



MÉTODOS DE PRIMER ORDEN
ANÁLISIS DE CONVERGENCIA

AXEL SIROTA

Universidad de Buenos Aires

Facultad de Ciencias Exactas y Naturales
Departamento de Matemática

Tesis de Licenciatura
Director de Tesis: Dr. Pablo Amster
Octubre 2018 – version 2.0

Ohana significa familia.
Y tu familia nunca te abandona, ni te olvida.
— Lilo & Stitch

A las 2 chicas que marcan mi Norte:
Patri y Sophia

RESUMEN

En esta tesis intentamos responder a la simple pregunta:

¿Bajo qué condiciones convergen los métodos de primer orden más utilizados en Machine Learning?

Para ello utilizamos herramientas de sistemas dinámicos y procesos estocásticos en pos de analizar la convergencia tanto en algoritmos determinísticos como estocásticos.

Notamos que aunque los algoritmos determinísticos gozan de una velocidad excepcional en el caso convexo, esto no se generaliza al caso no convexo donde la convergencia puede ser hasta de orden exponencial. Por otro lado, en el caso estocástico la misma naturaleza de éste garantiza una complejidad uniforme en ambos regímenes, aún en el rango del *big data*. Sumado a esto dimos motivos tanto teóricos como prácticos para la preferencia de los algoritmos estocásticos, relacionados a la velocidad de convergencia a entornos de la solución en el caso general.

AGRADECIMIENTOS

A la UBA, pues sin su educación gratuita, pública, laica y de calidad este título no sería posible

A mi director Pablo, porque a pesar de estar tan atareado y haber sido recientemente padre se animó a meterse en un tema totalmente nuevo, porque el es así

A los jurados Pablo Ferrari y Pablo De Napoli, porque se tomaron el trabajo y esfuerzo de analizar y evaluar esta Tesis

A Isa y Xime, porque siempre me apoyaron aun cuando me equivoqué de dirección de Iglesia

A los amigos de siempre y a los que se sumaron con los años: Martín, Ezequiel M, Ezequiel G y Pablo L; por estar cuando se necesita y las juntadas y visitas que motivan

A Medallia por la calidad humana inigualable, porque a pesar que recién ingrese supieron entenderme y darme el tiempo que necesité para recibirme

Sobre todo a mi equipo Uwe, Vic, Staffan y David, que nunca me dijeron nada y siempre se bancaron el trabajo extra para que logre esto

A Agustín y Matu, porque vieron algo en mí que yo no conocía que tenía, y sin ustedes no habría descubierto el área que ahora amo y me dedico

Finalmente y más importante : A la familia del corazon, porque lo importante no es la que a uno le toca, sino la que a uno se elige.

**A Miriam, porque siempre nos acompaño pase lo que pase,
A Mati y Laura, porque nos dieron un espacio y siempre nos divertimos,**

A Juan, por ser el modelo de padre que quiero seguir y siempre va a estar con nosotros,

A Sophi, porque indirectamente me inspiro a apurarme con la tesis y recibirme,

Y finalmente a mi compañera de vida, a Patri, porque te conocí por casualidad; pero no es casualidad que nunca te deje ir:

Le dejo a ella todas las virtudes de mis aciertos y cargo conmigo el peso de mis fallos.

CONTENTS

I	Introducción	1
1	INTRODUCCIÓN	3
2	PRELIMINARES	7
2.1	Preliminares de Machine Learning	7
2.2	Preliminares Matemáticos	15
3	RESUMEN DE RESULTADOS	23
3.1	Algoritmos de tipo Batch	23
3.2	Algoritmos estocásticos	24
II	Algoritmos de tipo batch	27
4	CONVERGENCIA PUNTUAL	29
4.1	Intuición	30
4.2	Caso discreto	31
4.3	Acerca de convexidad fuerte y funciones L-Lipschitz	33
5	TEOREMA DE LA VARIEDAD ESTABLE Y LOS PUNTOS FIJOS INESTABLES	37
5.1	Intuición	37
5.2	Puntos fijos inestables	40
6	CONVERGENCIA CTP A MÍNIMOS : CASO GENERAL	43
6.1	Descenso de Gradiente en Batch	43
6.2	Punto Próximo	44
6.3	Descenso por coordenadas	45
7	RESULTADOS NEGATIVOS	49
7.1	Ejemplos <i>patológicos</i>	49
7.2	Intuición acerca del teorema	50
7.3	Demostración del teorema	52
III	Algoritmos estocásticos	59
8	CONVERGENCIA EN L_1	61
8.1	Contexto	61
8.2	Algunos lemas fundamentales	62
8.3	Caso Fuertemente Convexo	64
8.4	Caso general	68

9	CONVERGENCIA CTP	73
9.1	Caso débilmente convexo	73
9.2	Caso no convexo	75

IV Epílogo 81

10	CONCLUSIONES	83
A	APÉNDICE	85
A.1	Proposiciones enunciadas	85
A.2	Demostraciones	85
	BIBLIOGRAFÍA	92
B	BIBLIOGRAFÍA	93

Parte I

Introducción

INTRODUCCIÓN

Las Matemáticas no conocen de razas ni fronteras geográficas; para las Matemáticas el mundo cultural es un solo país.

David Hilbert

Machine Learning es un conglomerado de técnicas estadísticas, probabilísticas y algorítmicas para resolver el problema universal siguiente:

Dado un conjunto de datos ya ocurridos, construir un algoritmo que aprenda los patrones intrínsecos para poder predecir datos futuros bajo la misma distribución.

Esto está contextualizado en la formalización matemática del problema y la identificación de una familia de funciones tal que aproximen a este, lo que conlleva un problema de optimización convexa o no convexa según el caso. A este proceso de modelizar el problema mediante una familia de funciones y obtener el óptimo mediante un algoritmo de optimización es lo que comunmente se llama *aprender* la distribución de los datos. En esta Tesis nos centramos en el análisis de convergencia de algoritmos de optimización que suelen aparecer en machine learning, los cuales se dividen en *batch* o *estocásticos* dependiendo la naturaleza de éste.

Dentro del espectro de los algoritmos de batch se encuentra el algoritmo de descenso más pronunciado -también conocido como gradiente, descenso de gradiente por batch o método de gradiente completo- (GD) que se define en el algoritmo 1.1 y sus variantes que veremos en el capítulo 6.

Vamos a representar a los algoritmos mediante un recuadro titulado. Tendrán un input definiendo el conocimiento *a priori* del problema y el algoritmo escrito en *pseudo-código*; es decir, una sucesión lógica de pasos que representan las fórmulas a utilizar y texto explicativo del paso a ejecutar.

Algorithmus 1.1 : Descenso de gradiente en batch (GD)

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

En este algoritmo F es la función objetivo a optimizar, $\{\alpha_k\}$ resultan los incrementos positivos, w_1 la condición inicial y X el conjunto de datos conocidos.

Por otro lado, en el espectro *estocástico* encontramos el algoritmo *descenso estocástico de gradiente generalizado* (DE) definido en 1.2:

Algorithmus 1.2 : Descenso Estocastico de Gradiente (DE)	
1	Input: $w_1 \in \mathbb{R}^d$, $(x_k, y_k) = \xi_j$ dato muestreado al azar
2	for $k \in \mathbb{N}$ do
3	$w_{k+1} \leftarrow w_k - \alpha_k g(w_k, (x_k, y_k))$

En este algoritmo $g : \mathbb{R}^d \times (\mathbb{R}^{d_x} \times \mathbb{R}^{d_y}) \rightarrow \mathbb{R}^d$ es un estimador del vector $\nabla F := \left(\frac{\partial F}{\partial x_1}, \dots, \frac{\partial F}{\partial x_n} \right)$, $\{\alpha_k\}$ resultan los incrementos positivos, w_1 la condición inicial y (x_k, y_k) un dato nuevo a analizar. Una variante muy utilizada en machine learning resulta de tomar $g(w_k, (x_k, y_k)) := \nabla F(w_k, x_k, y_k)$, caso conocido como *descenso estocástico de gradiente* (SG) propuesto por Robbins y Monro [22]. Cada iteración de este método es computacionalmente menos costosa que del algoritmo GD, ya que solo incluye el cálculo del gradiente $\nabla f_{i_k}(w_k)$ correspondiente a un dato en vez de tomar toda la muestra $X = \{(x_j, y_j)\}_{j \leq N}$. El método es notable porque la secuencia de iteración no está determinada únicamente por la función F , el punto de inicio w_1 , y la secuencia de incrementos $\{\alpha_k\}$, como lo haría en un algoritmo de optimización determinista. Por el contrario, $\{w_k\}$ es un proceso estocástico cuyo comportamiento está determinado por la secuencia aleatoria $\{(x_j, y_j)\}$.

En pos de analizar su convergencia analizamos los algoritmos determinísticos vía el sistema dinámico que inducen, mientras que a los algoritmos estocásticos vía el proceso estocástico que inducen.

Históricamente se intentaron realizar algoritmos que tengan heurísticas más y más complejas para poder memorizar todos los casos excepcionales que ocurren (un ejemplo de esto son los análisis de fraude donde el algoritmo intenta matchear las operaciones ocurridas contra una lista de reglas preseteadas que se determinaron como *sospechosas*), pero se observó tanto empírica como experimentalmente que los mejores resultados se obtuvieron mediante técnicas de machine learning; obteniendo mejor robustez y estabilidad de las soluciones. Estos resultados motivan la investigación tanto en convergencia como en desarrollo de mejores algoritmos para obtener soluciones óptimas.

Esta Tesis esta organizada según la categorización de algoritmos mencionada anteriormente.

La primera parte (I) refiere a la motivación tanto matemática, algorítmica y del área para analizar la convergencia de los algoritmos presentados, como a su vez los contenidos preliminares usados a lo largo del documento. En el último capítulo de esta parte (capítulo 3) se incluye un resumen de los resultados, pensando mayoritariamente

en el practicante del Machine Learning que quiere verificar rápidamente condiciones de convergencia para sus algoritmos.

La segunda parte (II) trata exclusivamente los algoritmos de *tipo batch*. En el Capítulo 4, utilizando la histórica referencia [15], analizamos la convergencia puntual del descenso de gradiente con condiciones de convexidad débil y luego la convergencia *lineal* con convexidad más fuerte. Luego en el capítulo 5 nos basamos en [7] para ver estos algoritmos como discretizaciones de sistemas dinámicos y gracias al 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 6 con variantes usualmente usadas en el área. Finalmente en el capítulo 7 nos basamos en [4] para ver que aunque se tiene convergencia *casi todo punto*, la complejidad algorítmica del descenso de gradiente es exponencial.

Estos resultados negativos en el caso determinístico nos motivan a estudiar los algoritmos estocásticos en la parte III. En el capítulo 8 nos basamos en [3] para analizar la convergencia en *norma L1* al mínimo (o a un entorno de éste). Mientras que en el capítulo 9, utilizando [2], estudiamos la convergencia *casi todo punto* tanto en los casos convexo como no convexo; aprovechando que el algoritmo induce un proceso estocástico que a su vez induce una *cuasi-martingala* convergente.

En general, en lugar de contrastar los DE y otros métodos basados en los resultados de experimentos numéricos -que pueden sesgar nuestra revisión hacia un conjunto de pruebas limitado y detalles de implementación- enfocamos nuestra atención en las compensaciones computacionales fundamentales y las propiedades teóricas de los métodos de optimización. Esto hace una clara yuxtaposición con la forma tradicional de investigación de los algoritmos en el área (sobre todo en los trabajos de hace algunas décadas) donde se proponen algoritmos y se evalúa su competencia mediante experimentos, para luego extrapolar los resultados al caso general sin validación extra. Una referencia interesante de esto es el algoritmo *Adam* que recientemente (ver [30]) se obtuvo resultados teóricos refutando su utilidad real.

PRELIMINARES

Las Matemáticas son la creación
mas bella y poderosa del espíritu
humano.

Stefan Banach

2.1 PRELIMINARES DE MACHINE LEARNING

En esta sección vamos a presentar en forma resumida el procedimiento por el cual se pasa desde un conjunto de datos a aprender a un problema de optimización, generalmente no convexo. De ninguna manera esta exposición pretender ser exhaustiva y se basa principalmente en los excelentes trabajos de [3] y [13].

2.1.1 *Procedimiento formal de machine learning.*2.1.1.1 *Fundamentos*

Nuestro objetivo es elegir una función de predicción que evite la memorización mecánica y, en su lugar, generalice los conceptos que se pueden aprender a partir de un conjunto dado de ejemplos. Formalmente, necesitamos encontrar una $h : \mathcal{X} \rightarrow \mathcal{Y}$ tal que, dado $x \in \mathcal{X}$, el valor $h(x)$ ofrece una predicción precisa sobre el valor verdadero de salida y . Para hacer esto, uno debe elegir la función de predicción h intentando minimizar el riesgo medido sobre una familia de funciones de predicción adecuadamente seleccionadas [24], llamadas \mathcal{H} .

Para formalizar esta idea, supongamos que los ejemplos se muestrean a partir de una función de distribución de probabilidad conjunta $P(x, y)$. En lugar de algo que simplemente minimiza el riesgo empírico (2.1), uno debe buscar encontrar h que arroje un pequeño *riesgo esperado* de clasificaciones erróneas sobre *todas las entradas posibles*, es decir, una h que minimice 2.2:

$$R_n(h) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}[h(x_i) \neq y] \quad (2.1)$$

$$R(h) = \mathbb{P}[h(x) \neq y] = \mathbb{E}[\mathbb{1}[h(x) \neq y]] \quad (2.2)$$

donde $\mathbb{P}[A]$ denota la probabilidad del evento A y $\mathbb{E}[A]$ denota la esperanza de la variable aleatoria indicadora del evento A . Tal con-

texto es *variacional* ya que estamos optimizando sobre un conjunto de funciones, y es *estocástico* ya que la función objetivo implica una esperanza.

Como en la práctica no se conoce *P a priori*, la única opción posible resulta ser transformar el problema a uno que dependa únicamente de los ejemplos $\{(x_i, y_i)\}_{i=1}^n$ utilizando la *ley de los grandes números* y teniendo en cuenta dos cuestiones principales que deben abordarse:

- Cómo elegir la familia parametrizada de funciones de predicción \mathcal{H}
- Cómo determinar (y encontrar) la función de predicción particular $h \in \mathcal{H}$ que es óptima.

2.1.1.2 Elección de una familia de funciones de predicción

La familia de funciones \mathcal{H} debe determinarse teniendo en cuenta tres *objetivos potencialmente competitivos*. En primer lugar, \mathcal{H} debe contener funciones de predicción tal que en el conjunto de entrenamiento puedan lograr un riesgo empírico numéricamente bajo, a fin de evitar el sesgo o el ajuste inadecuado de los datos. Esto se puede lograr seleccionando una familia rica de funciones o utilizando un conocimiento *a priori* para seleccionar una familia bien dirigida. En segundo lugar, la brecha entre el riesgo esperado y el riesgo empírico, es decir, $R(h) - R_n(h)$, debe ser pequeña para todo $h \in \mathcal{H}$. Por lo general, esta brecha disminuye cuando se utilizan más ejemplos de entrenamiento pero, debido al potencial sobreajuste, aumenta cuando usa familias de funciones más ricas. En tercer lugar, \mathcal{H} debe seleccionarse de manera que se pueda resolver eficientemente el problema de optimización correspondiente, cuya dificultad puede aumentar cuando se emplea una familia más rica de funciones y/o un conjunto de entrenamiento más amplio.

Nuestra observación sobre la brecha entre el riesgo esperado y el empírico puede entenderse utilizando la *ley de los grandes números*. La desigualdad Hoeffding [6] garantiza que dado $\eta > 0$ existe $N \in \mathbb{N}$ tal que con probabilidad de al menos $1 - \eta$, uno tiene

$$|R(h) - R_n(h)| \leq \sqrt{\frac{1}{2n} \log \left(\frac{2}{\eta} \right)} \quad \text{para un dado } h \in \mathcal{H} \quad \forall n \geq N \quad (2.3)$$

Esta desigualdad implica la *convergencia en probabilidad* del riesgo empírico y ofrece la explicación intuitiva de que la brecha disminuye a medida que se usan más ejemplos de entrenamiento. Sin embargo, esta explicación es insuficiente para nuestros propósitos ya que, en el contexto del machine learning, h no es una función fija, sino que es la variable sobre la cual se está optimizando!

Por esta razón, a menudo se recurre a la *ley uniforme de los grandes números* y al concepto de la dimensión Vapnik-Chervonenkis (VC) de \mathcal{H} , una medida de la *capacidad* de dicha familia de funciones [24], [13].

