mu > 0$  tal que para todo  $x \in \mathbb{R}^d$  vale:

$$\frac{1}{2} \|\nabla f(x)\|_2^2 \geq \mu (f(x) - f_{\inf}) \quad (3.9)$$

**Proposición 3.3.4** 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 3.3.5** 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:

$$\begin{aligned} h(\alpha x + (1 - \alpha)y) &= f(\alpha x + (1 - \alpha)y) + g(\alpha x + (1 - \alpha)y) \\ &\stackrel{3.3.2}{\leq} \underbrace{\alpha(f + g)(x) + (1 - \alpha)(f + g)(y)}_{3.3.2} - \frac{\mu\alpha(1 - \alpha)}{2} \|x - y\|_2^2 \\ &= \alpha h(x) + (1 - \alpha)h(y) - \frac{\mu\alpha(1 - \alpha)}{2} \|x - y\|_2^2 \end{aligned}$$

■

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

**Definición 3.3.6** 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:

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L \|y - x\|_2 \quad (3.10)$$

**Proposición 3.3.7** 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.  $f(\alpha x + (1 - \alpha)y) \geq \alpha f(x) + (1 - \alpha)f(y) - \frac{\alpha(1 - \alpha)}{2L} \|y - x\|_2^2$
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.  $f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y) - \frac{\alpha(1 - \alpha)}{2L} \|\nabla f(y) - \nabla f(x)\|_2^2$

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 [\[12\]](#)

**Teorema 3.3.8 (Convergencia lineal , Objetivos L-Lipshitz y PL-convexos)**

Sea  $F \in C^1$  tal que existe  $F_{inf}$  valor mínimo,  $F$  cumple [3.3.6](#) y [3.3.3](#); entonces el algortimo [3.1](#) con incremento fijo  $\alpha_k = \frac{1}{L}$  cumple:

$$f(w_k) - f_{inf} \leq \left(1 - \frac{\mu}{L}\right)^k (f(w_0) - f_{inf}) \quad (3.11)$$

**Demostración** Notemos que si usamos [3.3.7](#) 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 [3.3.3](#) tenemos:

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

Luego obtenemos:

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

■



## 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

### 4.1 INTUICIÓN

Del capítulo anterior ya sabemos que bajo condiciones de convexidad estándar el algoritmo 3.1 converge puntualmente. Nos surge entonces la pregunta:

**Bajo que casos el algoritmo 3.1 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 3.1) convergan a puntos silla. Sea  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  dada por  $f(x) = \frac{1}{2}x^T H x$  con  $H = \text{diag}(\lambda_1, \dots, \lambda_n)$ ; supongamos además que  $\lambda_1, \dots, \lambda_k > 0$  y  $\lambda_{k+1}, \dots, \lambda_n < 0$ .

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

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

Por lo tanto:

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

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

Recordemos que si  $g(x) = x - \alpha \nabla f(x)$  entonces el algoritmo 3.1 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:

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

Por lo tanto por inducción es fácil probar que:

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

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

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

Con lo que concluimos que:

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

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

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

**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:

$$\begin{aligned} \nabla f &= (x, y^3 - y) \\ g &= ((1 - \alpha)x, (1 + \alpha)y - \alpha y^3) \\ \nabla^2 f &= \begin{pmatrix} 1 & 0 \\ 0 & 3y^2 - 1 \end{pmatrix} \end{aligned} \tag{4.1}$$

De lo que vemos que los puntos críticos son:

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

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

**Lema 4.1.1** 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 4.1, luego:

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

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

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

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

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

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

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

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

## 4.2 PUNTOS FIJOS INESTABLES

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,  $g : \chi \rightarrow \chi$  y  $\chi$  es una  $d$ -variedad sin borde.

**Definición 4.2.1** Sea:

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

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

Con los resultados de 2 demostremos el teorema principal para analizar la convergencia de los algoritmos de tipo batch en el caso no convexo:

**Teorema 4.2.2** Sea  $g \in C^1(\chi)$  tal que  $\det(Dg(x)) \neq 0$  para todo  $x \in \chi$ , luego el conjunto de puntos iniciales que convergen por  $g$  a un punto fijo inestable tiene medida cero, i. e.:

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

**Demostración** Para cada  $x^* \in \mathcal{A}_g^*$  por 4.2.2 existe  $B_{x^*}$  un entorno abierto; es más,  $\bigcup_{x^* \in \mathcal{A}_g^*} B_{x^*}$  forma un cubrimiento abierto del cual existe un subcubrimiento numerable pues  $X$  es variedad, i. e.

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

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  $W := \left\{ x_0 : \lim_k x_k \in \mathcal{A}_g^* \right\}$  queremos ver que  $\mu(W) = 0$ .

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

**Lema 4.2.3**  $x_t \in \bigcap_{k \in \mathbb{N}} g^{-k}(B_{x_i^*})$  para todo  $t \geq T$

Pablo: Hace falta demostrar esto??

Si notamos  $S_i \triangleq \bigcap_{k \in \mathbb{N}} g^{-k}(B_{x_i^*})$ , entonces por 2.2.10 sabemos por un lado que es una subvariedad de  $W_{loc}^{cs}$  y por el otro que  $\dim(S_i) \leq \dim(W_{loc}^{cs}) = \dim(E_s) < d - 1$ ; por lo que por 2.2.8  $\mu(S_i) = 0$ .

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

$$\begin{aligned} \mu(W) &\leq \mu\left(\bigcup_{i \in \mathbb{N}} \bigcup_{k \in \mathbb{N}} g^{-k}(S_i)\right) \\ &\leq \sum_{i \in \mathbb{N}} \sum_{k \in \mathbb{N}} \mu(g^{-k}(S_i)) \\ &\stackrel{2.2.7}{=} 0 \end{aligned}$$

■

Para finalizar veamos un caso simple que nos encontraremos seguido:

**Corolario 4.2.4** *Bajo las mismas hipótesis que en 4.2.2 si agregamos que  $\chi^* \subseteq \mathcal{A}_g^*$  entonces  $\mu(W_g) = 0$*

**Demostración** Como  $\chi^* \subseteq \mathcal{A}_g^*$  entonces  $W_g \subseteq W$ , luego  $\mu(W_g) \leq \mu(W) \stackrel{4.2.2}{=} 0$ . ■