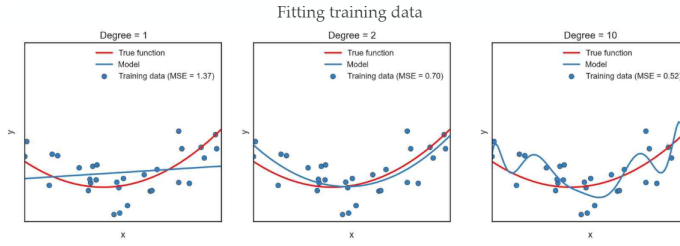


Figure 1: Fenómeno de sobreajuste en polinomios de alto grado

Un ejemplo de éste fenómeno se puede ver en la brecha entre el riesgo empírico y el riesgo esperado para un conjunto de polinomios de alto grado, donde al aumentar el grado puede ser mayor ya que su alta capacidad les permite sobreajustar un conjunto dado de datos de entrenamiento. (Ver Figura 1)

2.1.1.3 Minimización de riesgos estructurales

Un enfoque para elegir una función de predicción que ha demostrado ser ampliamente exitoso en la práctica es la *minimización del riesgo estructural* [25], [26]. En lugar de elegir una familia genérica de funciones de predicción, sobre las cuales sería difícil optimizar y estimar la brecha entre los riesgos empíricos y los esperados, se elige una *estructura*, es decir, una colección de familias de funciones anidadas. Por ejemplo, dicha estructura se puede formar como una colección de subconjuntos de una determinada familia \mathcal{H} de la siguiente manera: dada una función de preferencia Ω , se elijen varios valores de un *hiperparámetro* C , de acuerdo con cada uno de los cuales se obtiene el subconjunto $\mathcal{H}_C := \{h \in \mathcal{H} : \Omega(h) \leq C\}$. Dado un número fijo de ejemplos, el aumento de C reduce el riesgo empírico (es decir, el mínimo de $R_n(h)$ sobre $h \in \mathcal{H}_C$), pero, después de cierto punto, típicamente aumenta la brecha entre los riesgos esperado y empírico. Este fenómeno se ilustra en la Figura 2.

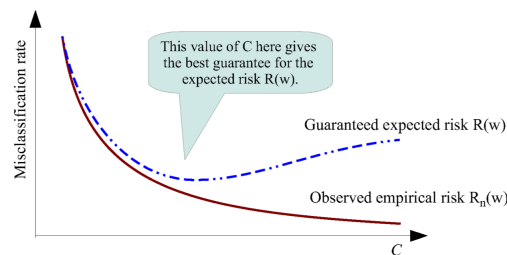


Figure 2: Fenómeno de distancia de los riesgos en función de la evolución de hiperparámetros

Luego, en vez de estimar la brecha entre el riesgo empírico y esperado teóricamente, se observa que una estimación aceptable de esa brecha para el valor óptimo de C se puede obtener dividiendo los datos disponibles en tres subconjuntos.

1. **Conjunto de entrenamiento:** Utilizado para elegir un subconjunto finito $\tilde{\mathcal{H}}_C \subset \mathcal{H}_C$ de valores que minimicen 2.1
2. **Conjunto de validación:** Utilizado para estimar 2.2 sobre todo $\tilde{h} \in \tilde{\mathcal{H}}_C$ y elegir una h_C óptima
3. **Conjunto de prueba:** Utilizado para estimar la brecha entre 2.1 y 2.2 para la h_C elegida

Suponiendo que se ha utilizado un rango suficientemente grande para C , a menudo se encuentra que la mejor solución no corresponde al mayor valor de C considerado; nuevamente, vea la Figura 2.

En general, el principio de minimización del riesgo estructural ha demostrado ser útil para muchas aplicaciones. En lugar de codificar el conocimiento como reglas formales de clasificación, uno lo codifica mediante preferencias para ciertas funciones de predicción sobre otras, luego explora el rendimiento de varias funciones de predicción que se han optimizado bajo la influencia de dichas preferencias.

2.1.2 Enunciados de problemas de optimización formal

Para continuar, debemos definir las funciones de predicción y pérdida que aparecen en las mediciones de riesgo esperado y empírico que se pretenden minimizar.

2.1.2.1 Funciones de predicción y pérdida

Durante esta Tesis vamos a considerar, como se mencionó anteriormente, que el conocimiento *a priori* del problema otorga una función h que parametrizaremos por un vector real $w \in \mathbb{R}^d$ sobre el cual se realizará la optimización; en vez de encarar el problema variacional. Formalmente, para un $h(\cdot; \cdot) : \mathbb{R}^{d_x} \times \mathbb{R}^d \rightarrow \mathbb{R}^{d_y}$ dado consideramos la familia de funciones de predicción

$$\mathcal{H} := \{h(\cdot; w) : w \in \mathbb{R}^d\} \quad (2.4)$$

Nuestro objetivo es encontrar la función de predicción en esta familia que minimice una función de pérdida dada $\ell : \mathbb{R}^{d_y} \times \mathbb{R}^{d_y} \rightarrow \mathbb{R}$. Esta mide, para un par de entrada-salida (x, y) , la pérdida $\ell(h(x; w), y)$ cuando $h(x; w)$ e y son las salidas predicha y verdadera respectivamente.

2.1.2.2 Riesgo esperado

Idealmente, el vector de parámetros ω se elige para minimizar la pérdida esperada en la que se incurriría con *cualquier* par de entrada-salida. Para expresar esta idea formalmente, suponemos que las pérdidas se miden con respecto a una distribución de probabilidad $P(x, y)$ que representa la verdadera relación entre las entradas y las salidas. Luego, la función objetivo que deseamos minimizar es

$$R(w) = \int_{\mathbb{R}^{d_x} \times \mathbb{R}^{d_y}} \ell(h(x; w), y) dP(x, y) = \mathbb{E}[\ell(h(x; w), y)] \quad (2.5)$$

Decimos que $R : \mathbb{R}^d \rightarrow \mathbb{R}$ produce el *riesgo esperado* (es decir, la pérdida esperada) dado un vector de parámetro w con una respectiva distribución de probabilidad P .

2.1.2.3 Riesgo empírico

Si bien puede ser deseable minimizar (2.5), tal objetivo es insostenible cuando no se cuenta con información completa sobre P . Por lo tanto, en la práctica, uno busca la solución de un problema que involucre una estimación del riesgo esperado R_n . En el aprendizaje supervisado, uno tiene acceso (ya sea de una vez o de manera incremental) a un conjunto de $n \in \mathbb{N}$ muestras de entrada y salida independientes $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$, con las cuales se puede definir la función de riesgo empírico $R_n : \mathbb{R}^d \rightarrow \mathbb{R}$ dada por la ecuación

$$R_n(\omega) = \frac{1}{n} \sum_{i=1}^n \ell(h(x_i; w), y_i) \quad (2.6)$$

En términos generales, la minimización de R_n puede considerarse el problema práctico de optimización de interés.

2.1.2.4 Notación simplificada

Para simplificar, representemos una muestra (o conjunto de muestras) por una variable aleatoria ξ ; por ejemplo, uno puede imaginar una realización de ξ como una muestra única (x, y) de $\mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$, o una realización de ξ podría ser un conjunto de muestras $\{(x_i, y_i)\}_{i \in S}$. Además, podemos referirnos a la pérdida incurrida para un dado (w, ξ) como $F(w; \xi)$, es decir, F es la composición de la función de pérdida y la función de predicción h .

Para referencia futura, usamos $\xi_{[i]}$ para denotar el i -ésimo elemento de un conjunto fijo de realizaciones de una variable aleatoria ξ , mientras que, comenzando en la parte III, utilizaremos $\xi_{[k]}$ para denotar el k -ésimo elemento de una secuencia de variables aleatorias.

2.1.3 Motivación para los métodos estocásticos

Antes de analizar las fortalezas de los métodos estocásticos, como el algoritmo DE, no se debe perder de vista el hecho de que los enfoques por batch poseen algunas ventajas intrínsecas. Primero, cuando uno ha reducido el problema estocástico de minimizar el riesgo esperado R para enfocarse exclusivamente en el problema determinista de minimizar el riesgo empírico R_n , el uso de información de gradiente completo en cada iteración abre la puerta para muchos métodos de optimización determinísticos. Es decir, en un enfoque por batch, uno tiene a su disposición la gran cantidad de técnicas de optimización no lineal que se han desarrollado en las últimas décadas, incluido el método de gradiente completo o gradiente de batch (1.1), pero también gradiente acelerado, gradiente conjugado, cuasi-Newton y métodos inexactos de Newton [17]. Segundo, debido a la estructura de suma de R_n , un método por batch puede beneficiarse fácilmente de la paralelización ya que la mayor parte del cálculo se basa en evaluaciones de R_n y ∇F . Los cálculos de estas cantidades pueden incluso realizarse de forma distribuida.

A pesar de estas ventajas, existen razones intuitivas, prácticas y teóricas para seguir un enfoque estocástico. En pos de eso contrastaremos la iteración DE característica (1.2) con la iteración de gradiente de batch completo (1.1).

2.1.3.1 Motivación intuitiva

En un nivel intuitivo, DE emplea información de manera más eficiente que un método por batch. Para ver esto, en una situación en la cual un conjunto de entrenamiento, notémoslo S , consta de diez copias de un conjunto de S_{sub} . Un minimizador de riesgo empírico para el conjunto mayor S está claramente dado por un minimizador para el conjunto más pequeño S_{sub} , pero si se aplicara un enfoque por batch para minimizar R_n sobre S , entonces cada iteración sería diez veces más costosa que si solo tuviese una copia de S_{sub} . En realidad, un conjunto de entrenamiento típicamente no consiste en duplicados exactos de datos de muestra, pero en muchas aplicaciones a gran escala los datos involucran una buena cantidad de redundancia (aproximada). Esto sugiere que usar todos los datos de muestra en cada iteración de optimización es ineficiente.

2.1.3.2 Motivación práctica

Los beneficios intuitivos que acabamos de describir se han observado repetidamente en la práctica, donde a menudo se encuentran ventajas muy reales de SG en muchas aplicaciones. Como ejemplo, la Figura

3 compara el rendimiento de un método **L-BFGS** por batch¹ [10] [16] y el método SG con un incremento constante (es decir, $\alpha_k = \alpha$ para todos $k \in \mathbb{N}$) en un problema de clasificación binario que utiliza una función objetivo de pérdida logística y los datos del conjunto de datos RCV1². La figura traza el riesgo empírico R_n en función del número de accesos de una muestra del conjunto de entrenamiento, es decir, el número de evaluaciones de un gradiente de muestra $\nabla f_{i_k}(w_k)$. Cada conjunto de n accesos consecutivos se llama *epoch*. El método por batch solo realiza un paso por *epoch*, mientras que SG realiza n pasos por *epoch*. La trama muestra el comportamiento en los primeros 10 *epochs*. La ventaja de SG es llamativa y representativa del comportamiento típico en la práctica.

Sin embargo, se debe tener en cuenta que para obtener un comportamiento tan eficiente, era necesario ejecutar SG varias veces usando diferentes opciones para α hasta que se identificara una buena opción para este problema en particular. Discutimos cuestiones teóricas y prácticas relacionadas con la elección de incrementos en nuestro análisis en la parte III.

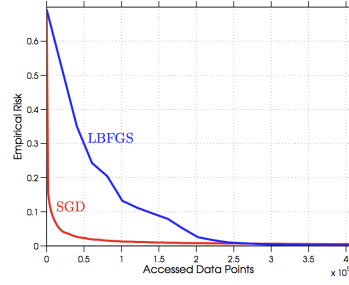


Figure 3: Riesgo empírico R_n en función del número de puntos de datos accedidos (ADP) para un método **L-BFGS** por batch y el método de gradiente estocástico (SG) en un problema de clasificación binaria con un objetivo de pérdida logística y el conjunto de datos RCV1. SG se ejecutó con un incremento fijo de $\alpha = 4$

2.1.3.3 Motivación Teórica

También se pueden citar argumentos teóricos para una preferencia de DE sobre un enfoque por batch. Ahora vamos a dar un pequeño resumen de estos argumentos, que se estudian con más profundidad y más detalle en la parte III:

- ¹ El algoritmo L-BFGS no resulta importante *per-se* en este trabajo, se lo menciona pues es un algoritmo de tipo batch de primer orden que resultó muy famoso en los '80 para problemas de clasificación
- ² Esta colección de datos contiene características de los documentos escritos originalmente en cinco idiomas diferentes y sus traducciones, en un conjunto común de 6 categorías, lo que lo convierte en un conjunto de datos muy útil a la hora de comparar algoritmos de clasificación de texto

- Es bien sabido que un enfoque por batch puede minimizar R_n a un ritmo rápido. Por ejemplo, si R_n es fuertemente convexo (ver 4.3.1) y uno aplica un método de descenso de gradiente por batch, entonces existe una constante $\rho \in (0, 1)$ tal que, para todo $k \in \mathbb{N}$, el error de entrenamiento satisface:

$$R_n(w_k) - R_n^* \leq \mathcal{O}(\rho^k) \quad (2.7)$$

Donde R_n^* denota el valor mínimo de R_n . La velocidad de convergencia exhibida aquí se conoce como *convergencia lineal* en la bibliografía de optimización [18] y *convergencia geométrica* en la comunidad de investigación de machine learning; simplemente nos referiremos a él como convergencia lineal. De (2.7), se puede concluir que, en el peor de los casos, el número total de iteraciones en las que el error de entrenamiento puede estar por encima de un valor dado de $\epsilon > 0$ es proporcional a $\log(\frac{1}{\epsilon})$. Esto significa que, con un costo por iteración proporcional a n (debido a la necesidad de calcular $\nabla F(w_k)$ para todo $k \in \mathbb{N}$), el trabajo total requerido para obtener ϵ -optimalidad para un método de gradiente por batch es proporcional a $n \log(\frac{1}{\epsilon})$.

- La velocidad de convergencia de un método estocástico básico es más lenta que para un método de gradiente por batch; por ejemplo, si R_n es estrictamente convexo y cada i_k se muestrea uniformemente desde en $\{1, \dots, n\}$, entonces, para todo $k \in \mathbb{N}$, las iteraciones DE satisfacen la propiedad de convergencia sublineal (ver 8.3.3):

$$\mathbb{E}[R_n(w_k) - R_n^*] = \mathcal{O}\left(\frac{1}{k}\right) \quad (2.8)$$

Sin embargo, es crítico notar que ni el costo de iteración ni el orden dependen del tamaño del conjunto de muestras n . Esto significa que el trabajo total requerido para obtener ϵ -optimalidad para DE es proporcional a $\frac{1}{\epsilon}$. Es cierto que esto puede ser mayor que $n \log(\frac{1}{\epsilon})$ para valores chicos de n y ϵ , pero la comparación favorece a DE cuando se pasa al régimen de *big data* donde n es grande y uno se encuentra limitado únicamente por un presupuesto de tiempo computacional.

- Otra característica importante de DE es que produce la misma velocidad de convergencia que en (2.8) para el error en el riesgo esperado, $R - R^*$, donde R^* es el valor mínimo de R . Específicamente, aplicando la iteración DE, pero con $g(w_k, \xi_k)$ reemplazado por $\nabla f(w_k; \xi_k)$ con cada ξ_k tomado independientemente de acuerdo con la distribución P , vale:

$$\mathbb{E} [R(w_k) - R^*] = \mathcal{O} \left(\frac{1}{k} \right) \quad (2.9)$$

Que nuevamente es una velocidad sublineal pero en el riesgo esperado, algo que es imposible estimar en los métodos de batch. Por lo tanto, deducimos que en el regimen de *big data* minimizar el riesgo empírico o el riesgo esperado es equivalente, lo que potencia la generalidad de las soluciones halladas por DE.

En resumen, existen argumentos intuitivos, prácticos y teóricos a favor de enfoques estocásticos sobre métodos por batch en optimización para el machine learning a gran escala. Sin embargo, no pretendemos que los métodos por batch no tengan lugar en la práctica; por ejemplo si en la Figura 3 se considerara un mayor número de *epochs*, entonces se vería que el algoritmo de batch eventualmente mejora al método estocástico y produce un menor error de entrenamiento. Esto motiva que muchos métodos propuestos recientemente intenten combinar las mejores propiedades de los algoritmos por batch y estocásticos. Además, la iteración de DE es difícil de paralelizar y requiere una comunicación excesiva entre nodos en una configuración de computación distribuida, proporcionando un mayor impulso para el diseño de algoritmos de optimización nuevos y mejorados. [14] [28]

2.2 PRELIMINARES MATEMÁTICOS

2.2.1 Espectro

Definición 2.2.1 Sea $f : X \rightarrow Y$, con X, Y espacio de Banach, $f \in L(X, Y)$; definimos el espectro de f de la siguiente manera:

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

Si este conjunto fuese vacío, entonces $R_\lambda = (\lambda - f)^{-1}$ sería holomorfa en \mathbb{C} y no constante, absurdo por Liouville. Luego, como es no vacío existe el supremo pues \mathbb{C} es completo; al supremo del espectro le decimos radio espectral:

$$\rho(f) = \sup \{|\alpha| : \alpha \in \sigma(f)\}$$

Proposición 2.2.2 Sea $f \in L(X, Y)$ un operador lineal, entonces:

$$\rho(f) = \lim_{n \rightarrow \infty} \|f^n\|^{\frac{1}{n}}$$

Teorema 2.2.3 (Teorema espectral para operadores compactos autoadjuntos)

Sea X espacio de Hilbert y $T \in L(X)$ compacto y autoadjunto, entonces T admite numerables autovalores distintos.

Es más si $\{\lambda_1, \dots, \lambda_n, \dots\} \subset \sigma_p(T) \cap \mathbb{R}^*$ y $P_n = P_{\ker(T - \lambda_n)}$ (donde notamos $P_{\ker(T - \lambda_n)}$ a la proyección ortogonal con respecto al autoespacio de T asociado al autovalor λ_n) luego $P_n P_m = P_m P_n = 0$ si $n \neq m$ y vale:

$$T = \sum_{n \in \mathbb{N}} \lambda_n P_n \quad (2.10)$$

Proposición 2.2.4 Sea $f \in L(X, Y)$ un operador compacto autoadjunto y h una función holomorfa en un entorno $U \subset \mathbb{C}$ con $\sigma(f) \subset U$, entonces:

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

Donde definimos $h(f) = \sum_{n \in \mathbb{N}} h(\lambda_n) P_n \in L(X, Y)$ pues f es compacto.

Corolario 2.2.5 Sea $A \in \mathbb{C}^{n \times n}$ una matriz diagonalizable con autovalores $\{\lambda_1, \dots, \lambda_n\}$ y $f \in \mathbb{C}_n[X]$ un polinomio con coeficiente complejos de grado n , entonces $f(A)$ tiene autovalores $\{f(\lambda_1), \dots, f(\lambda_n)\}$.

2.2.2 Variedades diferenciables y Teorema de la variedad estable

Cuando estudiemos los algoritmos de tipo batch es habitual analizar el 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.

2.2.2.1 Un repaso de Variedades

Definición 2.2.6 [Capítulo 1 de [9]] Dado un espacio topológico X decimos que es una variedad diferenciable de dimensión d si:

- X es Hausdorff
- Para todo cubrimiento por abiertos de X existe un subcubrimiento numerable
- 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, ϕ) 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 2.2.7 [Capítulo 6 de [9]] Dada una variedad de dimensión d que notaremos como X y el espacio de medida $(\mathbb{R}^d, \mathcal{B}, \mu)$, decimos que $E \subset X$ tiene medida cero si existe un atlas $\mathcal{A} = \{U_i, \phi^i\}_{i \in \mathbb{N}}$ tal que $\mu(\phi^i(E \cap U_i)) = 0$ para todo i . En este caso usamos el abuso de notación $\mu(E) = 0$.

Definición 2.2.8 [Capítulo 3 de [9]] 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 γ 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 y resulta invariante por cambios de base.

Proposición 2.2.9 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 2.2.10 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; es decir que $F, F^{-1} \in C^\infty(U)$.

Proposición 2.2.11 Sea $f : \mathbb{R}^d \rightarrow \mathbb{R}^d$ una función localmente Lipschitz, luego si $\mu(E) = 0$ vale que $\mu(f(E)) = 0$

Lema 2.2.12 Sea $E \subset X$ tal que $\mu(E) = 0$; si $\det(Dg(x)) \neq 0$ para todo $x \in X$, luego $\mu(g^{-1}(E)) = 0$

Demostración Sea $h = g^{-1}$ y (V_i, ψ^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 $\mathcal{A} := \{(U_i, \varphi^i)\}$ un atlas de X y notemos $E_i = E \cap U_i$; luego $E = \bigcup_{i \in \mathbb{N}} E_i = \bigcup_{i \in \mathbb{N}} \varphi^{i-1} \circ \varphi^i(E_i)$ por lo que:

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

Por hipótesis $\varphi^i(E_i)$ es de medida cero, luego como g es difeomorfismo local por 2.2.10 entonces $\psi \circ h \circ \varphi^{i-1} \in C^1$. Como $f \in C^1(\mathbb{R}^d)$ implica que es localmente Lipschitz concluimos por 2.2.11 que $\mu(\psi \circ h \circ \varphi^{i-1}(\varphi^i(E_i))) = 0$ para todo $i \in \mathbb{N}$. ■

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

Proposición 2.2.13 Sea $N \hookrightarrow M$ una subvariedad de dimensión $n < m$, luego para todo $U \subset N$ abierto relativo vale que $\mu(U) = 0$

2.2.2.2 Teorema de la variedad estable

Definición 2.2.14 Sea $f : \mathbb{R}^d \rightarrow \mathbb{R}$ tal que $f \in C^2$, (notando $\nabla^2 f(x) =$ Hessiano de f evaluado en x) y $x^* \in \mathbb{R}^d$, luego:

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

Notaremos χ^* al conjunto de puntos silla estrictos de f .

Teorema 2.2.15 Sea x^* un punto fijo de $g \in C^r(\chi)$ un difeomorfismo local. Supongamos que $E = E_s \oplus E_u$ donde

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

Entonces existe $W_{loc}^{cs} \hookrightarrow \chi$ una inyección difeomorfa 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 [23]

2.2.3 Procesos Estocásticos

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

2.2.3.1 Esperanza condicional

Dado un espacio de probabilidad (Ω, \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 Ω y una familia $(X_\gamma)_{\gamma \in C}$ tal que $X_\gamma : \Omega \mapsto \mathbb{R}$ definimos la σ algebra generada por las X_γ $\mathcal{F} = \sigma(\{X_\gamma\})$ como la menor σ álgebra (en el sentido de la inclusión) tal que todas las X_γ son \mathcal{F} medibles.

Recordemos además:

Teorema 2.2.16 (Teorema de la convergencia monótona) Sea (f_n) una sucesión positiva de elementos medibles en (Ω, Σ, μ) un espacio de medida tal que $f_n \nearrow f$; luego:

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

Teorema 2.2.17 (Teorema de la convergencia dominada) Sea (f_n) una sucesión de elementos medibles en (Ω, Σ, μ) un espacio de medida tal que $f_n \rightarrow f$ c.t.p.; 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, $w_k := w_{k-1} - \alpha_k g(w_k, \cdot) : \Omega \rightarrow \mathbb{R}$ y por hipótesis es \mathcal{F} medible; es más, podemos ver a $W : \mathbb{N}_+ \times \Omega \rightarrow \mathbb{R}$ dado por $W(k, \omega) := w_k$ como un proceso estocástico discreto.

Proposición 2.2.18 Sea (Ω, \mathcal{F}, P) un espacio de probabilidad y X una variable aleatoria tal que $X \in L^1(\Omega)$ (es decir $\mathbb{E}[|X|] < \infty$). Si \mathcal{G} es una sub- σ -álgebra de \mathcal{F} entonces existe Y 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 $Y = \tilde{Y}$ c.t.p..

Definición 2.2.19 Dados X, \mathcal{G} como en la proposición, a la variable aleatoria cuya existencia se prueba en 2.2.18 se 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:

1. **Unicidad c.t.p.**

Sea $X \in L^1$ e Y, \tilde{Y} dos versiones de $\mathbb{E}[X|\mathcal{G}]$ tal que no son iguales c.t.p., 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:

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

Luego $Y = \tilde{Y}$ c.t.p..

2. Existencia en L^2

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

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

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

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

3. 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 2.2.16 demuestra lo que necesitábamos. ■

Teorema 2.2.20 (Propiedades de la esperanza condicional) Sea (Ω, \mathcal{F}, P) un espacio de medida, $X \in L^1$ y \mathcal{G}, \mathcal{H} sub- σ -álgebras de \mathcal{F} , luego:

1. $\mathbb{E}[\mathbb{E}[X|\mathcal{G}]] = \mathbb{E}[X]$
2. Si X es \mathcal{G} medible entonces $X = \mathbb{E}[X|\mathcal{G}]$ c.t.p.
3. $\mathbb{E}[aX + bY|\mathcal{G}] = a\mathbb{E}[X|\mathcal{G}] + b\mathbb{E}[Y|\mathcal{G}]$
4. Si $X \geq 0$ c.t.p., entonces $\mathbb{E}[X|\mathcal{G}] \geq 0$ c.t.p.
5. Si $0 \leq X_n \nearrow X$ entonces $\mathbb{E}[X_n|\mathcal{G}] \nearrow \mathbb{E}[X|\mathcal{G}]$
6. Si $0 \leq X_n$ c.t.p. 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 $X_n \rightarrow X$ c.t.p. vale que $\mathbb{E}[X_n|\mathcal{G}] \rightarrow \mathbb{E}[X|\mathcal{G}]$ c.t.p..
8. Si \mathcal{H} es una sub- σ -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 $\mathbb{E}[ZX|\mathcal{G}] = Z\mathbb{E}[X|\mathcal{G}]$ c.t.p.
10. Si \mathcal{H} es independiente de $\sigma(\sigma(X), \mathcal{G})$ entonces:

$$\mathbb{E}[X|\sigma(\mathcal{G}, \mathcal{H})] = \mathbb{E}[X|\mathcal{G}] \text{ c.t.p.}$$

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

Demostración Una buena referencia resulta [27]. No se incluye la demostración por su extensión y porque no resulta crucial al presente trabajo.

2.2.3.2 Martingalas y Cuasi-martingalas

Definición 2.2.21 Dada una sucesión creciente (en el sentido de la inclusión) de σ 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 (Ω, \mathcal{F}, P) dada por $\mathcal{F}_n = \sigma(X_1, \dots, X_n)$

Definición 2.2.22 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
- $X_n \in L^1$
- $\mathbb{E}[X_n|\mathcal{F}_{n-1}] = X_{n-1}$ c.t.p.

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

$$\mathbb{V}_n(X) = \sup_{J \subset \{1, \dots, n\}} \left\{ \mathbb{E} \left[\sum_{\substack{i_k \in J \\ 1 \leq k \leq |J|}} |\mathbb{E}[X_{i_{k+1}} - X_{i_k} | \mathcal{F}_{i_k}]| \right] \right\} < \infty \quad (2.11)$$

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 acotada se

relacionan con funciones crecientes y decrecientes. En particular, mediante la descomposición de Jordan, cualquier función de variación acotada 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 [20], y significó el inicio de la extensión de gran parte de la teoría de submartingalas a cuasimartingalas.

Proposición 2.2.24 *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 2.2.20 resulta que $V_n(X) = |\mathbb{E}[X_n - X_0]| < \infty$. ■

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 2.2.25 *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 :*

$$\delta_k^u := \begin{cases} 1 & \text{si } \mathbb{E}[u_{k+1} - u_k | \mathcal{P}_k] > 0 \\ 0 & \text{si no} \end{cases} \quad (2.12)$$

Teorema 2.2.26 (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:*

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

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

Demostración Una buena revisión de la demostración se encuentra en el capítulo 9 de [12]

RESUMEN DE RESULTADOS

Si de verdad amas a alguien,
regálale un teorema; eso sí que es
para siempre

Eduardo Sanchez de Cabezón

A modo de síntesis, presentamos un resumen de los resultados vistos en este trabajo.

3.1 ALGORITMOS DE TIPO BATCH

3.1.1 Objetivos convexos

Teorema 3.1.1 Sea $F : \mathbb{R}^d \rightarrow \mathbb{R}$ tal que $F \in C^1$ la función objetivo, asumamos que F es débilmente convexo, w^* su mínimo y que existen $A, B \geq 0$ tal que para todo $w \in \mathbb{R}^d$ vale que:

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

Luego si consideramos el algoritmo de descenso de gradiente por batch tal que los incrementos $\{\alpha_k\}$ cumplen la condición Robbins - Monro entonces:

$$w_k \xrightarrow[k \rightarrow \infty]{} w^*$$

Teorema 3.1.2 Sea $F : \mathbb{R}^d \rightarrow \mathbb{R}$ la función objetivo tal que $F \in C^1$, F es L -Lipshitz y PL -convexa; entonces el algoritmo descenso de gradiente por batch con incremento fijo $\alpha_k = \frac{1}{L}$ cumple:

$$F(w_k) - F_{inf} \leq \left(1 - \frac{\mu}{L}\right)^k (F(w_1) - F_{inf})$$

3.1.2 Objetivos no convexos

Teorema 3.1.3 Sea $F : \mathbb{R}^d \rightarrow \mathbb{R}$ tal que $F \in C^2$ la función objetivo con Hessiano acotado con constante L , w^* algún mínimo local de F ; entonces el algoritmo descenso de gradiente por batch con incremento fijo $\alpha < \frac{1}{L}$ cumple:

$$w_k \xrightarrow[k \rightarrow \infty]{c.t.p.} w^*$$

Teorema 3.1.4 Sea $F : \mathbb{R}^d \rightarrow \mathbb{R}$ tal que $F \in C^2$ la función objetivo con Hessiano acotado con constante L , w^* algún mínimo local de F ; entonces el algoritmo punto próximo con incremento fijo $\alpha < \frac{1}{L}$ cumple:

$$w_k \xrightarrow[k \rightarrow \infty]{c.t.p.} w^*$$

Teorema 3.1.5 Sea $F : \mathbb{R}^d \rightarrow \mathbb{R}$ tal que $F \in C^2$ la función objetivo con Hessiano acotado por bloques con constante L_b , w^* algún mínimo local de F ; entonces el algoritmo descenso de gradiente por coordenadas con incremento fijo $\alpha < \frac{1}{L_b}$ cumple:

$$w_k \xrightarrow[k \rightarrow \infty]{c.t.p.} w^*$$

Teorema 3.1.6 Consideremos el algoritmo descenso de gradiente por batch con w_0 elegido uniformemente en $[-1, 1]^d$; luego existe $F : \mathbb{R}^d \mapsto \mathbb{R}$ función objetivo B -acotada, l -Lipshitz, μ -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 $k \leq e^{\Omega(d)}$

3.2 ALGORITMOS ESTOCÁSTICOS

3.2.1 Objetivos convexos

Teorema 3.2.1 Sea $F : \mathbb{R}^d \rightarrow \mathbb{R}$ tal que $F \in C^1$ la función objetivo tal que existe F_{\inf} valor mínimo, F es L -Lipshitz, F es fuertemente convexa y supongamos además que g tiene varianza acotada; entonces el algoritmo descenso estocástico de gradiente generalizado con incremento fijo $0 < \alpha_k = \alpha \leq \frac{\mu}{LM_G}$ cumple:

$$\mathbb{E} [F(w_k) - F_{\inf}] \xrightarrow[k \rightarrow \infty]{} \frac{\alpha LM}{2c\mu}$$

Teorema 3.2.2 Sea $F : \mathbb{R}^d \rightarrow \mathbb{R}$ tal que $F \in C^1$ la función objetivo tal que existe F_{\inf} valor mínimo, F es L -Lipshitz, F es fuertemente convexa; supongamos además que g tiene varianza acotada y que los incrementos α_k cumplen:

$$\alpha_k = \frac{\beta}{\gamma + k} \quad \text{para algún } \beta > \frac{1}{c\mu} \text{ y } \gamma > 0 \text{ tal que } \alpha_1 \leq \frac{\mu}{LM_G} \quad (3.1)$$

Luego el algoritmo descenso estocástico de gradiente generalizado cumple::

$$\mathbb{E} [R(w_k) - R^*] = \mathcal{O} \left(\frac{1}{k} \right)$$

Teorema 3.2.3 Sea $F : \mathbb{R}^d \rightarrow \mathbb{R}$ tal que $F \in C^1$ la función objetivo tal que existe F_{\inf} valor mínimo, F es L -Lipshitz, F es fuertemente convexa; supongamos además que g tiene varianza acotada geométricamente. Luego el algoritmo descenso estocástico de gradiente generalizado con incremento fijo $0 < \alpha_k = \alpha \leq \min \left\{ \frac{\mu}{L\mu_G^2}, \frac{1}{\mu} \right\}$ cumple:

$$\mathbb{E} [R(w_k) - R^*] = \mathcal{O}(\rho^k)$$

Teorema 3.2.4 Sea $F : \mathbb{R}^d \rightarrow \mathbb{R}$ tal que $F \in C^1$ la función objetivo tal que existe F_{\inf} valor mínimo, F es débilmente convexo; supongamos además que g tiene varianza acotada y que los incrementos cumplen la condición de Robbins Monro. Luego el algoritmo descenso estocástico de gradiente generalizado con incremento decrecientes cumple:

$$w_k \xrightarrow[k \rightarrow \infty]{c.t.p.} w^*$$

$$(w_k - w^*)^T \nabla F(w_k) \xrightarrow[k \rightarrow \infty]{c.t.p.} 0$$

3.2.2 Objetivos no convexos

Teorema 3.2.5 Sea $F : \mathbb{R}^d \rightarrow \mathbb{R}$ tal que $F \in C^1$ la función objetivo, F es L -Lipshitz y supongamos además que g tiene varianza acotada. Luego el algoritmo descenso estocástico de gradiente generalizado con incremento fijo $0 < \alpha_k = \alpha \leq \frac{\mu}{LM_G}$ cumple:

$$\mathbb{E} \left[\frac{1}{K} \sum_{k=1}^K \|\nabla F(w)_k\|^2 \right] \xrightarrow[k \rightarrow \infty]{} \frac{\alpha LM}{\mu}$$