## CONVERGENCIA CTP A MÍNIMOS : CASO GENERAL

### 5.1 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 4.2.2 demostraremos que el *descenso de gradiente en batch* tiene probabilidad cero de converger a puntos silla. Consideremos el algoritmo 3.1 con incrementos constantes  $\alpha_k = \alpha$ :

$$x_{k+1} = g(x_k) \triangleq x_k - \alpha \nabla f(x_k) \quad (5.1)$$

**Hipótesis 5.1.1** Asumamos que  $f \in \mathcal{C}^2$  y  $\|\nabla^2 f(x)\|_2 \leq L$

**Proposición 5.1.2** 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 por 2.1.3 los autovalores de  $Dg$  son  $\{1 - \alpha \lambda_i : \lambda_i \in \{\mu : \nabla^2 f(x^*)v = \mu v \text{ para algún } v \neq 0\}\}$ . Como  $x^* \in \chi^*$  existe  $\lambda_{j^*} < 0$  por lo que  $1 - \alpha \lambda_{j^*} > 1$ ; concluimos que  $x^* \in \mathcal{A}_g^*$ . ■

**Proposición 5.1.3** Bajo 5.1.1 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:

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

Luego por 5.1.1 tenemos que  $\alpha < \frac{1}{|\lambda_i|}$  y entonces  $1 - \alpha \lambda_i > 0$  para todo  $i \in \{1, \dots, d\}$ ; concluimos que  $\det(Dg(x)) > 0$ . ■

**Corolario 5.1.4** Sea  $g$  dada por el algoritmo 3.1, bajo 5.1.1 y  $\alpha < \frac{1}{L}$  se tiene que  $\mu(W_g) = 0$ .

**Demostración** Por 5.1.2 y 5.1.3 tenemos que vale 4.2.4 y concluimos que  $\mu(W_g) = 0$ . ■

## 5.2 PUNTO PRÓXIMO

El algoritmo de punto próximo esta dado por la iteración:

$$x_{k+1} = g(x_k) \triangleq \arg \min_{z \in \chi} f(z) + \frac{1}{2\alpha} \|x_k - z\|_2^2 \quad (5.2)$$

**Proposición 5.2.1** Bajo 5.1.1 y  $\alpha < \frac{1}{L}$  entonces vale:

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

*Probamos esto? Me parece un poco claro*

**Demostración** Veamos primero el siguiente lema:

**Lema 5.2.2** Bajo 5.1.1,  $\alpha < \frac{1}{L}$  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 5.2.2 podemos tomar límite, i. e.

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

Finalmente por diferenciación implícita obtenemos:

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

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

**Corolario 5.2.3** Sea  $g$  dado por el algoritmo de punto próximo con ecuación 5.2, bajo 5.1.1 y  $\alpha < \frac{1}{L}$  se tiene que  $\mu(W_g) = 0$ .

**Demostración** Por 5.2.1 tenemos que vale 4.2.4 y concluimos que  $\mu(W_g) = 0$ . ■

## 5.3 DESCENSO POR COORDENADAS

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

Consideremos el algoritmo 5.1:

<b>Algorithmus 5.1</b> : Descenso por coordenadas	
1	<b>Input:</b> $f \in C^1, \alpha > 0, x_0 \in \mathcal{X}$
2	<b>for</b> $k \in \mathbb{N}$ <b>do</b>
3	<b>for block</b> $i = 1, \dots, b$ <b>do</b>
4	<b>for index</b> $j \in S_i$ <b>do</b>
5	$y_k^{S_0} = x_k$ e $y_k^{S_i} = (x_{k+1}^{S_1}, \dots, x_{k+1}^{S_i}, x_k^{S_{i+1}}, \dots, x_k^{S_b})$
6	$x_{k+1}^j \leftarrow x_k^j - \alpha \frac{\partial f}{\partial x_j} (y_k^{S_{i-1}})$

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

**Lema 5.3.1** La iteración de Descenso por coordenadas esta dada por:

$$x_{k+1} = g(x_k) \triangleq g_d \circ g_{d-1} \circ \dots \circ g_1(x) \quad (5.3)$$

**Lema 5.3.2** Si  $g$  está dada por 5.3 entonces si notamos  $P_S = \sum_{i \in S} e_i e_i^T$  entonces:

$$Dg(x_k) = \prod_{i \in \{1, \dots, b\}} \left( Id - \alpha P_{b-i+1} \nabla^2 f(y_k^{S_{b-i}}) \right) \quad (5.4)$$

**Demostración** Notemos primero que:

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

Por lo tanto:

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

■

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

**Proposición 5.3.3** Bajo 6 y  $\alpha < \frac{1}{L_b}$  se tiene que  $\det(Dg(x)) \neq 0$

**Demostración** Basta probar que cada término de 5.4 es invertible, para eso:

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

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

**Proposición 5.3.4** Bajo 6 y  $\alpha < \frac{1}{L_{max}}$  se tiene que  $\chi^* \subseteq \mathcal{A}_g^*$

**Demostración** Sea  $x^* \in \chi^*$ ,  $H = \nabla^2 f(x^*)$ ,  $J = Dg(x^*) = \prod_{i \leq b} (Id_n - \alpha P_{S_{b-i+1}} H)$  e  $y_0$  el autovector correspondiente al menor autovalor de  $H$ . Vamos a probar que  $\|J^t y_0\|_2 \geq c(1 + \epsilon)^t$  por lo que  $\|J^t\|_2 \geq c(1 + \epsilon)^t$ , luego por 2.1.2:

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

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

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

**Afirmación 5.3.5** Sea  $y_t \in \text{Ran}(H)$ , luego existe  $i \in \{1, \dots, b\}$  y  $\delta > 0$  tal que  $\alpha \sum_{j \in S_i} |e_j^T H z_i| \geq \delta \|z_i\|_2$

**Lema 5.3.6** 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:

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



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

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

$$z_{i+1}^T H z_{i+1} < z_i^T H z_i - \alpha \left\| \sum_{j \in S_i} (e_j^T H z_i) e_j \right\|_2^2 \quad (5.5)$$

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

$$w^T H w \geq \lambda_{\min}(H) \|w\|_2^2 \geq -L_b \|w\|_2^2 \quad (5.6)$$

Luego si usamos 5.3.5, 5.6 y Cauchy-Schwartz existe  $i \in \{1, \dots, b\}$  y  $\delta > 0$  tal que:

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

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

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

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

■

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

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

Luego:

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

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

■

**Corolario 5.3.7** Sea  $g$  dado por el algoritmo de descenso por coordenadas con ecuación 5.3, bajo 6 y  $\alpha < \frac{1}{L_b}$  se tiene que  $\mu(W_g) = 0$ .

**Demostración** Por 5.3.3 y 5.3.4 tenemos que vale 4.2.4 y concluimos que  $\mu(W_g) = 0$ . ■

## RESULTADOS NEGATIVOS

Ya vimos de 5.1.4, 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?**

### 6.1 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 3.1 con  $\alpha_k = \alpha = \frac{1}{4}$  nos queda:

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

Luego si tomamos  $\epsilon > 0$  y  $w_0 = (x_1^0, x_2^0)$  uniformemente en  $w_0 \in \tilde{U} = [-1,1] \times \left[ -\frac{3}{2}^{-e^{\frac{1}{\epsilon}}}, \frac{3}{2}^{-e^{\frac{1}{\epsilon}}} \right]$  entonces el algoritmo 3.1 necesita  $k \geq e^{\frac{1}{\epsilon}}$  pasos para que  $w_k \notin U$ . Concluimos que el algoritmo es exponencial en converger a cualquier mínimo si  $w_0 \in \tilde{U}$ . ■

**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 3.1 tomamos  $\alpha_k = \alpha = \frac{1}{4}$  tendríamos la siguiente dinámica:

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

Luego si tomamos  $R > 0$  grande y  $w_0 = (x_1^0, x_2^0)$  uniformemente en  $w_0 \in \tilde{U} = [-R-1, -R+1] \times [-1, 1]$  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  $U = [-1, 1]^2$ ; concluimos que el algoritmo es exponencial en converger a cualquier mínimo si  $w_0 \in \tilde{U}$ . ■

**Teorema 6.1.1 (Convergencia exponencial, Inicialización uniforme en el cubo)**

Consideremos el algoritmo 3.1 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  $B, l, \mu \in \text{poly}(d)$  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:

$$f(x_1, x_2) = \begin{cases} -\gamma x_1^2 + Lx_2^2 & \text{si } (x_1, x_2) \in [0, 1] \times [0, 1] \\ L(x_1 - 2)^2 - \gamma x_2^2 & \text{si } (x_1, x_2) \in [1, 3] \times [0, 1] \\ L(x_1 - 2)^2 + L(x_2 - 2)^2 & \text{si } (x_1, x_2) \in [1, 3] \times [1, 3] \end{cases} \quad (6.1)$$

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  $w_0 = (x_1^0, x_2^0) \in U$  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  $k_2 = Ck_1$  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 3.1:

$$\begin{aligned} (x_1^{k+1}, x_2^{k+1}) &= \begin{cases} (x_1^k, x_2^k) - \alpha (-2\gamma x_1^k, 2Lx_2^k) & \text{si } x_1 \leq 1 \\ (x_1^k, x_2^k) - \alpha (2L(x_1^k - 2), -2\gamma x_2^k) & \text{si } x_1 \geq 1, x_2 \leq 1 \\ (x_1^k, x_2^k) - \alpha (2L(x_1^k - 2), 2L(x_2^k - 2)) & \text{si } x_1 \geq 1, x_2 \geq 1 \end{cases} \\ &= \begin{cases} ((1 + 2\alpha\gamma)x_1^k, (1 - \alpha 2L)x_2^k) & \text{si } x_1 \leq 1 \\ ((1 - 2L\alpha)x_1^k + 4L\alpha, (1 + 2\alpha\gamma)x_2^k) & \text{si } x_1 \geq 1, x_2 \leq 1 \\ ((1 - 2L\alpha)x_1^k + 4L\alpha, (1 - 2L\alpha)x_2^k + 4L\alpha) & \text{si } x_1 \geq 1, x_2 \geq 1 \end{cases} \end{aligned}$$

Luego evaluando en  $k_1$  y  $k_2$ :

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

Concluimos que:

$$k_2 \geq \frac{2\alpha(L + \gamma)k_1 - \log(x_2^0)}{2\alpha\gamma} \geq \frac{L + \gamma}{\gamma} k_1 \quad (6.3)$$

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 proxivamente; 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 complejidad del algoritmo 3.1 sea exponencial. Sea  $D_0$  dado por:

$$\begin{aligned} D_0 &= \bigcup_{i=1}^{d+1} \{x \in \mathbb{R}^d : 6\tau \geq x_1, \dots, x_{i-1} \geq 2\tau; 2\tau \geq x_i \geq 0; \tau \geq x_{i+1}, \dots, x_d \geq 0\} \\ &:= \bigcup_{i=1}^{d+1} D_i \end{aligned} \quad (6.4)$$

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:

$$f|_{D_i}(x) = \begin{cases} \sum_{j=1}^{i-1} L(x_j - 4\tau)^2 - \gamma x_i^2 + \sum_{j=i+1}^d Lx_j^2 - (i-1)\eta \\ \triangleq f_{i,1}(x) \text{ si } x \in D_{i,1} \\ \\ \sum_{j=1}^{i-1} L(x_j - 4\tau)^2 + g(x_i, x_{i+1}) + \sum_{j=i+2}^d Lx_j^2 - (i-1)\eta \\ \triangleq f_{i,2}(x) \text{ si } x \in D_{i,2} \end{cases} \quad (6.5)$$

Donde nuevamente  $\eta$  esta pendiente de definición y  $g : \mathbb{R}^2 \rightarrow \mathbb{R}$  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:

$$f|_{D_d}(x) = \begin{cases} \sum_{j=1}^{d-1} L(x_j - 4\tau)^2 - \gamma x_d^2 - (d-1)\eta \\ \triangleq f_{d,1}(x) \text{ si } x \in D_{d,1} \\ \\ \sum_{j=1}^{d-1} L(x_j - 4\tau)^2 + g_1(x_d) - (d-1)\eta \\ \triangleq f_{d,2}(x) \text{ si } x \in D_{d,2} \end{cases} \quad (6.6)$$

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:

$$f|_{D_{d+1}} = \sum_{j=1}^d L(x_j - 4\tau)^2 - d\eta \triangleq f_{d+1,1} \quad (6.7)$$

**Lema 6.1.2** 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:

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

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. Concluimos 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  $T_k^\tau$**

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  $T_k = \min_{x_k^t \geq 2\tau}$  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  $T_k^\tau$  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  $|x^k - x^{k+1}| \geq 2\alpha\gamma\tau$  por lo que:

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

**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 [3.1](#) sabemos que en  $D_{1,2}$  vale la relación:

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

Tenemos que:

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

**Paso 4 - El algoritmo 3.1 se queda confinado a  $D_0$ :**

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

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

Separaremos el caso  $j = k+1$ , donde el lema 6.1.2 nos dice que:

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

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

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

Y concluimos 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  $T_k^\tau$ :

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

Por el otro, usando el mismo argumento 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  $\alpha < \frac{1}{2L}$ :



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

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

Inductivamente:

$$T_d \geq T_d - T_d^\tau \geq \left( \frac{L + \gamma}{\gamma} \right)^{d-1} (T_1 - T_1^\tau) \geq \frac{1}{2\alpha\gamma} \left( \frac{L + \gamma}{\gamma} \right)^{d-1} \geq \left( \frac{L + \gamma}{\gamma} \right)^{d-1} \quad (6.8)$$

**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 3.1 necesita tiempo exponencial para converger al mínimo, ataquemos el caso  $x^0 \in [-1, 1]^d \cap D_0^c$

Para  $a = 0, \dots, 2^d - 1$  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:

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

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

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

$$\begin{aligned} \tilde{D}_{i,1} &= \{x \in \mathbb{R}^d : 6\tau \geq |x_1|, \dots, |x_{i-1}| \geq 2\tau, |x_i| \leq \tau, |x_{i+1}|, \dots, |x_d| \leq \tau\} \\ \tilde{D}_{i,2} &= \{x \in \mathbb{R}^d : 6\tau \geq |x_1|, \dots, |x_{i-1}| \geq 2\tau, \tau \leq |x_i| \leq 2\tau, |x_{i+1}|, \dots, |x_d| \leq \tau\} \\ \tilde{D}_{d+1} &= \{x \in \mathbb{R}^d : 6\tau \geq |x_1|, \dots, |x_d| \geq 2\tau\} \end{aligned}$$

Luego definimos:

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

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 3.1 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  $(4\tau, \dots, 4\tau)$ .