Teorema 3.2.6 Sea $F : \mathbb{R}^d \rightarrow \mathbb{R}$ tal que $F \in C^1$ la función objetivo, F es L -Lipshitz, supongamos además que g tiene varianza acotada y que los incrementos cumplen la condición de Robbins Monro. Luego si notamos $A_K := \sum_{k=1}^K \alpha_k$, el algoritmo descenso estocástico de gradiente generalizado con incrementos decrecientes cumple:

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

Teorema 3.2.7 Sea $F : \mathbb{R}^d \rightarrow \mathbb{R}$ tal que $F \in C^2$ la función objetivo, F es L -Lipshitz y $w \mapsto \|\nabla F(w)\|_2^2$ sea L -Lipshitz; supongamos además que g tiene varianza acotada y que los incrementos cumplen la condición de Robbins Monro. Luego el algoritmo descenso estocástico de gradiente generalizado con incrementos decrecientes cumple:

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

Teorema 3.2.8 Sea $F : \mathbb{R}^d \rightarrow \mathbb{R}$ función objetivo y g un estimador insesgado de ∇F tal que ambos cumplen:

1. $F \in C^3$

2. Existe $w^* \in \mathbb{R}^d$ tal que $F_{inf} = F(w^*) \leq F(w)$ para todo $w \in U$ entorno.

3. $F(w) \geq 0$ para todo $w \in \mathbb{R}^d$

4. Los incrementos cumplen la condición de Robbins Monro

5. Existe $B_2 \geq 0$ tal que:

$$\mathbb{E} \left[\|g(w_k, \xi_k)\|^2 \right] \leq B_2 \|w\|^2$$

6. Para 3,4 existen $A_j, B_j \geq 0$ tal que:

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

7. Existe $D > 0$ tal que:

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

Luego el algoritmo descenso estocástico de gradiente generalizado con incrementos decrecientes cumple:

$$F(w_k) \xrightarrow[k \rightarrow \infty]{c.t.p.} F_\infty$$

$$\nabla F(w_k) \xrightarrow[k \rightarrow \infty]{c.t.p.} 0$$

Parte 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 patológicos donde aunque el algoritmo converga la cantidad de iteraciones para alcanzar un entorno del mínimo va a ser exponencial

“El libro de la naturaleza esta escrito
en el lenguaje de las Matemáticas”

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 1.1].

Un buen inicio es analizar la convergencia puntual del descenso de gradiente y bajo qué condiciones se da. Para esto recordemos que dado una función de pérdida objetivo uno estima el *riesgo esperado* por el *riesgo empírico* (Ver 2.6), luego considerando esto el algoritmo a analizar resulta:

Algoritmus 4.1 : Descenso de gradiente en batch	
1	Input: $F : \mathbb{R}^d \rightarrow \mathbb{R}$, $f : \mathbb{R}^d \rightarrow \mathbb{R}$ tal que $f \in C^1$, $\alpha_k > 0$, $N \in \mathbb{N}$, $w_1 \in \mathbb{R}^d$, $X = \{\xi_j\}_{j \leq N}$ muestra aleatoria i.i.d.
2	for $k \in \mathbb{N}$ do
3	$w_{k+1} \leftarrow w_k - \alpha_k \sum_{j=1}^N \nabla F(\xi_j) = w_k - \alpha_k \nabla R_N(w_k)$

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

Definición 4.0.1 Decimos que $F : \mathbb{R}^d \rightarrow \mathbb{R}$, $f : \mathbb{R}^d \rightarrow \mathbb{R}$ tal 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 $\inf_{\|w - w^*\|^2 > \epsilon} (w - w^*)^T \nabla F(w) > 0$

Observación Si $F : \mathbb{R}^d \rightarrow \mathbb{R}$, tal que $F \in C^1$ es convexa, entonces es débilmente convexa.

En efecto, como F es convexa existe $w^* \in \mathbb{R}^d$ mínimo global y como $F \in C^1$ entonces $\nabla F(w^*) = 0$. Entonces F cumple:

$$(\nabla F(x) - \nabla F(y))^T (x - y) > 0$$

Para todos $x, y \in \mathbb{R}^d$; en particular si $y = w^*$ vemos que F es débilmente convexa.

Observación Notemos que existen funciones no convexas tal que cumplen 4.0.1. En efecto, sea $M > \epsilon > 0$ y basta con tomar:

$$f(x) = \begin{cases} x^2 & \text{si } |x| < M - \epsilon \\ g_+(x) & \text{si } M - \epsilon < x < M + \epsilon \\ g_-(x) & \text{si } -M - \epsilon < x < -M + \epsilon \\ \sqrt{x} & \text{si } |x| > M + \epsilon \end{cases}$$

Con g_+, g_- funciones tal que f resulte C^1 [Para ver la existencia basta ver [A](#)]. Luego f es débilmente convexa pero no convexa pues \sqrt{x} no resulta convexa en $(-\infty, -M - \epsilon) \cup (M + \epsilon, \infty)$

4.1 INTUICIÓN

A fin de ganar intuición acerca del proceso de cómo probar la convergencia del algoritmo [4.1](#) vamos a analizar el caso continuo. En el caso continuo, tenemos que demostrar que la solución $w(t)$ de la ecuación diferencial [4.1](#) tiene límite w^* y además que w^* es mínimo de F .

$$\frac{dw}{dt} = -\nabla F(w) \quad (4.1)$$

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

1. Vamos a definir $h \in C^1(\mathbb{R})$ una *función de Lyapunov*
2. Vamos a verificar computando $\frac{dh}{dt}$ que es una función monótona decreciente y acotada, por lo que converge
3. Vamos a probar que $h(t) \xrightarrow{t \rightarrow \infty} 0$.

Proposición 4.1.1 (Objetivo débilmente convexo, Versión continua)
 Sea $f : \mathbb{R}^d \rightarrow \mathbb{R}$ tal que $f \in C^1$ que cumple [4.0.1](#) y supongamos que el algoritmo [4.1](#) 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 Sigamos la estrategia definida anteriormente:

Paso 1 Definamos la función de Lyapunov:

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

Paso 2 Notemos que:

$$\frac{dh}{dt} = 2(w(t) - w^*) \frac{dw}{dt} = -2(w(t) - w^*) \nabla F(w) \underbrace{\leq}_{\substack{\text{4.0.1} \\ w \neq w^*}} 0 \quad (4.2)$$

Luego como $h(t) \geq 0$ y $\frac{dh}{dt} \leq C < 0$ para $C \in \mathbb{R}$ constante, existe $h_{inf} \geq 0$ tal que $h(t) \searrow h_{inf}$

Paso 3 Como $h(t) \searrow h_{inf}$ entonces existe una sucesión $\{t_k\}_{k \in \mathbb{N}} \subset \mathbb{R}_+$ tal que $\frac{dh}{dt}(t_k) \xrightarrow{k \rightarrow \infty} 0$.

Supongamos por el absurdo que $h_{inf} > 0$, luego existe $\tilde{\epsilon} > 0$ y $K \in \mathbb{N}$ tal que para todo $k \geq K$ vale que $h(t_k) = (w(t_k) - w^*)^2 > \tilde{\epsilon}$. Si juntamos entonces el paso anterior con la definición 4.0.1 llegamos a un absurdo; luego concluimos que $h_{inf} = 0$ por lo que:

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

■

4.2 CASO DISCRETO

Ahora sí, analicemos la convergencia del algoritmo 4.1 para el caso discreto. Para esto enunciamos un lema útil cuya demostración referimos al lector al Apéndice:

Lema 4.2.1 Sea $\{u_k\}_{k \in \mathbb{N}} \subset \mathbb{R}$ una sucesión tal que $u_k \geq 0$ para todo k . Luego si:

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

Con:

$$(x)_+ = \begin{cases} x & \text{si } x > 0 \\ 0 & \text{si no} \end{cases}$$

$Y (x)_-$ definida análogamente. Luego:

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

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 4.1, decimos que los incrementos $\{\alpha_k\}$ cumplen la condición de Robbins - Monro (ver [22]) si:

$$\sum_{k=1}^{\infty} \alpha_k = \infty \quad \text{y} \quad \sum_{k=1}^{\infty} \alpha_k^2 < \infty \quad (4.3)$$

Teorema 4.2.2 (Objetivo débilmente convexo, incrementos decrecientes)

Sea $F : \mathbb{R}^d \rightarrow \mathbb{R}$ tal que $F \in C^1$, cumple la definición 4.0.1 y existen $A, B \geq 0$ tal que para todo $w \in \mathbb{R}^d$ vale que:

$$\|\nabla F(w)\|^2 \leq A + B \|w - w^*\|^2 \quad (4.4)$$

Luego si consideramos el algoritmo 4.1 con incrementos $\{\alpha_k\}$ que cumplen la condición 4.3 entonces:

$$w_k \xrightarrow[k \rightarrow \infty]{} w^* \quad (4.5)$$

Demostración Hagamos los 3 pasos análogos a la demostración de la proposición 4.1.1:

Paso 1 Sea $h_k = \|w_k - w^*\|^2$.

Paso 2 Análogo a la demostración de la proposición 4.1.1 notemos que:

$$\begin{aligned} h_{k+1} &= \|w_{k+1} - w^*\|^2 \\ &= \langle (w_k - w^*) - \alpha_k \nabla F(w_k), (w_k - w^*) - \alpha_k \nabla F(w_k) \rangle \\ &= \|w_k - w^*\|^2 - 2\alpha_k (w_k - w^*)^T \nabla F(w_k) + \alpha_k^2 \|\nabla F(w_k)\|^2 \end{aligned}$$

Luego:

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

A diferencia de antes la naturaleza discreta del algoritmo lleva a un término positivo en las variaciones, pero si usamos la condición 4.3 y la hipótesis 4.4 entonces:

$$\begin{aligned} h_{k+1} - h_k &= \underbrace{-2\alpha_k (w_k - w^*)^T \nabla F(w_k)}_{\substack{\leq 0 \\ \text{por definición 4.0.1}}} + \alpha_k^2 \|\nabla F(w_k)\|^2 \\ &\leq \alpha_k^2 \|\nabla F(w_k)\|^2 \\ &\leq \alpha_k^2 (A + B h_k) \end{aligned}$$

Por lo que:

$$h_{k+1} - (1 + \alpha_k^2 B) h_k \leq \alpha_k^2 A \quad (4.6)$$

Definamos ahora las sucesiones auxiliares:

$$\mu_k = \prod_{j=1}^{k-1} \frac{1}{1 + \alpha_j^2 B} \quad (4.7a)$$

$$h'_k = \mu_k h_k \quad (4.7b)$$

Notemos que $\log(\mu_k) = -\sum_{j=1}^{k-1} \log\left(1 + \underbrace{\alpha_j^2 B}_{\geq 0}\right) \geq -B \sum_{j=1}^{k-1} \alpha_j^2 \geq -B \sum_{j=1}^{\infty} \alpha_j^2$, por lo que μ_k es una sucesión decreciente acotada inferiormente por $e^{-B \sum_{j=1}^{\infty} \alpha_j^2}$, luego $\mu_k \searrow \mu_{\infty} > 0$. Ahora si volvemos a la ecuación 4.6 del paso anterior y observamos que $\mu_{k+1} = \frac{\mu_k}{1 + B\alpha_k^2}$ 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 4.2.1 concluimos que $\{h'_k\}$ converge; como $\underbrace{\mu_k}_{\geq 0} \rightarrow \mu_{\infty} > 0$ entonces $\{h_k\}$ converge.

Paso 3 De la ecuación 4.6 del paso anterior 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 la definición 4.0.1 concluimos que existe $M > 0$ tal que $M \leq (w_k - w^*) \nabla F(w_k)$ para todo $k \geq K$. Por la condición 4.3 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 4.2.3 Sea $F \in C^2$, asumamos 4.0.1 y que $\|\nabla^2 F\|_2^2 \leq L$; si consideramos el algoritmo 4.1 tal que incrementos $\{\alpha_k\}$ cumplen 4.3 entonces:

$$w_k \xrightarrow[k \rightarrow \infty]{} w^* \quad (4.8)$$

4.3 ACERCA DE CONVEXIDAD FUERTE Y FUNCIONES L-LIPSCHITZ

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-Lipschitz** y **Convexidad fuerte**.

Definición 4.3.1 Sea $f : \mathbb{R}^d \rightarrow \mathbb{R}$ tal que $f \in C^1$ y $\mu > 0$, decimos que es fuertemente convexa o μ -convexa si para todos $x, y \in \mathbb{R}^d$ vale:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2} \|y - x\|_2^2 \quad (4.9)$$

Proposición 4.3.2 Sea $f : \mathbb{R}^d \rightarrow \mathbb{R}$ tal que $f \in C^1$ y $\mu > 0$, 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 4.3.3 Decimos que una función $f : \mathbb{R}^d \rightarrow \mathbb{R}$ tal que $f \in C^1$ es PL-convexa, o cumple la condición de Polyak-Lojasiewicz (ver [\[19\]](#), [\[11\]](#)) si existe $\mu > 0$ tal que para todo $x \in \mathbb{R}^d$ vale:

$$\frac{1}{2} \|\nabla f(x)\|_2^2 \geq \mu (f(x) - f_{\inf}) \quad (4.10)$$

Proposición 4.3.4 Sea $f : \mathbb{R}^d \rightarrow \mathbb{R}$ tal que $f \in C^1$ una función μ -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](#)

Definición 4.3.5 Dada una función objetivo $f : \mathbb{R}^d \rightarrow \mathbb{R}$ tal que $f \in C^1$ definimos el problema de optimización regularizado en L2 cuando se minimiza $\tilde{f} = f + \lambda \|x\|_2^2$ para $\lambda > 0$.

En Machine Learning es usual resolver el problema de optimización regularizado en vez de el problema original. Esto se debe mayormente al siguiente resultado, que aunque no lo usemos *per-se* en esta tesis, es de sumo interés:

Corolario 4.3.6 Sea $h = f + g$ donde g es fuertemente convexa y f es convexa, entonces h es fuertemente convexa. En particular, si f es convexa entonces $\tilde{f} = f + \lambda \|x\|_2^2$ es fuertemente convexo.

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

$$\begin{aligned}
 h(\alpha x + (1 - \alpha)y) &= f(\alpha x + (1 - \alpha)y) + g(\alpha x + (1 - \alpha)y) \\
 &\stackrel{\text{Por proposición 4.3.2}}{\leq} \alpha(f + g)(x) + (1 - \alpha)(f + g)(y) - \frac{\mu\alpha(1 - \alpha)}{2} \|x - y\|_2^2 \\
 &= \alpha h(x) + (1 - \alpha)h(y) - \frac{\mu\alpha(1 - \alpha)}{2} \|x - y\|_2^2
 \end{aligned}$$

■

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

Definición 4.3.7 Sea $f : \mathbb{R}^d \rightarrow \mathbb{R}$ tal que $f \in C^1$, decimos que es L -Lipschitz global si existe $L > 0$ tal que para todos $x, y \in \mathbb{R}^d$ vale:

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L \|y - x\|_2 \quad (4.11)$$

Proposición 4.3.8 Sea $f : \mathbb{R}^d \rightarrow \mathbb{R}$ tal que $f \in C^1$, 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 μ -convexa entonces las 8 propiedades son equivalentes

Demostración Ver [A](#)

Con todas estas propiedades, probemos el resultado histórico de [\[19\]](#)

Teorema 4.3.9 (Convergencia lineal, Objetivos L-Lipschitz y PL-convexos)

Sea $F : \mathbb{R}^d \rightarrow \mathbb{R}$ tal que $F \in C^1$, existe F_{inf} valor mínimo, F es L -Lipschitz y PL -convexa; entonces el algortimo [4.1](#) con incremento fijo $\alpha_k = \frac{1}{L}$ cumple:

$$F(w_k) - F_{inf} \leq \left(1 - \frac{\mu}{L}\right)^k (F(w_1) - F_{inf}) \quad (4.12)$$

Demostración Notemos que si usamos 4.3.8 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 la definición 4.3.3 tenemos:

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

Luego obtenemos:

$$F(w_{k+1}) - F_{inf} \leq \left(1 - \frac{\mu}{L}\right) (F(w_k) - F_{inf}) \leq \left(1 - \frac{\mu}{L}\right)^k (F(w_1) - F_{inf})$$

■

TEOREMA DE LA VARIEDAD ESTABLE Y LOS PUNTOS FIJOS INESTABLES

El matemático que no tenga algo de poeta nunca podrá ser un matemático completo.