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

Por A.1.2 sabemos que existe  $F : \mathbb{R}^d \rightarrow \mathbb{R}$  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.



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

Jerry Bona

## 7.1 CONTEXTO

En esta parte vamos a analizar la convergencia en  $L_1$  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  $\xi$ , cuya distribución desconocida es  $P$ , luego  $F$  se representa como:

$$F(w) = \begin{cases} R(w) = \mathbb{E}[f(w, \xi)] \\ \text{o} \\ R_n(w) = \frac{1}{n} \sum_{i=1}^n f_i(w) \end{cases} \quad (7.1)$$

Sea el algoritmo estocástico 7.1

**Algorithmus 7.1 : Descenso Estocastico (DE)**

```

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

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

$$g(w_k, \xi_k) = \begin{cases} \nabla f(w_k, \xi_k) \\ \frac{1}{n_k} \sum_{i=1}^{n_k} \nabla f(w_k, \xi_{k,i}) \\ H_k \frac{1}{n_k} \sum_{i=1}^{n_k} \nabla f(w_k, \xi_{k,i}) \end{cases} \quad (7.2)$$

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

Para iniciar el análisis de la convergencia, lo mínimo que necesitamos es que el gradiente se mantenga controlado, por lo tanto recordemos la condición 3.3.6:

**Hipótesis 7.1.1 (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 7.1.1, 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 3.3.7

## 7.2 ALGUNOS LEMAS FUNDAMENTALES

Con el contexto claro, veamos algunos lemas que van a ser clave en la demostración de la convergencia L1 del algoritmo 7.1.

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

**Lema 7.2.1** Bajo 7.1.1 las iteraciones de 7.1 satisfacen que para todo  $k \in N$ :

$$\mathbb{E}_{\xi_k} [F(w_{k+1})] - F(w_k) \leq -\alpha_k \nabla F(w_k)^T \mathbb{E}_{\xi_k} [g(w_k, \xi_k)] + \frac{1}{2} \alpha_k^2 \mathbb{E}_{\xi_k} [\|g(w_k, \xi_k)\|_2^2] \quad (7.3)$$

**Demostración** Notemos que por 7.1.1 vale que:

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

Luego tomando esperanza de ambos lados y recordando 2.3.5:

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

■

**Observación** Notemos que si  $g(w_k, \xi_k)$  es un estimador insesgado de  $\nabla F(w_k)$  entonces de 7.2.1:

$$\mathbb{E}_{\xi_k} [F(w_{k+1})] - F(w_k) \leq -\alpha_k \|\nabla F(w_k)\|^2 + \frac{1}{2} \alpha_k^2 \mathbb{E}_{\xi_k} [\|g(w_k, \xi_k)\|_2^2] \quad (7.4)$$

Luego entonces para controlar la convergencia de 7.1 también hay que poner suposiciones sobre el segundo momento de  $g$ , luego si definimos:

$$\mathbb{V}_{\xi_k} [g(w_k, \xi_k)] := \mathbb{E}_{\xi_k} [\|g(w_k, \xi_k)\|_2^2] - \|\mathbb{E}_{\xi_k} [g(w_k, \xi_k)]\|_2^2 \quad (7.5)$$

Asumamos:

**Hipótesis 7.2.2 (Acotaciones al primer y segundo momento de  $g$ )** Supongamos que dada  $F$  función objetivo y  $g$  la estimación del gradiente en 7.1 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  $k \in \mathbb{N}$  valen:

$$\nabla F(w_k)^T \mathbb{E}_{\xi_k} [g(w_k, \xi_k)] \geq \mu \|\nabla F(w_k)\|_2^2 \quad (7.6a)$$

Y

$$\|\mathbb{E}_{\xi_k} [g(w_k, \xi_k)]\|_2 \leq \mu_G \|\nabla F(w_k)\|_2 \quad (7.6b)$$

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

$$\mathbb{V}_{\xi_k} [g(w_k, \xi_k)] \leq M + M_V \|\nabla F(w_k)\|_2^2 \quad (7.7)$$

**Observación** Notemos que si  $g$  es un estimador insesgado de  $\nabla F$  entonces 7.6a y 7.6b 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  $\xi_k$  entonces tanto 7.6a como 7.6b valen.

**Observación** Bajo 7.2.2 y por 7.5 tenemos que:

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

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

**Lema 7.2.3** Bajo 7.2.2 y 7.1.1 las iteraciones de 7.1 satisfacen para todo  $k \in \mathbb{N}$ :

$$\mathbb{E}_{\xi_k} [F(w_{k+1})] - F(w_k) \leq -\mu \alpha_k \|\nabla F(w_k)\|_2^2 + \frac{1}{2} \alpha_k^2 L \mathbb{E}_{\xi_k} [\|g(w_k, \xi_k)\|_2^2] \quad (7.8a)$$

$$\mathbb{E}_{\xi_k} [F(w_{k+1})] - F(w_k) \leq -\left(\mu - \frac{1}{2} \alpha_k L M_G\right) \alpha_k \|\nabla F(w_k)\|_2^2 + \frac{1}{2} \alpha_k^2 L M \quad (7.8b)$$

**Demostración** Por 7.2.1 y 7.6a vale que:

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

Que es 7.8a; luego por 7.2 obtenemos 7.8b. ■

**Corolario 7.2.4** Bajo 7.2.2 y 7.1.1 las iteraciones de 7.1 satisfacen para todo  $k \in \mathbb{N}$  que  $\{w_k\}$  es una cadena de Markov de primer orden.

### 7.3 CASO FUERTEMENTE CONVEXO

Consideremos primero los casos de convexidad donde sabemos que el mínimo existe y es único, por lo tanto asumamos por ahora 3.3.1:

**Hipótesis 7.3.1 (Convexidad fuerte)** Supongamos que la función objetivo  $F : \mathbb{R}^d \mapsto \mathbb{R}$  es fuertemente convexa, es decir que cumple 3.3.1.

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 3.3.1 y 6 vale que  $c \leq L$

**Lema 7.3.2 (Fuertemente convexa es PL-convexa)** Supongamos que  $F$  cumple 3.3.1, luego para todo  $w \in \mathbb{R}^d$  vale que:

$$2c (F(w) - F_{inf}) \leq \|\nabla F(w)\|_2^2 \quad (7.9)$$

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

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

Se puede verificar que  $z_* := w - \frac{1}{c} \nabla F(w)$  cumple que  $q(z_*) = F(w) - \frac{1}{2c} \|\nabla F(w)\|_2^2 \leq q(z)$  para todo  $z \in \mathbb{R}^d$ ; luego por 7.3.2 se tiene:

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

■

Ya estamos en condiciones de demostrar nuestro primer resultado de convergencia para 7.1 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 7.8b se ve que el segundo término es constante.

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

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



**Teorema 7.3.3 (Objetivo fuertemente convexo, Incremento constante)**

Supongamos 7.1.1, 7.2.2 y 3.3.1; además supongamos que dado 7.1  $\alpha_k = \alpha > 0$  constante tal que:

$$0 < \alpha \leq \frac{\mu}{LM_G} \quad (7.10)$$

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

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

**Demostración** Usando 7.2.3 con 7.10 y 7.3.2 tenemos para todo  $k \in \mathbb{N}$  que:

$$\begin{aligned} \mathbb{E}_{\xi_k} [F(w_{k+1}) - F(w_k)] &\stackrel{7.2.3}{\leq} -(\mu - \tfrac{1}{2}\alpha LM_G) \alpha \|\nabla F(w_k)\|_2^2 + \tfrac{1}{2}\alpha^2 LM \\ &\stackrel{7.10}{\leq} -\tfrac{1}{2}\alpha\mu \|\nabla F(w_k)\|_2^2 + \tfrac{1}{2}\alpha^2 LM \\ &\stackrel{7.3.2}{\leq} -\alpha\mu c (F(w_k) - F_{inf}) + \tfrac{1}{2}\alpha^2 LM \end{aligned}$$

Luego si restamos  $F_{inf}$  y tomamos esperanza total (definida en 7.3):

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

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:

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

■

**Observación** Notemos que si  $g$  es un estimador insesgado de  $\nabla F$  entonces  $\mu = M_G = 1$  por lo que  $\alpha \in [0, \frac{1}{L})$  que es la condición que pedimos en 5.1.4.

**Observación** Notemos además que si  $M = 0$  (o sea el algoritmo 7.1 no tiene ruido) entonces la convergencia es lineal, recuperando el resultado de 3.3.8.

**Observación** Notemos finalmente que hay un compromiso entre el primer y segundo término de 7.3.3 donde a un  $\alpha$  más cercano a  $\frac{\mu}{LM_G}$  acelera la convergencia del primer término, pero a costa de un entorno final de mayor volumen.

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  $\mathbb{E}[F(w_{k_1}) - F_{inf}] \leq \frac{\alpha_1 LM}{2c\mu}$ . Luego se tomaba  $\alpha_2 = \frac{\alpha_1}{2}$  y se seguía inductivamente.

**Teorema 7.3.4 (Objetivo fuertemente convexo, Incremento decreciente)**

Supongamos 7.1.1, 7.2.2 y 3.3.1; además supongamos que dado 7.1  $\alpha_k$  cumple:

$$\alpha_k = \frac{\beta}{\gamma + k} \quad \text{para algún } \beta > \frac{1}{c\mu} \text{ y } \gamma > 0 \text{ tal que } \alpha_1 \leq \frac{\mu}{LM_G} \quad (7.11)$$

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

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

Donde:

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

**Demostración** Notemos primero que por 7.11 para todo  $k \in \mathbb{N}$  vale:

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

Luego por 7.2.3 y 7.3.2 uno tiene para todo  $k \in \mathbb{N}$ :

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

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

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

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

*Porque vale  $k=1$ ?*

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

$\underbrace{\leq}_{(\gamma+k)^2 \geq (\gamma+k+1)(\gamma+k-1)}$

■

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  $L_1$ , cabría preguntarse (inspirados en la observación del caso  $\alpha$  fijo con  $M = 0$ ) si con el ruido existente pero controlado podemos mantener la convergencia en  $L_1$ .

**Teorema 7.3.5 (Objetivo Fuertemente Convexo, Reducción del Ruido)**

Supongamos que valen 7.1.1, 7.2.2 y 3.3.1 pero reforcemos 7.7 a la existencia de una constante  $M \geq 0$  y  $\xi \in (0, 1)$  tal que para todo  $k \in \mathbb{N}$ :

$$\mathbb{V}_{\xi_k} [g(w_k, \xi_k)] \leq M \xi^{k-1} \quad (7.12)$$

Supongamos además que 7.1 tiene  $\alpha_k = \alpha$  para todo  $k \in \mathbb{N}$  satisfaciendo:

$$0 < \alpha \leq \min \left\{ \frac{\mu}{L \mu_G^2}, \frac{1}{\mu} \right\} \quad (7.13)$$

Luego vale:

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

Donde:

$$\omega := \max \left\{ \frac{\alpha LM}{c \mu}, F(w_1) - F_{inf} \right\} \quad (7.14a)$$

$$\rho := \max \left\{ 1 - \frac{\alpha c \mu}{2}, \xi \right\} < 1 \quad (7.14b)$$

**Demostración** Por 7.8a vale que:

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

Luego si juntamos 7.5, 7.6b, 7.13 y 7.12 entonces:

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

Por lo tanto:

$$\mathbb{E} [F(w_{k+1}) - F_{inf}] \leq (1 - c\mu\alpha) \mathbb{E} [F(w_k) - F_{inf}] + \frac{1}{2}\alpha^2 L M \xi^{k-1} \quad (7.15)$$

*Por que vale k=1??*

Probemos ahora la identidad por inducción. Para esto, notemos que el caso  $k = 1$  vale por la definición de  $\omega$ ; luego si asumimos que vale para algún  $k \geq 1$  entonces de 7.15, 7.14a y 7.14b:

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

■

#### 7.4 CASO GENERAL

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

**Teorema 7.4.1 (Objetivo no convexo, Incrementos fijos)** *Asumiendo 7.1.1 y 7.2.2 y suponiendo que en 7.1 tenemos  $\alpha_k = \alpha$  tal que:*

$$0 < \alpha \leq \frac{\mu}{LM_G} \quad (7.16)$$

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

$$\mathbb{E} \left[ \sum_{k=1}^K \|\nabla F(w_k)\|_2^2 \right] \leq \frac{K\alpha LM}{\mu} + \frac{2(F(w_1) - F_{inf})}{\mu\alpha} \quad (7.17a)$$

$$\mathbb{E} \left[ \frac{1}{K} \sum_{k=1}^K \|\nabla F(w)_k\|_2^2 \right] \leq \frac{\alpha LM}{\mu} + \frac{2(F(w_1) - F_{inf})}{K\mu\alpha} \quad (7.17b)$$

$$\xrightarrow{K \rightarrow \infty} \frac{\alpha LM}{\mu}$$