Karl Weierstrass

5.1 INTUICIÓN

Del capítulo anterior ya sabemos que bajo condiciones de convexidad estándar el algoritmo 4.1 converge puntualmente. Nos surge entonces la pregunta:

Bajo qué casos el algoritmo 1.1 converge (en alguna forma) con objetivos no convexos?

En el caso no convexo existen puntos extremos no óptimos entre los cuales se encuentran los puntos silla, máximos y mínimos locales "grandes" (E.g. Mínimos locales 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. Intuitivamente esto es claro, ya que asumiendo suficiente regularidad, si tomamos un máximo local w^* y un entorno de éste $B \ni w^*$ entonces como el gradiente apunta a la dirección de máximo crecimiento y este es en la dirección a w^* , la iteración de descenso de gradiente cumplirá que $\|w^* - \alpha \nabla F(w_k)\|^2$ es una sucesión creciente en B .

Usemos un caso modelo para ejemplificar por qué no es probable que los métodos de primer orden (entre ellos el algoritmo 4.1) 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 $\lambda_{k+1}, \dots, \lambda_n < 0$ con $k < n$.

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:

$$\nabla f(x) = \sum_{i=1}^n \lambda_i x_i e^i = 0 \iff x = 0$$

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

Recordemos que si $g(x) = x - \alpha \nabla f(x)$ entonces el algoritmo 4.1 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:

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

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:

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

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

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

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:

$$\begin{aligned} \nabla f &= (x, y^3 - y) \\ g &= ((1 - \alpha)x, (1 + \alpha)y - \alpha y^3) \\ \nabla^2 f &= \begin{pmatrix} 1 & 0 \\ 0 & 3y^2 - 1 \end{pmatrix} \end{aligned} \tag{5.1}$$

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 5.1.1 Para $f(x, y) = \frac{1}{2}x^2 + \frac{1}{4}y^4 - \frac{1}{2}y^2$ resulta que $E_s = \langle (t, 0, \dots, 0) / t \in \mathbb{R} \rangle := W_s$

Asumiendo el resultado por un momento, dado que $\dim_{\mathbb{R}^2}(E_s) = 1 < 2$ entonces $|E_s| = 0$ 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 5.1, 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:

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

Luego, bajo esas condiciones g_y resulta una contracción y como \mathbb{R} es completo g_y converge a un punto fijo y_f .

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

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

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

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

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

Definición 5.2.1 *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 del capítulo 2 demostremos el teorema principal para analizar la convergencia de los algoritmos de tipo batch en el caso no convexo:

Teorema 5.2.2 *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 Sea $x^* \in \mathcal{A}_g^*$, luego por el teorema de la función inversa existe B_{x^*} un entorno abierto de χ tal que g resulta difeomorfismo local; es más, $\bigcup_{x^* \in \mathcal{A}_g^*} B_{x^*}$ forma un cubrimiento abierto de \mathcal{A}_g^* 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 5.2.3 $x_t \in \bigcap_{k \in \mathbb{N}} g^{-k}(B_{x_i^*})$ para todo $t \geq T$

Demostración [Del lema] Sea $k_0 \in \mathbb{N}$ y $\tilde{t} \geq T$, luego $g^{k_0}(x_{\tilde{t}}) = x_{\tilde{t}+k_0} \in B_{x_i^*}$ pues $\tilde{t} + k_0 \geq T$. Como k_0 era arbitrario y \tilde{t} era arbitrario sujeto a ser mayor que T concluimos que $x_t \in \bigcap_{k \in \mathbb{N}} g^{-k}(B_{x_i^*})$ para todo $t \geq T$. ■

Si notamos $S_i \triangleq \bigcap_{k \in \mathbb{N}} g^{-k}(B_{x_i^*})$, entonces por 2.2.15 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 2.2.13 $\mu(S_i) = 0$.

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:

$$\begin{aligned} \mu(W) &\leq \mu\left(\bigcup_{i \in \mathbb{N}} \bigcup_{k \in \mathbb{N}} g^{-k}(S_i)\right) \\ &\leq \sum_{i \in \mathbb{N}} \sum_{k \in \mathbb{N}} \mu(g^{-k}(S_i)) \\ &\stackrel{2.2.12}{=} 0 \end{aligned}$$

■

Para finalizar veamos un caso simple que nos encontraremos seguido:

Corolario 5.2.4 Bajo las mismas hipótesis que en 5.2.2 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 $\mu(W_g) \leq \mu(W) \stackrel{5.2.2}{=} 0$. ■

CONVERGENCIA CTP A MÍNIMOS : CASO GENERAL

6.1 DESCENSO DE GRADIENTE EN BATCH

La diferencia entre los matemáticos y los físicos es que luego de que los físicos prueban un resultado grande piensan que es fantástico, pero luego que los matemáticos lo hacen piensan que es trivial.

Richard Feynman

Como una aplicación del teorema en 5.2.2 demostraremos que el *descenso de gradiente en batch* tiene probabilidad cero de converger a puntos silla. Consideremos el algoritmo 4.1 con incrementos constantes $\alpha_k = \alpha$:

$$x_{k+1} = g(x_k) \triangleq x_k - \alpha \nabla f(x_k) \quad (6.1)$$

Hipótesis 6.1.1 Asumamos que $f \in \mathcal{C}^2$ y $\|\nabla^2 f(x)\|_2 \leq L$

Proposición 6.1.2 Todo punto silla estricto de f es un punto fijo inestable de g , es decir $\chi^* \subseteq \mathcal{A}_g^*$.

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 2.2.4 los autovalores de Dg son $\{1 - \alpha \lambda_i : \lambda_i \text{ autovalor de } \nabla^2 f\}$. Como $x^* \in \chi^*$ existe $\lambda_{j^*} < 0$ por lo que $1 - \alpha \lambda_{j^*} > 1$; concluimos que $x^* \in \mathcal{A}_g^*$. ■

Proposición 6.1.3 Bajo 6.1.1 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 6.1.1 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 6.1.4 Sea g dada por el algoritmo 4.1, bajo 6.1.1 y $\alpha < \frac{1}{L}$ se tiene que $\mu(W_g) = 0$.

Demostración Por 6.1.2 y 6.1.3 tenemos que vale 5.2.4 y concluimos que $\mu(W_g) = 0$. ■

6.2 PUNTO PRÓXIMO

El algoritmo de punto próximo esta dado por la iteración:

$$x_{k+1} = g(x_k) \triangleq \arg \min_{z \in \chi} f(z) + \frac{1}{2\alpha} \|x_k - z\|_2^2 \quad (6.2)$$

Proposición 6.2.1 Bajo 6.1.1 y $\alpha < \frac{1}{L}$ entonces vale:

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

Demostración Veamos primero el siguiente lema:

Lema 6.2.2 Bajo 6.1.1, $\alpha < \frac{1}{L}$ y $x \in \chi$ entonces $f(z) + \frac{1}{2\alpha} \|x - z\|_2^2$ es estrictamente convexa, por lo que $g \in \mathcal{C}^1(\chi)$

Demostración [Del lema] En efecto, por hipótesis sumado a 4.3.8 y 4.3.6 concluimos que $g \in \mathcal{C}^1(\chi)$. ■

Por lo tanto por 6.2.2 podemos tomar límite, i. e.

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

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 λ_i autovalores de $\nabla^2 f(x^*)$; luego como $\alpha < \frac{1}{L}$ se tiene que $\det(Dg(x)) \neq 0$. ■

Corolario 6.2.3 Sea g dado por el algoritmo de punto próximo con ecuación 6.2, bajo 6.1.1 y $\alpha < \frac{1}{L}$ se tiene que $\mu(W_g) = 0$.

Demostración Por 6.2.1 tenemos que vale 5.2.4 y concluimos que $\mu(W_g) = 0$. ■

6.3 DESCENSO POR COORDENADAS

Sea S_1, \dots, S_b una partición disjunta de $\{1, \dots, d\}$ donde d y b son parámetros del método, por ejemplo si $d = 4$ y $b = 2$ una posible opción es $S_1 = \{1, 3\}$ y $S_2 = \{2, 4\}$.

Consideremos el algoritmo 6.1:

Algorithmus 6.1 : Descenso por coordenadas	
1	Input: $f \in C^1$, $\alpha > 0$, $x_0 \in \chi$
2	for $k \in \mathbb{N}$ do
3	for block $i = 1, \dots, b$ do
4	for index $j \in S_i$ do
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 6.3.1 La iteración de Descenso por coordenadas esta dada por:

$$x_{k+1} = g(x_k) \triangleq g_b \circ g_{b-1} \circ \dots \circ g_1(x_k) \quad (6.3)$$

Lema 6.3.2 Si g está dada por la ecuación 6.3 y $P_S = \sum_{i \in S} e_i e_i^T$ entonces:

$$Dg(x_k) = \prod_{i \in \{1, \dots, b\}} \left(Id - \alpha P_{b-i+1} \nabla^2 f(y_k^{S_{b-i}}) \right) \quad (6.4)$$

Demostración Notemos primero que:

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

Por lo tanto:

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

■

Hipótesis 6.3.3 Sea $f \in C^2$ y $\nabla^2 f|_S$ la submatriz compuesta de las filas y columnas indexadas por S . Por ejemplo si $f \in C^2(\mathbb{R}^3)$ y $S = \{1, 2\}$

$$\text{entonces } \nabla^2 f|_S = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 x_2} \\ \frac{\partial^2 f}{\partial x_2 x_1} & \frac{\partial^2 f}{\partial x_2^2} \end{pmatrix}$$

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

Proposición 6.3.4 Sea $f \in C^2$ y asumamos 6.3.3, luego 6.1 con $\alpha < \frac{1}{L_b}$ cumple que $\det(Dg(x)) \neq 0$

Demostración Basta probar que cada término de 6.4 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 6.3.5 Bajo 6.3.3 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 2.2.2:

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

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 6.3.6 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 6.3.7 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 Ver A

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 6.3.8 Sea g dado por el algoritmo de descenso por coordenadas con ecuación 6.3, bajo 6.3.3 y $\alpha < \frac{1}{L_b}$ se tiene que $\mu(W_g) = 0$.

Demostración Por 6.3.4 y 6.3.5 tenemos que vale 5.2.4 y concluimos que $\mu(W_g) = 0$. ■

RESULTADOS NEGATIVOS

La esencia de las Matemáticas yace
en su libertad.

Georg Cantor

Ya vimos de 6.1.4, que el descenso de gradiente, con cualquier inicialización aleatoria razonable, tiene probabilidad 1 de converger a mínimos locales, 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?

7.1 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 4.1 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 4.1 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 $f \in C^2((-\infty, 1] \times [-1, 1])$ 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 conectando las regiones $(-\infty, -2] \times [-1, 1]$ y $[-1, 1] \times [-1, 1]$ de modo que $f \in C^2$ y $\frac{\partial f}{\partial x_2}|_{[-2, -1] \times [-1, 1]} < M$ con $M > 0$ constante. Notemos que esta h existe por A.1.1.

Luego si para el algoritmo 4.1 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 toma tiempo al menos exponencial en converger a cualquier mínimo si $w_0 \in \tilde{U}$. ■

7.2 INTUICIÓN ACERCA DEL TEOREMA

Definición 7.2.1 Dado $B \in \mathbb{R}^d$ decimos que $B \in \text{poly}(d)$ si existe $p \in \mathbb{R}[X]$ tal que $p(d) = B$. Asimismo decimos que una iteración de un algoritmo w_k esta a $\Omega(f(k))$ de w^* si existe $K \in \mathbb{N}$ tal que $|w^* - w_k| \geq Kf(k)$

Teorema 7.2.2 (Convergencia exponencial, Inicialización uniforme en el cubo)

Consideremos el algoritmo 4.1 con w_0 elegido uniformemente en $[-1, 1]^d$; luego existe $f : \mathbb{R}^d \mapsto \mathbb{R}$ B -acotada, l -Lipschitz, μ -Lipschitz 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:

$$f(x_1, x_2) = \begin{cases} -\gamma x_1^2 + Lx_2^2 & \text{si } (x_1, x_2) \in [0, 1] \times [0, 1] \\ L(x_1 - 2)^2 - \gamma x_2^2 & \text{si } (x_1, x_2) \in [1, 3] \times [0, 1] \\ L(x_1 - 2)^2 + L(x_2 - 2)^2 & \text{si } (x_1, x_2) \in [1, 3] \times [1, 3] \end{cases} \quad (7.1)$$

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 están bien definidos y que $k_2 \geq k_1 \geq 0$. Vamos a probar 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; y esta propiedad se va a mantener en el caso general y nos permitirá concluir el resultado. En pos de esto, veamos como va a ser la iteración del algoritmo 1.1:

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

Luego evaluando en k_1 y k_2 :

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

concluimos que:

$$k_2 \geq \frac{2\alpha(L + \gamma)k_1 - \log(x_2^0)}{2\alpha\gamma} \geq \frac{L + \gamma}{\gamma} k_1 \quad (7.3)$$

Esta f que presentamos tiene varios problemas:

1. No es continua ni mucho menos C^2
2. No podemos asegurar que f sea l -Lipschitz o μ -Lipschitz en el hessiano
3. Los puntos críticos están en el borde del dominio, lo que no es ideal
4. No está definida en todo \mathbb{R}^d

La clave va a ser usar splines para resolver los primeros puntos, esparcir f para hacer los puntos extremos 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.

7.3 DEMOSTRACIÓN DEL TEOREMA

Demostración Segmentemos esta demostración en diferentes secciones:

Paso 1 - Definiciones

Fijemos 4 constantes: $L \geq e, \gamma \geq 1, \tau \geq e, \eta$ 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 1.1 sea exponencial. Sea D_0 dado por:

$$\begin{aligned} D_0 &= \bigcup_{i=1}^{d+1} \{x \in \mathbb{R}^d : 6\tau \geq x_1, \dots, x_{i-1} \geq 2\tau; 2\tau \geq x_i \geq 0; \tau \geq x_{i+1}, \dots, x_d \geq 0\} \\ &:= \bigcup_{i=1}^{d+1} D_i \end{aligned} \quad (7.4)$$

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:

$$f|_{D_i}(x) = \begin{cases} \sum_{j=1}^{i-1} L(x_j - 4\tau)^2 - \gamma x_i^2 + \sum_{j=i+1}^d Lx_j^2 - (i-1)\eta \\ \triangleq f_{i,1}(x) \text{ si } x \in D_{i,1} \\ \\ \sum_{j=1}^{i-1} L(x_j - 4\tau)^2 + g(x_i, x_{i+1}) + \sum_{j=i+2}^d Lx_j^2 - (i-1)\eta \\ \triangleq f_{i,2}(x) \text{ si } x \in D_{i,2} \end{cases} \quad (7.5)$$

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

Para $i = d$ definamos:

$$f|_{D_d}(x) = \begin{cases} \sum_{j=1}^{d-1} L(x_j - 4\tau)^2 - \gamma x_d^2 - (d-1)\eta \\ \triangleq f_{d,1}(x) \text{ si } x \in D_{d,1} \\ \\ \sum_{j=1}^{d-1} L(x_j - 4\tau)^2 + g_1(x_d) - (d-1)\eta \\ \triangleq f_{d,2}(x) \text{ si } x \in D_{d,2} \end{cases} \quad (7.6)$$

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

$$f|_{D_{d+1}} = \sum_{j=1}^d L(x_j - 4\tau)^2 - d\eta \triangleq f_{d+1,1} \quad (7.7)$$

Lema 7.3.1 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 están acotados. Concluimos que ambas son B -acotadas y μ -Lipschitz con $B, \mu \in \text{poly}(d)$

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

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

Paso 2- Cota superior a T_k^τ

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

Sea $T_0 = 0$ y definamos $T_k = \min_{x_k^t \geq 2\tau} t$ el tiempo de escape de $D_{k,2}$; notemos que como $x^0 \in D_{1,1}$ vale que $T_k \geq 0$ para todo k y esta bien

definido. Definamos además T_k^τ como la cantidad de iteraciones que x^k esta en $D_{k,2}$ antes de escapar; como del lema $\frac{\partial g}{\partial x_k}(x_k, x_{k+1}) \leq -2\gamma\tau$ tenemos que $|x^k - x^{k+1}| \geq 2\alpha\gamma\tau$ 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 4.1 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 4.1 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 7.3.1 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^τ :

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

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

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

Inductivamente:

$$T_d \geq T_d - T_d^\tau \geq \left(\frac{L + \gamma}{\gamma} \right)^{d-1} (T_1 - T_1^\tau) \geq \frac{1}{2\alpha\gamma} \left(\frac{L + \gamma}{\gamma} \right)^{d-1} \geq \left(\frac{L + \gamma}{\gamma} \right)^{d-1} \quad (7.8)$$

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 4.1 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:

$$\begin{aligned} D_a &= \bigcup_{i=1}^d \left\{ x \in \mathbb{R}^d : x_i \geq 0 \text{ si } i \in a_2(0), x_i \leq 0 \text{ si no,} \right. \\ &\quad \left. 6\tau \geq |x_1|, \dots, |x_{i-1}| \geq 2\tau, |x_i| \leq 2\tau, |x_{i+1}|, \dots, |x_d| \leq \tau \right\} \end{aligned}$$

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

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 4.1 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 $\{x^k\} \subset D$. ■

Parte 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.