**Demostración** Recordemos 7.8a y si tomamos esperanza total e imponemos 7.16 tenemos:

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

Luego como por 7.2.2 tenemos que  $F_{inf} \leq \mathbb{E}[F(w_k)]$  para todo  $k \in \mathbb{N}$  vale:

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

■

**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  $\left\{ \|\nabla F(w_k)\|_2^2 \right\}_{k \in \mathbb{N}} \xrightarrow{k \rightarrow \infty} 0$ , que es el resultado obtenido en [11].

En cambio, cuando  $M \neq 0$  aunque no podemos acotar  $\|\nabla F(w_k)\|_2^2$  per-se, podemos decir de 7.17b 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 (7.3.3) donde a menor incremento el entorno es menor (el algoritmo 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 3.3:

**Teorema 7.4.2 (Objetivo no convexo, Incrementos decrecientes)** *Asumiendo 7.1.1 y 7.2.2, suponiendo además que en 7.1 los  $\{\alpha_k\}$  satisfacen 3.3; si notamos  $A_K := \sum_{k=1}^K \alpha_k$  vale para todo  $k \in \mathbb{N}$ :*

$$\lim_{K \rightarrow \infty} \mathbb{E} \left[ \sum_{k=1}^K \alpha_k \|\nabla F(w)_k\|_2^2 \right] < \infty \quad (7.18a)$$

$$\mathbb{E} \left[ \frac{1}{A_K} \sum_{k=1}^K \alpha_k \|\nabla F(w)_k\|_2^2 \right] \xrightarrow{K \rightarrow \infty} 0 \quad (7.18b)$$

**Demostración** Como  $\alpha_k \rightarrow 0$  por 3.3 entonces podemos asumir sin pérdida de generalidad que  $\alpha_k LM_G \leq \mu$  para todo  $k \in \mathbb{N}$ , luego:

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

Luego como por 7.2.2 tenemos que  $F_{inf} \leq \mathbb{E}[F(w_k)]$  para todo  $k \in \mathbb{N}$  vale:

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

Luego:

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

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

Por lo que 7.18a esta probado. Finalmente como por 3.3 tenemos que  $A_K \rightarrow \infty$  se tiene 7.18b. ■

**Corolario 7.4.3** Asumiendo 7.1.1 y 7.2.2, suponiendo además que en 7.1 los  $\{\alpha_k\}$  satisfacen 3.3 entonces :

$$\liminf_{k \rightarrow \infty} \mathbb{E}[\|\nabla F(w_k)\|_2^2] = 0 \quad (7.19)$$

**Corolario 7.4.4** Bajo las mismas hipótesis de 7.4.2 sea  $k(K) \in \{1, \dots, K\}$  un índice aleatorio elegido con probabilidades respectivas  $\{\alpha_k\}_{k=1}^K$ ; luego  $\|\nabla F(w_k)\|_2 \rightarrow 0$  en probabilidad.

**Demostración** Sea  $\epsilon > 0$ , luego de 7.18a y la desigualdad de Markov:

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

Porque aca es la  
 esperanza de la  
 esperanza  
 condicional?

**Teorema 7.4.5 (Objetivo no convexo regular, Incrementos decrecientes)**

Bajo las mismas hipótesis de 7.4.2 si además pedimos que  $F \in C^2$  y que  $w \mapsto \|\nabla F(w)\|_2^2$  sea  $l$ -Lipshitz entonces:

$$\lim_{k \rightarrow \infty} \mathbb{E}[\|\nabla F(w_k)\|_2^2] = 0 \quad (7.20)$$

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

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

Si tomamos esperanza condicional a  $\xi_k$  y usamos 7.1.1, 7.2.2 entonces:

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

Luego obtenemos tomando esperanza total:

$$\mathbb{E} [G(w_{k+1})] - \mathbb{E} [G(w_k)] \leq 2\alpha_k L_{\mu_G} \mathbb{E} [\|\nabla F(w_k)\|_2^2] + \frac{1}{2} \alpha_k^2 L_G (M + M_V \mathbb{E} [\|\nabla F(w_k)\|_2^2]) \quad (7.21)$$

Notemos que existe  $K \in \mathbb{N}$  tal que  $\alpha_k^2 \leq \alpha_k$  y luego por 7.4.2 el lado derecho cumple:

$$\lim_{N \rightarrow \infty} 2L_{\mu_G} \underbrace{\sum_{k=K}^{K+N} \mathbb{E} [\alpha_k \|\nabla F(w_k)\|_2^2]}_{7.18a} + \frac{1}{2} L_G \left( M \underbrace{\sum_{k=K}^{K+N} \alpha_k^2}_{3.3} + M_V \underbrace{\sum_{k=K}^{K+N} \mathbb{E} [\alpha_k^2 \|\nabla F(w_k)\|_2^2]}_{7.18a} \right) = 0$$

Sean:

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

Luego como en 7.21 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; concluimos que  $\mathbb{E} [G(w_k)]$  debe ser convergente, y por 7.4.3 tenemos  $\mathbb{E} [\|\nabla F(w_k)\|_2^2] = \mathbb{E} [G(w_k)] \rightarrow 0$ .

■





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  $L_1$  de 7.1 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 3.3. Por simpleza en los cálculos vamos a asumir en este capítulo que  $g$  es un estimador insesgado de  $\nabla F(w_k)$ .

## 8.1 CASO DÉBILMENTE CONVEXO

**Hipótesis 8.1.1 (Acotaciones al segundo momento de  $g$ )** Supongamos que dada  $F$  función objetivo cumple 3.0.1 y  $g$  la estimación insesgada del gradiente en 7.1 vale que existen  $A, B \geq 0$  tales que:

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

**Observación** Notemos que si  $F$  cumple 7.1.1 y  $g$  cumple 7.2.2 entonces automáticamente cumplen 8.1.1.

### Teorema 8.1.2 (Objetivo débilmente convexo, incrementos decrecientes)

Supongamos 3.0.1, 8.1.1; además supongamos que dado 7.1  $\alpha_k$  cumple 3.3:

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

$$w_k \xrightarrow[ctp]{k \rightarrow \infty} w^* \quad (8.1a)$$

$$(w_k - w^*) \nabla F(w_k) \xrightarrow[ctp]{k \rightarrow \infty} 0 \quad (8.1b)$$

**Demostración** Vayamos de a pasos como cuando demostramos 3.2.2.

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

Paso 2 Análogamente a casos anteriores notemos que:

$$h_{k+1} - h_k = -2\alpha_k (w_k - w^*) g(w_k, \xi_k) + \alpha_k^2 g(w_k, \xi_k)^2 \quad (8.2)$$

*Como definimos bien la filtración? Es saber toda esa información a tiempo k, como lo notamos?*

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