El axioma de elección es obviamente cierto, el principio de buen ordenamiento es obviamente falso y quien puede decidir sobre el lema de Zorn?

Jerry Bona

8.1 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 los datos (x, y) como extraídos de una variable aleatoria ξ , cuya distribución desconocida es P , luego F se representa como:

$$F(w) = \begin{cases} R(w) = \mathbb{E}[f(w, \xi)] \\ \text{o} \\ R_n(w) = \frac{1}{n} \sum_{i=1}^n f_i(w) \end{cases} \quad (8.1)$$

Sea el algoritmo estocástico [8.1](#)

Algorithmus 8.1 : 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 ξ_k puede ser un único par (x_i, y_i) como en el *Descenso por gradiente estocástico* o una muestra $S_n = \{(x_i, y_i)\}_{i \leq n}$ como en *mini-batch descenso por gradiente estocástico*; a su vez, $g(w_k, \xi_k)$ puede ser varias estimaciones del gradiente como por ejemplo:

$$g(w_k, \xi_k) = \begin{cases} \nabla f(w_k, \xi_k) \\ \frac{1}{n_k} \sum_{i=1}^{n_k} \nabla f(w_k, \xi_{k,i}) \\ H_k \frac{1}{n_k} \sum_{i=1}^{n_k} \nabla f(w_k, \xi_{k,i}) \end{cases} \quad (8.2)$$

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 4.3.7:

Hipótesis 8.1.1 (F es L -Lipshitz) La función a optimizar $F : \mathbb{R}^d \rightarrow \mathbb{R}$ es de clase C^1 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 8.1.1, 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 4.3.8

8.2 ALGUNOS LEMAS FUNDAMENTALES

Con el contexto claro, veamos algunos lemas que van a ser clave en la demostración de la convergencia L1 del algoritmo 8.1.

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

Lema 8.2.1 Bajo 8.1.1 las iteraciones de 8.1 satisfacen que para todo $k \in N$:

$$\mathbb{E}_{\xi_k} [F(w_{k+1})] - F(w_k) \leq -\alpha_k \nabla F(w_k)^T \mathbb{E}_{\xi_k} [g(w_k, \xi_k)] + \frac{1}{2} \alpha_k^2 \mathbb{E}_{\xi_k} [\|g(w_k, \xi_k)\|_2^2] \quad (8.3)$$

Demostración Notemos que por 8.1.1 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 2.2.20:

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

■

Observación Notemos que si $g(w_k, \xi_k)$ es un estimador insesgado de $\nabla F(w_k)$ entonces de 8.2.1:

$$\mathbb{E}_{\xi_k} [F(w_{k+1})] - F(w_k) \leq -\alpha_k \|\nabla F(w_k)\|^2 + \frac{1}{2} \alpha_k^2 \mathbb{E}_{\xi_k} [\|g(w_k, \xi_k)\|_2^2] \quad (8.4)$$

Luego entonces para controlar la convergencia de 8.1 también hay que poner suposiciones sobre el segundo momento de g , luego si definimos:

$$\mathbb{V}_{\xi_k} [g(w_k, \xi_k)] := \mathbb{E}_{\xi_k} [\|g(w_k, \xi_k)\|_2^2] - \|\mathbb{E}_{\xi_k} [g(w_k, \xi_k)]\|_2^2 \quad (8.5)$$

Asumamos:

Hipótesis 8.2.2 (Acotaciones al primer y segundo momento de g) Supongamos que dada F función objetivo y g la estimación del gradiente en 8.1 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:

$$\nabla F(w_k)^T \mathbb{E}_{\xi_k} [g(w_k, \xi_k)] \geq \mu \|\nabla F(w_k)\|_2^2 \quad (8.6a)$$

Y

$$\|\mathbb{E}_{\xi_k} [g(w_k, \xi_k)]\|_2 \leq \mu_G \|\nabla F(w_k)\|_2 \quad (8.6b)$$

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

$$\mathbb{V}_{\xi_k} [g(w_k, \xi_k)] \leq M + M_V \|\nabla F(w_k)\|_2^2 \quad (8.7)$$

Observación Notemos que si g es un estimador insesgado de ∇F entonces 8.6a y 8.6b valen con $\mu_G = \mu = 1$.

Observación Bajo 8.2.2 y por 8.5 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 8.2.3 Bajo 8.2.2 y 8.1.1 las iteraciones de 8.1 satisfacen para todo $k \in \mathbb{N}$:

$$\mathbb{E}_{\xi_k} [F(w_{k+1})] - F(w_k) \leq -\mu \alpha_k \|\nabla F(w_k)\|_2^2 + \frac{1}{2} \alpha_k^2 L \mathbb{E}_{\xi_k} [\|g(w_k, \xi_k)\|_2^2] \quad (8.8a)$$

$$\mathbb{E}_{\xi_k} [F(w_{k+1})] - F(w_k) \leq -\left(\mu - \frac{1}{2} \alpha_k L M_G\right) \alpha_k \|\nabla F(w_k)\|_2^2 + \frac{1}{2} \alpha_k^2 L M \quad (8.8b)$$

Demostración Por 8.2.1 y 8.6a 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 8.8a; luego por 8.2 obtenemos 8.8b. ■

Corolario 8.2.4 Bajo 8.2.2 y 8.1.1 las iteraciones de 8.1 satisfacen para todo $k \in \mathbb{N}$ que $\{w_k\}$ es una cadena de Markov de primer orden.

8.3 CASO FUERTEMENTE CONVEXO

Consideremos primero los casos de convexidad donde sabemos que el mínimo existe y es único, por lo tanto asumamos por ahora 4.3.1:

Hipótesis 8.3.1 (Convexidad fuerte) Supongamos que la función objetivo $F : \mathbb{R}^d \mapsto \mathbb{R}$ es fuertemente convexa, es decir que cumple 4.3.1.

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

Ya estamos en condiciones de demostrar nuestro primer resultado de convergencia para 8.1 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 8.8b se ve que el segundo término es constante.

Dado w_k que depende de ξ_1, \dots, ξ_{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 8.3.2 (Objetivo fuertemente convexo, Incremento constante)

Supongamos 8.1.1, 8.2.2 y 4.3.1; además supongamos que dado 8.1 $\alpha_k = \alpha > 0$ constante tal que:

$$0 < \alpha \leq \frac{\mu}{LM_G} \quad (8.9)$$

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 8.2.3 con 8.9 y 4.3.8 tenemos para todo $k \in \mathbb{N}$ que:

$$\begin{aligned}
\mathbb{E}_{\xi_k} [F(w_{k+1}) - F(w_k)] &\stackrel{\text{8.2.3}}{\leq} -\left(\mu - \frac{1}{2}\alpha LM_G\right) \alpha \|\nabla F(w_k)\|_2^2 + \frac{1}{2}\alpha^2 LM \\
&\stackrel{\text{8.9}}{\leq} -\frac{1}{2}\alpha\mu \|\nabla F(w_k)\|_2^2 + \frac{1}{2}\alpha^2 LM \\
&\stackrel{\text{4.3.8}}{\leq} -\alpha\mu c (F(w_k) - F_{inf}) + \frac{1}{2}\alpha^2 LM
\end{aligned}$$

Luego si restamos F_{inf} y tomamos esperanza total (definida en 8.3):

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

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 ∇F entonces $\mu = M_G = 1$ por lo que $\alpha \in [0, \frac{1}{L})$ que es la condición que pedimos en 6.1.4.

Observación Notemos además que si $M = 0$ (o sea el algoritmo 8.1 no tiene ruido) entonces la convergencia es lineal, recuperando el resultado de 4.3.9.

Observación Notemos finalmente que hay un compromiso entre el primer y segundo término de 8.3.2 donde a un α más cercano a $\frac{\mu}{LM_G}$ acelera la convergencia del primer término, pero a costa de un entorno final de mayor volumen.

Luego esto llevó a varios investigadores a tomar un enfoque experimental 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 8.3.3 (Objetivo fuertemente convexo, Incremento decreciente)

Supongamos 8.1.1, 8.2.2 y 4.3.1; además supongamos que dado 8.1 α_k cumple:

$$\alpha_k = \frac{\beta}{\gamma + k} \quad \text{para algún } \beta > \frac{1}{c\mu} \text{ y } \gamma > 0 \text{ tal que } \alpha_1 \leq \frac{\mu}{LM_G} \quad (8.10)$$

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

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

Luego por 8.2.3 y 4.3.8 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.

Para el caso $k = 1$ notemos que por definición $\eta \geq (\gamma + 1) (F(w_1) - F_{inf})$ por lo que trivialmente:

$$\mathbb{E} [(F(w_1) - F_{inf})] \leq \mathbb{E} \left[\frac{\eta}{\gamma + 1} \right] = \frac{\eta}{\gamma + k}$$

Si asumimos que vale el resultado para algún $k \geq 1$ entonces:

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

$(\gamma + k)^2 \geq (\gamma + k + 1)(\gamma + k - 1)$

■

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 α fijo con $M = 0$) si con el ruido existente pero controlado podemos mantener la convergencia en L_1 .

Teorema 8.3.4 (Objetivo Fuertemente Convexo, Reducción del Ruido)

Supongamos que valen 8.1.1, 8.2.2 y 4.3.1 pero reforcemos 8.7 a la existencia de una constante $M \geq 0$ y $\zeta \in (0, 1)$ tal que para todo $k \in \mathbb{N}$:

$$\mathbb{V}_{\xi_k} [g(w_k, \xi_k)] \leq M\zeta^{k-1} \quad (8.11)$$

Supongamos además que 8.1 tiene $\alpha_k = \alpha$ para todo $k \in \mathbb{N}$ satisfaciendo:

$$0 < \alpha \leq \min \left\{ \frac{\mu}{L\mu_G^2}, \frac{1}{\mu} \right\} \quad (8.12)$$

Luego vale:

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

Donde:

$$\omega := \max \left\{ \frac{\alpha LM}{c\mu}, F(w_1) - F_{inf} \right\} \quad (8.13a)$$

$$\rho := \max \left\{ 1 - \frac{\alpha c\mu}{2}, \zeta \right\} < 1 \quad (8.13b)$$

Demostración Por 8.8a 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 8.5, 8.6b, 8.12 y 8.11 entonces:

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

Por lo tanto:

$$\mathbb{E} [F(w_{k+1}) - F_{inf}] \leq (1 - c\mu\alpha) \mathbb{E} [F(w_k) - F_{inf}] + \frac{1}{2}\alpha^2 LM\zeta^{k-1} \quad (8.14)$$

Probemos ahora el resultado por inducción.

Para el caso $k = 1$ notemos que por definición $\omega \geq F(w_1) - F_{inf}$ por lo que:

$$\mathbb{E} [\mathcal{F}(w_1) - F_{inf}] \leq \mathbb{E} [\omega] = \omega = \omega \rho^0 = \omega \rho^{k-1}$$

Si asumimos que vale para algún $k \geq 1$ entonces de 8.14, 8.13a y 8.13b:

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

■

8.4 CASO GENERAL

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

Teorema 8.4.1 (Objetivo no convexo, Incrementos fijos) *Asumiendo 8.1.1 y 8.2.2 y suponiendo que en 8.1 tenemos $\alpha_k = \alpha$ tal que:*

$$0 < \alpha \leq \frac{\mu}{LM_G} \quad (8.15)$$

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

$$\mathbb{E} \left[\sum_{k=1}^K \|\nabla F(w)_k\|_2^2 \right] \leq \frac{K\alpha LM}{\mu} + \frac{2(F(w_1) - F_{inf})}{\mu\alpha} \quad (8.16a)$$

$$\begin{aligned} \mathbb{E} \left[\frac{1}{K} \sum_{k=1}^K \|\nabla F(w)_k\|_2^2 \right] &\leq \frac{\alpha LM}{\mu} + \frac{2(F(w_1) - F_{inf})}{K\mu\alpha} \\ &\xrightarrow{K \rightarrow \infty} \frac{\alpha LM}{\mu} \end{aligned} \quad (8.16b)$$

Demostración Recordemos 8.8a y si tomamos esperanza total e imponemos 8.15 tenemos:

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

Luego como por 8.2.2 tenemos que $F_{inf} \leq \mathbb{E}[F(w_k)]$ para todo $k \in \mathbb{N}$ vale:

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

■

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 [15].

En cambio, cuando $M \neq 0$ aunque no podemos acotar $\|\nabla F(w_k)\|_2^2$ *per-se*, podemos decir de 8.16b 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 (8.3.2) 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* 4.3:

Teorema 8.4.2 (Objetivo no convexo, Incrementos decrecientes) *Asumiendo 8.1.1 y 8.2.2, suponiendo además que en 8.1 los $\{\alpha_k\}$ satisfacen 4.3; si notamos $A_K := \sum_{k=1}^K \alpha_k$ vale para todo $k \in \mathbb{N}$:*

$$\lim_{K \rightarrow \infty} \mathbb{E} \left[\sum_{k=1}^K \alpha_k \|\nabla F(w_k)\|_2^2 \right] < \infty \quad (8.17a)$$

$$\mathbb{E} \left[\frac{1}{A_K} \sum_{k=1}^K \alpha_k \|\nabla F(w_k)\|_2^2 \right] \xrightarrow{K \rightarrow \infty} 0 \quad (8.17b)$$

Demostración Como $\alpha_k \rightarrow 0$ por 4.3 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 8.2.2 tenemos que $F_{inf} \leq \mathbb{E}[F(w_k)]$ para todo $k \in \mathbb{N}$ vale:

$$F_{inf} - \mathbb{E}[F(w_1)] \leq \mathbb{E}[F(w_{K+1})] - \mathbb{E}[F(w_1)] \leq -\frac{1}{2}\mu \sum_1^K \alpha_k \mathbb{E}[\|\nabla F(w_k)\|_2^2] + \frac{1}{2}LM \sum_{k=1}^K \alpha_k^2$$

Luego:

$$\sum_1^K \alpha_k \mathbb{E}[\|\nabla F(w_k)\|_2^2] \leq \frac{2(\mathbb{E}[F(w_1) - F_{inf}])}{\mu} + \frac{LM}{\mu} \underbrace{\sum_{k=1}^K \alpha_k^2}_{\xrightarrow{K \rightarrow \infty} C < \infty}$$

Por lo que 8.17a esta probado. Finalmente como por 4.3 tenemos que $A_K \rightarrow \infty$ podemos ver que existe $D > 0$ constante tal que:

$$\mathbb{E} \left[\frac{1}{A_K} \sum_{k=1}^K \alpha_k \|\nabla F(w_k)\|_2^2 \right] \leq \frac{D}{A_K} \xrightarrow{K \rightarrow \infty} 0$$

se tiene 8.17b. ■

Corolario 8.4.3 Asumiendo 8.1.1 y 8.2.2, suponiendo además que en 8.1 los $\{\alpha_k\}$ satisfacen 4.3 entonces :

$$\liminf_{k \rightarrow \infty} \mathbb{E}[\|\nabla F(w_k)\|_2^2] = 0 \quad (8.18)$$

Corolario 8.4.4 Bajo las mismas hipótesis de 8.4.2 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 8.17a y la desigualdad de Markov:

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

■

Teorema 8.4.5 (Objetivo no convexo regular, Incrementos decrecientes)

Bajo las mismas hipótesis de 8.4.2 si además pedimos que $F \in C^2$ y que $w \mapsto \|\nabla F(w)\|_2^2$ sea l -Lipshitz entonces:

$$\lim_{k \rightarrow \infty} \mathbb{E}[\|\nabla F(w_k)\|_2^2] = 0 \quad (8.19)$$

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

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

Si tomamos esperanza condicional a ξ_k y usamos 8.1.1, 8.2.2 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:

$$\mathbb{E} [G(w_{k+1})] - \mathbb{E} [G(w_k)] \leq 2\alpha_k L \mu_G \mathbb{E} [\|\nabla F(w_k)\|_2^2] + \frac{1}{2} \alpha_k^2 L_G (M + M_V \mathbb{E} [\|\nabla F(w_k)\|_2^2]) \quad (8.20)$$

Notemos que existe $K \in \mathbb{N}$ tal que $\alpha_k^2 \leq \alpha_k$ y luego por 8.4.2 el lado derecho cumple:

$$\lim_{N \rightarrow \infty} 2L\mu_G \underbrace{\sum_{k=K}^{K+N} \mathbb{E} [\alpha_k \|\nabla F(w_k)\|_2^2]}_{8.17a} + \frac{1}{2} L_G \left(M \underbrace{\sum_{k=K}^{K+N} \alpha_k^2}_{4.3} + M_V \underbrace{\sum_{k=K}^{K+N} \mathbb{E} [\alpha_k^2 \|\nabla F(w_k)\|_2^2]}_{8.17a} \right) = 0$$

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 8.20 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 8.4.3 tenemos $\mathbb{E} [\|\nabla F(w_k)\|_2^2] = \mathbb{E} [G(w_k)] \rightarrow 0$.