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

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

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

Ahora si incorporamos 8.1.1 obtenemos:

$$\mathbb{E}[h_{k+1} - (1 + \alpha_k^2 B) h_k | \mathcal{P}_k] \leq -2\alpha_k (w_k - w^*) \nabla F(w_k) + \alpha_k^2 A \quad (8.3)$$

Definamos ahora las sucesiones auxiliares:

$$\mu_k = \prod_{j=1}^{k-1} \frac{1}{1 + \alpha_j^2 B} \quad (8.4a)$$

$$h'_k = \mu_k h_k \quad (8.4b)$$

Por lo que replicando las operaciones en 3.2.2 llegamos a:

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

Luego, recordando 2.3.10 notemos que:

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

Por el mismo motivo que en 3.2.2 concluimos que:

$$h'_k \geq 0 \quad (8.5a)$$

$$\sum_{k=1}^{\infty} \mathbb{E} \left[ \delta_k^{h'} (h'_{k+1} - h'_k) \right] < \infty \quad (8.5b)$$

Por 2.3.11 concluimos que  $h'_k$  converge ctp; como  $\underbrace{\mu_k}_{\geq 0} \rightarrow \mu_{\infty} > 0$  entonces  $\{h_k\}$  converge ctp.

Paso 3 Como  $h_k$  converge ctp, de 8.3 concluimos que:

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

Supongamos que  $\mathbb{P} \left( \left\{ \lim_k h_k > 0 \right\} \right) > \tilde{\epsilon}$ , luego  $\mathbb{P} (\{ \alpha_k (w_k - w^*) \nabla F(w_k) > C \alpha_k \}) > \tilde{\epsilon}$  lo que implicaría por 3.3 que  $\mathbb{P} \left( \left\{ \sum_{k=1}^{\infty} \alpha_k (w_k - w^*) \nabla F(w_k) = \infty \right\} \right) > \tilde{\epsilon}$ ; concluimos entonces que:

$$w_k \xrightarrow[\text{ctp}]{k \rightarrow \infty} w^* \quad (8.6a)$$

$$(w_k - w^*) \nabla F(w_k) \xrightarrow[\text{ctp}]{k \rightarrow \infty} 0 \quad (8.6b)$$

■

## 8.2 CASO NO CONVEXO

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

**Hipótesis 8.2.1 (Hipótesis caso no convexo)** Sea  $F$  una función de costo objetivo y supongamos que el algoritmo 7.1 cumple que  $g$  es un estimador insesgado, ie:  $\mathbb{E} [g(w_k, \xi_k)] = \nabla F(w_k)$ , 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  $\tilde{F} = F - F_{\inf} \geq 0$ )
4. Sean  $\{\alpha_k\}$  los incrementos del algoritmo 7.1, entonces estos cumplen 3.3
5. Para  $j = 2, 3, 4$  existen  $A_j, B_j \geq 0$  tal que:

$$\mathbb{E} \left[ \|g(w_k, \xi_k)\|_2^j \right] \leq A_j + B_j \|w\|_2^j \quad (8.7)$$

6. Existe  $D > 0$  tal que:

$$\inf_{(w)^2 > D} w \nabla F(w) > 0 \quad (8.8)$$

## 8.2.1 Acotación global del algoritmo

Usemos 8.8 para probar que existe un entorno  $w_1 \in U \subset \chi$  tal que si  $\{w_k\}$  son las iteraciones del algoritmo 7.1 con  $F, g$  cumpliendo 8.2.1, entonces  $\{w_k\} \subset U$ .

**Teorema 8.2.2 (Acotación global del algoritmo estocástico insesgado)**

Sea  $F$  función de costo objetivo y  $g$  un estimador insesgado de  $\nabla F$  tal que ambos cumplen 8.2.1, luego existe un entorno  $w_1 \in U \subset \chi$  tal que si  $\{w_k\}$  son las iteraciones del algoritmo 7.1 entonces  $\{w_k\} \subset U$ .

**Demostración** Nuevamente probemos esto en tres pasos:

Paso 1 Sea  $D$  el parámetro de horizonte dado por 8.8 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:

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

Paso 2 Notemos que por la definición vale:

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

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 Lya-punov  $\{f_k\}$ :

$$\begin{aligned} f_{k+1} - f_k &\leq (-2\alpha_k w_k g(w_k, \xi_k) + \alpha_k^2 g(w_k, \xi_k)^2) \phi'(w_k^2) \\ &\quad + 4\alpha_k^2 (w_k g(w_k, \xi_k))^2 - 4\alpha_k^3 w_k g(w_k, \xi_k)^3 \\ &\quad + 4\alpha_k^4 g(w_k, \xi_k)^4 \end{aligned} \tag{8.9}$$

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:

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

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  $\mathcal{P}_k$ :

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

Si ahora incluimos 8.7 entonces esto implica que existen  $A, B \geq 0$  tal que:

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

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 8.8 por lo que deducimos:

$$\mathbb{E}[f_{k+1} - f_k | \mathcal{P}_k] \leq \alpha_k^2 (A + B f_k) \quad (8.10)$$

Ahora siguiendo los mismo pasos que al demostrar 8.1.2 definiendo  $\mu_k, f'_k$  y usando 2.3.11 concluimos 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 8.9 es una igualdad y deducimos que:

$$\sum_{k=1}^{\infty} \alpha_k w_k \nabla F(w_k) \phi'(w_k^2) < \infty \quad ctp \quad (8.11)$$

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

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

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

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

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

## 8.2.2 Convergencia del algoritmo

**Teorema 8.2.3 (Convergencia a puntos extremales, Caso inesgado)**

Sea  $F$  función de costo objetivo y  $g$  un estimador inesgado de  $\nabla F$  tal que ambos cumplen 8.2.1, si  $\{w_k\}$  son las iteraciones del algoritmo 7.1 entonces valen:

$$F(w_k) \xrightarrow[ctp]{k \rightarrow \infty} F_\infty \quad (8.12a)$$

$$\nabla F(w_k) \xrightarrow[ctp]{k \rightarrow \infty} 0 \quad (8.12b)$$

**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 8.2.2 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:

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

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