■

En un nivel puramente formal, uno podría decir que la teoría de probabilidades es el estudio de los espacios de medida con medida total 1, pero sería como decir que la teoría de números es el estudio de secuencias de dígitos que terminan.

Terence Tao

Ahora que ya analizamos la convergencia en L_1 de 8.1 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 4.3. Por simplicidad en los cálculos vamos a asumir en este capítulo que g es un estimador insesgado de $\nabla F(w_k)$.

9.1 CASO DÉBILMENTE CONVEXO

Hipótesis 9.1.1 (Acotaciones al segundo momento de g) Supongamos que dada F función objetivo cumple 4.0.1 y g la estimación insesgada del gradiente en 8.1 vale que existen $A, B \geq 0$ tales que:

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

Observación Notemos que si F cumple 8.1.1 y g cumple 8.2.2 entonces automáticamente cumplen 9.1.1.

Teorema 9.1.2 (Objetivo débilmente convexo, incrementos decrecientes)

Supongamos 4.0.1, 9.1.1; además supongamos que dado 8.1 α_k cumple 4.3:

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

$$w_k \xrightarrow[c.t.p.]{k \rightarrow \infty} w^* \quad (9.1a)$$

$$(w_k - w^*)^T \nabla F(w_k) \xrightarrow[c.t.p.]{k \rightarrow \infty} 0 \quad (9.1b)$$

Demostración Vayamos de a pasos como cuando demostramos 4.2.2.

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

Paso 2 Análogamente a casos anteriores notemos que:

$$h_{k+1} - h_k = -2\alpha_k (w_k - w^*)^T g(w_k, \xi_k) + \alpha_k^2 \|g(w_k, \xi_k)\|^2 \quad (9.2)$$

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 ξ_k ; luego si tomamos la esperanza condicional a esta filtración:

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

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

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

Ahora si incorporamos 9.1.1 obtenemos:

$$\mathbb{E}[h_{k+1} - (1 + \alpha_k^2 B) h_k | \mathcal{P}_k] \leq -2\alpha_k (w_k - w^*)^T \nabla F(w_k) + \alpha_k^2 A \quad (9.3)$$

Definamos ahora las sucesiones auxiliares:

$$\mu_k = \prod_{j=1}^{k-1} \frac{1}{1 + \alpha_j^2 B} \quad (9.4a)$$

$$h'_k = \mu_k h_k \quad (9.4b)$$

Por lo que replicando las operaciones en 4.2.2 llegamos a:

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

Luego, recordando 2.2.25 notemos que:

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

Por el mismo motivo que en 4.2.2 concluimos que:

$$h'_k \geq 0 \quad (9.5a)$$

$$\sum_{k=1}^{\infty} \mathbb{E} \left[\delta_k^{h'} (h'_{k+1} - h'_k) \right] < \infty \quad (9.5b)$$

Por 2.2.26 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 9.3 concluimos que:

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

Supongamos que $\mathbb{P} \left(\left\{ \lim_k h_k > 0 \right\} \right) > \tilde{\epsilon}$, luego $\mathbb{P} \left(\left\{ \alpha_k (w_k - w^*)^T \nabla F(w_k) > C\alpha_k \right\} \right) > \tilde{\epsilon}$ lo que implicaría por 4.3 que $\mathbb{P} \left(\left\{ \sum_{k=1}^{\infty} \alpha_k (w_k - w^*)^T \nabla F(w_k) = \infty \right\} \right) > \tilde{\epsilon}$; concluimos entonces que:

$$w_k \xrightarrow[\text{c.t.p.}]{k \rightarrow \infty} w^* \quad (9.6a)$$

$$(w_k - w^*)^T \nabla F(w_k) \xrightarrow[\text{c.t.p.}]{k \rightarrow \infty} 0 \quad (9.6b)$$

■

9.2 CASO NO CONVEXO

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

Hipótesis 9.2.1 (Hipótesis caso no convexo) Sea F una función de costo objetivo y supongamos que el algoritmo 8.1 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 8.1, entonces estos cumplen 4.3
5. Existe $B_2 \geq 0$ tal que:

$$\mathbb{E} \left[\|g(w_k, \xi_k)\|_2^2 \right] \leq B_2 \|w\|_2^2 \quad (9.7)$$

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

$$\mathbb{E} \left[\|g(w_k, \xi_k)\|_2^j \right] \leq A_j + B_j \|w\|_2^j \quad (9.8)$$

7. Existe $D > 0$ tal que:

$$\inf_{(w)^2 > D} w \nabla F(w) > 0 \quad (9.9)$$

9.2.1 Acotación global del algoritmo

Usemos 9.9 para probar que existe un entorno $w_1 \in U \subset \chi$ tal que si $\{w_k\}$ son las iteraciones del algoritmo 8.1 con F, g cumpliendo 9.2.1, entonces $\{w_k\} \subset U$.

Teorema 9.2.2 (Acotación global del algoritmo estocástico insesgado)

Sea F función de costo objetivo y g un estimador insesgado de ∇F tal que ambos cumplen 9.2.1, luego existe un entorno $w_1 \in U \subset \mathbb{R}^d$ tal que si $\{w_k\}$ son las iteraciones del algoritmo 8.1 entonces $\{w_k\} \subset U$.

Demostración Nuevamente probemos esto en tres pasos:

Paso 1 Sea D el parámetro de horizonte dado por 9.9 y definamos $\phi : \mathbb{R} \mapsto \mathbb{R}$ 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:

$$\phi(y) - \phi(x) \leq (y - x) \frac{d\phi}{dx}(x) + (y - x)^2$$

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 $\{f_k\}$:

$$\begin{aligned} f_{k+1} - f_k &\leq \left(-2\alpha_k w_k^T g(w_k, \xi_k) + \alpha_k^2 \|g(w_k, \xi_k)\|^2 \right) \frac{d\phi}{dx}(\|w_k\|^2) \\ &\quad + 4\alpha_k^2 (w_k^T g(w_k, \xi_k))^2 - 4\alpha_k^3 w_k^T g(w_k, \xi_k) \|g(w_k, \xi_k)\|^2 \\ &\quad + 4\alpha_k^4 \|g(w_k, \xi_k)\|^4 \end{aligned} \quad (9.10)$$

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

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

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

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

Si ahora incluimos 9.8 y notamos que $\frac{d\phi}{dx}(\|w_k\|^2) \leq \|w_k\|^2$ entonces:

$$\begin{aligned} \mathbb{E}[f_{k+1} - f_k | \mathcal{P}_k] &\leq -2\alpha_k w_k^T \nabla F(w_k) \frac{d\phi}{dx}(\|w_k\|^2) \\ &\quad + \alpha_k^2 \frac{d\phi}{dx}(\|w_k\|^2) B_2 \|w_k\|^2 \\ &\quad + 4\alpha_k^2 B_2 \|w_k\|^2 \|w_k\|^2 \\ &\quad - \underbrace{4\alpha_k^3 \|w_k\| \mathbb{E}[\|g(w_k, \xi_k)\|^3]}_{\leq 0} \\ &\quad + 4\alpha_k^4 (A_4 + B_4 \|w_k\|^4) \\ &\leq -2\alpha_k w_k^T \nabla F(w_k) \frac{d\phi}{dx}(\|w_k\|^2) \\ &\quad + \alpha_k^2 \left(5B_2 \|w_k\|^4 + 4\alpha_k^2 B_4 \|w_k\|^4 + 4 + 4 \underbrace{\alpha_k^2}_{\leq \tilde{A}_0 \text{ por 4.3}} A_4 \right) \end{aligned}$$

Luego existen $A_0, B_0 \geq 0$ constantes tal que:

$$\mathbb{E}[f_{k+1} - f_k | \mathcal{P}_k] \leq -2\alpha_k w_k^T \nabla F(w_k) \frac{d\phi}{dx}(\|w_k\|^2) + \alpha_k^2 (A_0 + B_0 \|w_k\|^4)$$

Lo que implica que existen $A, B \geq 0$ tal que:

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

Notemos ahora que si $\|w_k\|^2 < D$ entonces $\frac{d\phi}{dx}(\|w_k\|^2) = 0$, y si $\|w_k\|^2 \geq D$ entonces $-2\alpha_k w_k^T \nabla F(w_k) \frac{d\phi}{dx}(\|w_k\|^2) < 0$ por 9.9 por lo que deducimos:

$$\mathbb{E} [f_{k+1} - f_k | \mathcal{P}_k] \leq \alpha_k^2 (A + Bf_k) \quad (9.11)$$

Ahora siguiendo los mismo pasos que al demostrar 9.1.2 definiendo μ_k, f'_k y usando 2.2.26 concluimos que $\{f_k\}$ converge casi todo punto.

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 9.10 es una igualdad y deducimos que:

$$\sum_{k=1}^{\infty} \alpha_k w_k^T \nabla F(w_k) \frac{d\phi}{dx}(\|w_k\|^2) < \infty \quad ctp \quad (9.12)$$

Pero por otro lado como $f_{inf} > 0$, $\sum_{k=1}^{\infty} \alpha_k = \infty$ y $0 < \lim_{k \rightarrow \infty} \frac{d\phi}{dx}(\|w_k\|^2)$ existe $\tilde{\epsilon} > 0$ y $M > 0$ tal que:

$$\begin{aligned} \tilde{\epsilon} &< \mathbb{P} \left(\left\{ \sum_{k=1}^{\infty} \alpha_k w_k^T \nabla F(w_k) \frac{d\phi}{dx}(\|w_k\|^2) = \infty \right\} \right) \\ &= \mathbb{P} \left(\left\{ \sum_{k=1}^{\infty} \alpha_k w_k^T \nabla F(w_k) \frac{d\phi}{dx}(\|w_k\|^2) > M \liminf \frac{d\phi}{dx}(\|w_k\|^2) \sum_{k=1}^{\infty} \alpha_k \right\} \right) \end{aligned}$$

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

9.2.2 Convergencia del algoritmo

Teorema 9.2.3 (Convergencia a puntos extremales, Caso insesgado)

Sea F función de costo objetivo y g un estimador insesgado de ∇F tal que ambos cumplen 9.2.1, si $\{w_k\}$ son las iteraciones del algoritmo 8.1 entonces valen:

$$F(w_k) \xrightarrow[c.t.p.]{k \rightarrow \infty} F_{\infty} \quad (9.13a)$$

$$\nabla F(w_k) \xrightarrow[c.t.p.]{k \rightarrow \infty} 0 \quad (9.13b)$$

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

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

Paso 2 Por 9.2.2 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:

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

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

$$\mathbb{E} [h_{k+1} - h_k | \mathcal{P}_k] \leq \underbrace{-2\alpha_k \|\nabla F(w_k)\|^2}_{\leq 0} + \alpha_k^2 K_1 \underbrace{\mathbb{E} [g(w_k, \xi_k)]}_{\leq K_2 \text{ por 9.2.2}} \quad (9.14)$$

Lo que implica:

$$\mathbb{E} [h_{k+1} - h_k | \mathcal{P}_k] \leq \alpha_k^2 K_1 K_2 \quad (9.15)$$

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 2.2.26 obtenemos:

$$F(w_k) \xrightarrow[\text{c.t.p.}]{k \rightarrow \infty} F_{\infty} \quad (9.16)$$

Paso 3 Si retomamos 9.14, reordenamos, sumamos hasta k y tomamos esperanza tenemos:

$$\begin{aligned}
2 \sum_{k=1}^K \alpha_k \|\nabla F(w_k)\|^2 &\leq \sum_{k=1}^K \mathbb{E} [h_{k+1} - h_k] + K_2 K_1 \sum_{k=1}^K \alpha_k^2 \\
&= \mathbb{E} \left[\sum_{k=1}^K h_{k+1} - h_k \right] + K_2 K_1 \sum_{k=1}^K \alpha_k^2 \\
&= \mathbb{E} [h_{K+1}] + K_2 K_1 \sum_{k=1}^K \alpha_k^2 \\
&\xrightarrow[\text{2.2.20}]{\text{ctp}} F_\infty + K_2 K_1 \sum_{k=1}^\infty \alpha_k^2
\end{aligned} \tag{9.17}$$

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

$$\begin{aligned}
g_{k+1} - g_k &= \nabla g(w_k) (w_{k+1} - w_k) + \frac{1}{2} \nabla^2 g(w_k) (w_{k+1} - w_k)^2 \\
&\leq \underbrace{-2\alpha_k \nabla F(w_k)^T K_4 g(w_k, \xi_k) + K_3 \alpha_k^2 \|g(w_k, \xi_k)\|^2}_{\text{por 9.2.2}}
\end{aligned}$$

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

$$\mathbb{E} [g_{k+1} - g_k | \mathcal{P}_k] \leq 2\alpha_k K_4 \|\nabla F(w_k)\|^2 + K_2 K_3 \alpha_k^2 \tag{9.18}$$

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

$$\begin{aligned}
\sum_{k=1}^\infty \mathbb{E} [\delta_k^g (g_{k+1} - g_k)] &= \sum_{k=1}^\infty \mathbb{E} [\delta_k^g \mathbb{E} [(g_{k+1} - g_k) | \mathcal{P}_k]] \\
&\leq \underbrace{K_4 \sum_{k=1}^\infty \alpha_k \|\nabla F(w_k)\|^2}_{9.17} + \underbrace{K_2 K_3 \sum_{k=1}^\infty \alpha_k^2}_{4.3} \\
&< \infty
\end{aligned} \tag{9.19}$$

Nuevamente por 2.2.26 concluimos que $\{g_k\}$ converge *ctp*, y por 9.17 este límite es 0; luego:

$$g_k \xrightarrow[\text{c.t.p.}]{k \rightarrow \infty} 0 \tag{9.20a}$$

$$\nabla F(w_k) \xrightarrow[\text{c.t.p.}]{k \rightarrow \infty} 0 \tag{9.20b}$$

■

Parte IV

Epílogo

CONCLUSIONES

A modo de conclusión observemos que el estudio de algoritmos estocásticos está viviendo un nuevo auge en su análisis motivado por los siguientes puntos:

1. Los métodos determinísticos sufren la lentitud natural de tener un solo *epoch* por set de datos.
2. El descenso de gradiente tiene complejidad exponencial para alcanzar el mínimo.
3. Los métodos estocásticos bajo condiciones de convexidad muy leves convergen rápidamente en esperanza y *casi todo punto*.
4. Esa convergencia es independiente de n , lo que los habilita para el régimen del *big data*.
5. Los algoritmos determinísticos tradicionales están perdiendo preferencia por sus propiedades teóricas desfavorables.

Además en [8] y [29] se observan las siguientes propiedades muy interesantes:

1. El algoritmo de PGD (una variante de GD con ruido isométrico) escapa puntos silla en tiempo polinomial.
2. El algoritmo SGLD (otra variante de GD) escapa mínimos locales con valor absoluto grande.

Quedan como posibles líneas futuras de investigación:

- Analizar qué tan restrictivas son las condiciones que pedimos en esta Tesis a la función objetivo, así como algoritmos que permitan la validación de las hipótesis con la menor cantidad de derivadas a calcular posibles.
- Profundizar el estudio de la complejidad y convergencia a mínimos en los algoritmos mixtos que surgen de intentar juntar características de los algoritmos más usuales.
- Estudiar si SG, u otros DE, cumplen las propiedades vistas en [8] y [29]; es de particular interés un algoritmo que cumpla ambas y además no sea costoso computacionalmente o admita un cálculo distribuido.

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 [21]

Teorema A.1.2 *Sea $E \subset \mathbb{R}^d$ un compacto y $f \in C^m(E)$. Si dotamos a $C^m(E)$ de la norma $\|f\|_{C^m} = \sup \{\|\partial^\alpha f\|_\infty : |\alpha| \leq m\}$ entonces existe $T \in L(C^m(E), C^m(\mathbb{R}^d))$ extensión de f 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 4.2.1] 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. ■

Demostración [De 4.3.2]

1. $[1 \iff 2]$ Notemos que g es convexa si y sólo si:

$$\begin{aligned} g(y) &\geq g(x) + \nabla g(x)^T (y - x) \\ \iff f(y) - \frac{\mu}{2} \|y\|^2 &\geq f(x) - \frac{\mu}{2} \|x\|^2 + (\nabla f(x) - \mu x)^T (y - x) \\ \iff f(y) &\geq f(x) + \nabla f(x)^T (y - x) - \mu \left(\frac{\|x\|^2 - \|y\|^2}{2} + x^T (y - x) \right) \\ \iff f(y) &\geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2} (\|y\|^2 + \|x\|^2) - \mu x^T y \\ \iff f(y) &\geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2} \|y - x\|^2 \end{aligned}$$

2. $[2 \iff 3]$ Notemos que g es convexa si y sólo si:

$$\begin{aligned} (\nabla g(x) - \nabla g(y))^T (x - y) &\geq 0 \\ \iff (\nabla f(x) - \mu x - [\nabla f(y) - \mu y])^T (x - y) &\geq 0 \\ \iff (\nabla f(x) - \nabla f(y) - \mu [x - y])^T (x - y) &\geq 0 \\ \iff (\nabla f(x) - \nabla f(y))^T (x - y) - \mu (x - y)^T (x - y) &\geq 0 \\ \iff (\nabla f(x) - \nabla f(y))^T (x - y) &\geq \mu \|x - y\|^2 \end{aligned}$$

3. $[2 \iff 4]$ Notemos que g es convexa si y sólo si:

$$\begin{aligned} g(\alpha x + (1 - \alpha)y) &\leq \alpha g(x) + (1 - \alpha)g(y) \\ \iff f(\alpha x + (1 - \alpha)y) &\leq \frac{\mu}{2} \|\alpha x + (1 - \alpha)y\|^2 + \alpha \left(f(x) - \frac{\mu}{2} \|x\|^2 \right) + \\ &\quad (1 - \alpha) \left(f(y) - \frac{\mu}{2} \|y\|^2 \right) \\ \iff f(\alpha x + (1 - \alpha)y) &\leq [\alpha f(x) + (1 - \alpha)f(y)] \\ &\quad + \frac{\mu}{2} \left(\|\alpha x + (1 - \alpha)y\|^2 \right. \\ &\quad \left. - \left\{ \alpha \|x\|^2 + (1 - \alpha)\|y\|^2 \right\} \right) \\ \iff f(\alpha x + (1 - \alpha)y) &\leq \alpha f(x) + (1 - \alpha)f(y) - \frac{\alpha(1 - \alpha)\mu}{2} \|y - x\|^2 \end{aligned}$$

■

Demostración [De 4.3.4]

1. Dado $x \in \mathbb{R}^d$ sea:

$$q(z) = f(x) + \nabla f(x)^T (z - x) + \frac{1}{2} \mu \|z - x\|_2^2$$

Se puede verificar que $z_* := x - \frac{1}{\mu} \nabla f(x)$ cumple que $q(z_*) = f(x) - \frac{1}{2\mu} \|\nabla f(x)\|_2^2 \leq q(z)$ para todo $z \in \mathbb{R}^d$; luego por 4.3.1 se tiene:

$$f_{\inf} \geq f(x) + \nabla f(x)^T (z_* - x) + \frac{1}{2} \mu \|z_* - x\|_2^2 \geq f(x) - \frac{1}{2\mu} \|\nabla f(x)\|_2^2$$

2. Por Cauchy-Schwartz y usando 4.3.2:

$$\|\nabla f(x) - \nabla f(y)\| \|x - y\| \geq (\nabla f(x) - \nabla f(y))^T (x - y) \geq \mu \|x - y\|^2$$

3. Consideremos $\phi_x(z) = f(z) - \nabla f(x)^T z$, luego notemos que es μ -fuertemente convexa; en efecto:

$$(\nabla \phi_x(z_1) - \nabla \phi_x(z_2))^T (z_1 - z_2) = (\nabla f(z_1) - \nabla f(z_2))^T (z_1 - z_2) \geq \mu \|z_1 - z_2\|^2$$

Luego si aplicamos la desigualdad PL a $\phi_x(z)$ evaluada en $z^* = x$ entonces:

$$\begin{aligned} (f(y) - \nabla f(x)^T y) - (f(x) - \nabla f(x)^T x) &= \phi_x(y) - \phi_x(x) \\ &\leq \frac{1}{2\mu} \|\nabla \phi_x(y)\|^2 \\ &= \frac{1}{2\mu} \|\nabla f(y) - \nabla f(x)\|^2 \end{aligned}$$

4. Notemos que por el punto anterior:

$$\begin{aligned} f(y) &\leq f(x) + \nabla f(x)^T (y - x) + \frac{1}{2\mu} \|\nabla f(y) - \nabla f(x)\|^2 \\ f(x) &\leq f(y) + \nabla f(y)^T (x - y) + \frac{1}{2\mu} \|\nabla f(x) - \nabla f(y)\|^2 \end{aligned}$$

Luego si las sumamos y reordenamos:

$$(\nabla f(x) - \nabla f(y))^T (x - y) \leq \frac{1}{\mu} \|\nabla f(x) - \nabla f(y)\|^2$$

Demostración [De 4.3.8]

1. $[2 \iff 3]$ Notemos que g es convexa si y sólo si:

$$\begin{aligned}
 g(y) &\geq g(x) + \nabla g(x)^T (y - x) \\
 \iff \frac{L}{2} \|y\|^2 - f(y) &\geq \frac{L}{2} \|x\|^2 - f(x) + (Lx - \nabla f(x))^T (y - x) \\
 \iff \frac{L}{2} \|y\|^2 - f(y) &\geq \frac{L}{2} \|x\|^2 - f(x) + Lx^T y - L \|x\|^2 - \nabla f(x)^T (y - x) \\
 \iff f(y) &\leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} (\|y\|^2 + \|x\|^2 - 2x^T y) \\
 \iff f(y) &\leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} \|y - x\|^2
 \end{aligned}$$

2. $[2 \iff 4]$ Sale igual que $2 \iff 3$ de 4.3.2

3. $[2 \iff 5]$ Sale igual que $2 \iff 4$ de 4.3.2

4. $[1 \implies 4]$

$$(\nabla f(x) - \nabla f(y))^T (x - y) \leq \|\nabla f(x) - \nabla f(y)\| \|x - y\| \leq L \|x - y\|^2$$

5. $[7 \implies 1]$

$$\begin{aligned}
 \|\nabla f(x) - \nabla f(y)\|^2 &\leq L (\nabla f(x) - \nabla f(y))^T (x - y) \\
 &\leq L \|\nabla f(x) - \nabla f(y)\| \|x - y\|
 \end{aligned}$$

6. $[8 \implies 6]$ Notemos que si cambiamos de lugar x, y en 8 entonces:

$$\begin{aligned}
 f(y) &\geq f(x) + \frac{f(x + \alpha(y - x)) - f(x)}{\alpha} + \frac{1 - \alpha}{2L} \|\nabla f(x) - \nabla f(y)\|^2 \\
 \downarrow \alpha \rightarrow 0 \quad \downarrow \\
 f(y) &\geq f(x) + \nabla f(x)^T (y - x) + \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|^2
 \end{aligned}$$

7. $[6 \implies 8]$ Sea $z = \alpha x + (1 - \alpha)y$, luego:

$$\begin{aligned}
 f(y) &\geq f(z) + \nabla f(z)^T (y - z) + \frac{1}{2L} \|\nabla f(z) - \nabla f(y)\|^2 \\
 f(x) &\geq f(z) + \nabla f(z)^T (x - z) + \frac{1}{2L} \|\nabla f(z) - \nabla f(x)\|^2
 \end{aligned}$$

Luego si multiplicamos la primera por α , la segunda por $1 - \alpha$ y las sumamos tenemos:

$$\begin{aligned}
 f(\alpha x + (1 - \alpha)y) &\leq \alpha f(x) + (1 - \alpha)f(y) - \frac{\alpha}{2L} \|\nabla f(z) - \nabla f(y)\|^2 \\
 &\quad - \frac{1 - \alpha}{2L} \|\nabla f(z) - \nabla f(x)\|^2 \\
 &\leq \alpha f(x) + (1 - \alpha)f(y) - \frac{\alpha(1 - \alpha)}{2L} \|\nabla f(y) - \nabla f(x)\|^2
 \end{aligned}$$

Pues $\|x - y\|^2 \alpha(1 - \alpha) \leq \alpha \|x\|^2 + (1 - \alpha) \|y\|^2$

8. [2 \implies 6] Si f es convexa entonces consideremos $\phi_x(z) = f(z) - \nabla f(x)^T z$ que es mínimo en $z^* = x$ pues f es convexa. Además por hipótesis $h(z) = \frac{L}{2} \|z\|^2 - \phi_x(z)$ es convexa por lo que:

$$\min_{z \in \mathbb{R}^d} \phi_x(z) \leq \min_{z \in \mathbb{R}^d} \left\{ \phi_x(y) + \nabla \phi_x(y)^T (z - y) + \frac{L}{2} \|z - y\|^2 \right\}$$

Luego reordenando:

$$\begin{aligned} f(y) - f(x) - \nabla f(x)^T (y - x) &= \phi_x(y) - \phi_x(x) \\ &\geq \frac{1}{2L} \|\nabla \phi_x(y)\|^2 \\ &= \frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|^2 \end{aligned}$$

■

Demostración [De 6.3.7] Manteniendo la notación previa a la afirmación:

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

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

$$z_{i+1}^T H z_{i+1} < z_i^T H z_i - \alpha \left\| \sum_{j \in S_i} (e_j^T H z_i) e_j \right\|_2^2 \quad (\text{A.4})$$

Por otro lado sabemos que para todo w vale:

$$w^T H w \geq \lambda_{\min}(H) \|w\|_2^2 \geq -L_b \|w\|_2^2 \quad (\text{A.5})$$

Luego si usamos 6.3.6, A.5 y Cauchy-Schwartz existe $i \in \{1, \dots, b\}$ y $\delta > 0$ tal que:

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

Tomando $\epsilon = \frac{\delta^2}{d\alpha L_b}$ 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:

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

■

Demostración [De 7.3.1]

Primero construyamos g_1 . Sea $i \in \{1, \dots, d\}$, luego si $x_i = \tau$ entonces $\frac{\partial f_{i,1}}{\partial x_i} = -2\gamma\tau$, $\frac{\partial^2 f_{i,1}}{\partial x_i^2} = -2\gamma$; mientras que si $x_i = 2\tau$ entonces $\frac{\partial f_{i+1,1}}{\partial x_i} = -4L\tau$ y $\frac{\partial^2 f_{i+1,1}}{\partial x_i^2} = 2L$. Como sabemos que $L > \gamma, 0 > -2\gamma\tau > -4L\tau$ y $2L > \frac{-4L\tau - (-2\gamma\tau)}{\tau}$ entonces por A.1.1 existe $p(x_i) \in \mathbb{R}_3[x]$ tal que:

$$\begin{aligned} p(\tau) &= -2\gamma\tau \\ p(2\tau) &= -4L\tau \\ p'(\tau) &= -2\gamma \\ p'(2\tau) &= 2L \\ p|_{[\tau, 2\tau]} &\leq -2\gamma\tau \end{aligned}$$

Definamos entonces:

$$\begin{aligned}
g_1(x_i) &= (\int p)(x_i) - (\int p)(\tau) - \gamma\tau^2 \\
\eta &= -g_1(2\tau) + 4L\tau^2 \\
g_2(x_i) &= -\gamma - \frac{10(L+\gamma)(x_i-2\tau)^3}{\tau^3} \\
&\quad - \frac{15(L+\gamma)(x_i-2\tau)^4}{\tau^4} - \frac{6(L+\gamma)(x_i-2\tau)^5}{\tau^5}
\end{aligned}$$

Luego notemos que g_1, η cumplen las condiciones de borde para x_i automáticamente; por otro lado si consideramos x_{i+1} notemos que:

$$g_2'(x_i) = -\frac{30(L+\gamma)(x_i-2\tau)^2(x_i-\tau)^2}{\tau^5}$$

Luego:

$$\begin{aligned}
g_2|_{[\tau, 2\tau]} &\geq -\gamma \\
g_2'|_{[\tau, 2\tau]} &\leq 0 \\
g_2(\tau) &= L \\
g_2(2\tau) &= -\gamma \\
g_2'(\tau) &= 0 \\
g_2'(2\tau) &= 0 \\
g_2''(\tau) &= 0 \\
g_2''(2\tau) &= 0
\end{aligned}$$

Luego g_2 cumple las condiciones de contorno en x_{i+1} y además al tener derivadas cero en los límites no contribuye a las condiciones de x_i . Concluimos que con estas definiciones g cumple lo pedido . ■

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] Wassily Hoeffding. *PROBABILITY INEQUALITIES FOR SUMS OF BOUNDED RANDOM VARIABLES*. 1962.
 - [7] G. Piliouras M. Simchowitz M.I. Jordan "J. Lee I. Panageas and B. Recht". "First-order methods almost always avoid saddle points." In: (2017).
 - [8] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, and Michael I. Jordan. "How to Escape Saddle Points Efficiently." In: *ICML*. 2017.
 - [9] John M. Lee. "Introduction to Smooth Manifolds." In: (2000).
 - [10] D. C. Liu and J. Nocedal. "On the Limited Memory BFGS Method for Large Scale Optimization." In: *Math. Program.* 45.3 (Dec. 1989), pp. 503–528. ISSN: 0025-5610. DOI: [10.1007/BF01589116](https://doi.org/10.1007/BF01589116). URL: <http://dx.doi.org/10.1007/BF01589116>.
 - [11] S. Łojasiewicz. *A topological property of real analytic subsets*. Coll. du CNRS, Les equations aux derivees partielles, 1963, pp. 87–89.
 - [12] M. Metivier. *Semi-Martingales*. Walter de Gruyter, 1983.
 - [13] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2012. ISBN: 026201825X, 9780262018258.
 - [14] Elad Hazan Naman Agarwal Brian Bullins. "Second-Order Stochastic Optimization for Machine Learning in Linear Time." In: (2017).
 - [15] Yurii Nesterov. *Introductory Lectures on Convex Optimization*. Vol. 87. Springer Science & Business Media, 2004.

- [16] Jorge Nocedal. "Updating Quasi-Newton Matrices with Limited Storage." English. In: *Mathematics of Computation* 35 (July 1980), pp. 773–782. DOI: [10.1090/S0025-5718-1980-0572855-7](https://doi.org/10.1090/S0025-5718-1980-0572855-7).
- [17] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. second. New York, NY, USA: Springer, 2006.
- [18] James M. Ortega and Werner C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2000. ISBN: 0-89871-461-3.
- [19] B. T. Polyak. *Gradient methods for minimizing functionals*. Zh. Vychisl. Mat. Fiz, 1963, pp. 643–653.
- [20] K. MURALI RAO. "QUASI-MARTINGALES." In: *Mathematica Scandinavica* 24.1 (1969), pp. 79–92. ISSN: 00255521, 19031807. URL: <http://www.jstor.org/stable/24489871>.
- [21] 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.
- [22] H. Robbins and S. Monro. *A Stochastic Approximation Model*. Vol. 22(3). *The Annal of Mathematical Statistics*, 1951, pp. 400–407.
- [23] Michael Schub. *Global Stability of Dynamical Systems*. Springer Science and Business media, 1987.
- [24] V. N. Vapnik and A. Ya. Chervonenkis. "On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities." In: *Theory of Probability and its Applications* 16.2 (1971), pp. 264–280. DOI: [10.1137/1116025](https://doi.org/10.1137/1116025). URL: <http://link.aip.org/link/?TPR/16/264/1>.
- [25] V. N. Vapnik and A. Ya. Chervonenkis. *Theory of Pattern Recognition [in Russian]*. USSR: Nauka, 1974.
- [26] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.
- [27] David Williams. *Probability with Martingales*. Cambridge University Press, 1991. DOI: [10.1017/CB09780511813658](https://doi.org/10.1017/CB09780511813658).
- [28] Gersende Fort Yves F Atchade and Eric Moulines. "On stochastic proximal gradient algorithms." In: (2014).
- [29] Yuchen Zhang, Percy Liang, and Moses Charikar. "A Hitting Time Analysis of Stochastic Gradient Langevin Dynamics." In: *COLT*. 2017.
- [30] Wilson et.al. "The Marginal Value of Adaptive Gradient Methods in Machine Learning." In: (2017).