$$\mathbb{E} [h_{k+1} - h_k | \mathcal{P}_k] \leq \underbrace{-2\alpha_k (\nabla F(w_k))^2}_{\leq 0} + \alpha_k^2 K_1 \underbrace{\mathbb{E} [g(w_k, \xi_k)^2]}_{\leq K_2 \text{ por 8.2.2}} \quad (8.13)$$

Lo que implica:

$$\mathbb{E} [h_{k+1} - h_k | \mathcal{P}_k] \leq \alpha_k^2 K_1 K_2 \quad (8.14)$$

Luego como tenemos de esto que:

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

Por lo que por 2.3.11 obtenemos:

$$F(w_k) \xrightarrow[ctp]{k \rightarrow \infty} F_\infty \quad (8.15)$$

Paso 3 Si retomamos 8.13, reordenamos, sumamos hasta  $k$  y tomamos esperanza tenemos:

$$\begin{aligned}
 2 \sum_{k=1}^K \alpha_k (\nabla F(w_k))^2 &\leq \sum_{k=1}^K \mathbb{E} [h_{k+1} - h_k] + K_2 K_1 \sum_{k=1}^K \alpha_k^2 \\
 &= \mathbb{E} \left[ \sum_{k=1}^K h_{k+1} - h_k \right] + K_2 K_1 \sum_{k=1}^K \alpha_k^2 \\
 &= \mathbb{E} [h_{K+1}] + K_2 K_1 \sum_{k=1}^K \alpha_k^2 \\
 &\xrightarrow[\text{2.3.5}]{\text{ctp}} F_\infty + K_2 K_1 \sum_{k=1}^{\infty} \alpha_k^2
 \end{aligned} \tag{8.16}$$

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

$$\begin{aligned}
 g_{k+1} - g_k &= \nabla g(w_k) (w_{k+1} - w_k) + \frac{1}{2} \nabla^2 g(w_k) (w_{k+1} - w_k)^2 \\
 &\quad \underbrace{\leq}_{\text{por 8.2.2}} -2\alpha_k \nabla F(w_k) K_4 g(w_k, \xi_k) + K_3 \alpha_k^2 g(w_k, \xi_k)^2
 \end{aligned}$$

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

$$\mathbb{E} [g_{k+1} - g_k | \mathcal{P}_k] \leq 2\alpha_k K_4 (\nabla F(w_k))^2 + K_2 K_3 \alpha_k^2 \tag{8.17}$$

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

$$\begin{aligned}
 \sum_{k=1}^{\infty} \mathbb{E} [\delta_k^g (g_{k+1} - g_k)] &= \sum_{k=1}^{\infty} \mathbb{E} [\delta_k^g \mathbb{E} [(g_{k+1} - g_k) | \mathcal{P}_k]] \\
 &\leq K_4 \underbrace{\sum_{k=1}^{\infty} \alpha_k (\nabla F(w_k))^2}_{8.16} + K_2 K_3 \underbrace{\sum_{k=1}^{\infty} \alpha_k^2}_{3.3} \\
 &< \infty
 \end{aligned} \tag{8.18}$$

Nuevamente por 2.3.11 concluimos que  $\{g_k\}$  converge *ctp*, y por 8.16 este límite es 0; luego:

$$g_k \xrightarrow[\text{ctp}]{k \rightarrow \infty} 0 \tag{8.19a}$$

$$\nabla F(w_k) \xrightarrow[\text{ctp}]{k \rightarrow \infty} 0 \tag{8.19b}$$

■





Part IV

Apéndice



## APÉNDICE

## 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:*

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

Donde:

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

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:

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

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

**Demostración** Ver [14]

**Teorema A.1.2** *Sea  $E \subset \mathbb{R}^d$  y dotemos a  $C^m(E)$  de la norma  $\|f\|_{C^m} = \sup \{\|\partial^\alpha f|_E\|_\infty : |\alpha| \leq m\}$ , si  $E$  es cerrado en  $\mathbb{R}^d$  entonces existe  $T \in L(C^m(E), C^m(\mathbb{R}^d))$  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 3.3.2] TODO

**Demostración** [De 3.3.4] TODO

**Demostración** [De 3.3.7] **TODO**

**Demostración** [De 3.2.1] Definamos:

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

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

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:

$$0 \leq u_k = u_0 + S_k^+ + S_k^- \leq u_0 + S_\infty^+ \quad (\text{A.3a})$$

$$-u_0 - S_\infty^+ \leq S_k^- \leq 0 \quad (\text{A.3b})$$

Luego como  $S_{k+1}^- \leq S_k^-$  concluimos 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.  $\blacksquare$

## BIBLIOGRAFÍA

---

- [1] Cheng Alan. «The Whitney extension theorem in high dimensions.» In: (2015).
- [2] Leon Bottou. «Online Learning and Stochastic Approximations.» In: (May 1999).
- [3] Leon Bottou, Frank E. Curtis, and Jorge Nocedal. «Optimization Methods for Large-Scale Machine Learning.» In: 60 (June 2016).
- [4] Simon S. Du, Chi Jin, Jason D. Lee, Michael I. Jordan, Barnabas Póczos, and Aarti Singh. «Gradient Descent Can Take Exponential Time to Escape Saddle Points.» In: *NIPS*. 2017.
- [5] Donald L. Fisk. «Quasi-Martingales.» In: *Transactions of the American Mathematical Society* 120.3 (1965), pp. 369–389. ISSN: 00029947. URL: <http://www.jstor.org/stable/1994531>.
- [6] G. Piliouras M. Simchowitz M.I. Jordan "J. Lee I. Panageas and B. Recht". «First-order methods almost always avoid saddle points.» In: (2017).
- [7] Donald E. Knuth. «Computer Programming as an Art.» In: *Communications of the ACM* 17.12 (1974), pp. 667–673.
- [8] John M. Lee. «Introduction to Smooth Manifolds.» In: (2000).
- [9] S. Łojasiewicz. *A topological property of real analytic subsets*. Coll. du CNRS, Les équations aux dérivées partielles, 1963, pp. 87–89.
- [10] M. Metivier. *Semi-Martingales*. Walter de Gruyter, 1983.
- [11] Yurii Nesterov. *Introductory Lectures on Convex Optimization*. Vol. 87. Springer Science & Business Media, 2004.
- [12] B. T. Polyak. *Gradient methods for minimizing functionals*. Zh. Vy-chisl. Mat. Fiz, 1963, pp. 643–653.
- [13] K. MURALI RAO. «QUASI-MARTINGALES.» In: *Mathematica Scandinavica* 24.1 (1969), pp. 79–92. ISSN: 00255521, 19031807. URL: <http://www.jstor.org/stable/24489871>.
- [14] 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.
- [15] H. Robbins and S. Monro. *A Stochastic Approximation Model*. Vol. 22(3). The Annal of Mathematical Statistics, 1951, pp. 400–407.
- [16] Michael Schub. *Global Stability of Dynamical Systems*. Springer Science and Business media, 1987.

- [17] David Williams. *Probability with Martingales*. Cambridge University Press, 1991. DOI: [10.1017/CB09780511813658](https://doi.org/10.1017/CB09780511813658).