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
Septiembre 2018 – version 1.0

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 que condiciones convergen los métodos de primer orden usuales?** Para ello utilizamos herramientas de sistemás dinámicos y procesos estocásticos para 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; mientras que en el caso estocástico la misma naturaleza de este garantiza una complejidad uniforme en ambos casos, aun en el rango del *big data*. Sumado a esto dimos motivos tanto teóricos como prácticos para la preferencia de los algoritmos estocásticos referido a la velocidad de convergencia a entornos de la solución en el caso general.

Queda como posible línea futura analizar que tan restrictivas son estas condiciones asi como profundizár el estudio de la complejidad y convergencia a mínimos en los algoritmos mixtos que surgen de intentar juntar características de los algoritmos más usuales.

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

— Donald E. Knuth [8]

AGRADECIMIENTOS

Agradecimientos para todos

CONTENTS

Part I

Introducción

1

INTRODUCCIÓN

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

1.1 INTRODUCTION

La promesa de la inteligencia artificial ha sido un tema de interés público y privado durante décadas. A partir de la década de 1950, hubo grandes esperanzas de que las técnicas clásicas de inteligencia artificial basadas en lógica, representación del conocimiento, razonamiento y planificación daría lugar a un software revolucionario que podría, entre otras cosas, comprender el lenguaje, controlar robots y proporcionar asesoramiento experto. Aunque los avances basados en tales técnicas pueden estar al alcance en el futuro, muchos comenzaron a dudar de estos enfoques clásicos, optando por enfocar sus esfuerzos en el diseño de sistemas basados en técnicas estadísticas, como en la rápida evolución y expansión del campo del *machine learning*.

El machine learning y los sistemas inteligentes que han surgido de él, como los motores de búsqueda, las plataformas de recomendación y el software de reconocimiento de voz e imagen, se han convertido en una parte indispensable de la sociedad moderna. Enraizados en las estadísticas y basados en gran medida en la eficacia de los algoritmos numéricos, las técnicas de machine learning capitalizan las plataformas informáticas cada vez más potentes del mundo y la disponibilidad de conjuntos de datos de gran tamaño. Además, dado que los frutos de sus esfuerzos se han vuelto tan fácilmente accesibles para el público a través de diversas modalidades, como *la nube*, el interés en el machine learning continuará su aumento dramático, produciendo impactos sociales, económicos y científicos.

Uno de los pilares del machine learning es la *optimización matemática*, que, en este contexto, implica el cálculo numérico de parámetros para un sistema diseñado para tomar decisiones basadas en datos aún no vistos. Es decir, de acuerdo con los datos actualmente disponibles, estos parámetros se eligen para ser óptimos con respecto a un problema de aprendizaje dado. El éxito de ciertos métodos de optimización para el machine learning ha inspirado a grandes números de científicos en diversas comunidades de investigación para abordar problemas de aprendizaje automático aún más desafiantes, y para diseñar nuevos métodos que sean más ampliamente aplicables.

Mientras que los métodos tradicionales basados en gradiente pueden ser efectivos para resolver problemas de aprendizaje a pequeña escala en los que se puede usar un enfoque por *batch*, en el contexto del machine learning a gran escala la estrategia central de interés ha sido un algoritmo estocástico: el método del *desenso estocástico del gradiente* (DE) propuesto por Robbins y Monro [22]. Debido a este papel central desempeñado por DE, discutimos sus propiedades teóricas y prácticas fundamentales en unos pocos contextos de interés. 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.

1.2 PROCEDIMIENTO FORMAL DE MACHINE LEARNING.

Un proceso de machine learning estándar conlleva la selección de una función de predicción h mediante la resolución de un problema de optimización. Continuando con nuestro trabajo, es necesario formalizar nuestra presentación discutiendo en mayor detalle los principios detrás del proceso de selección, enfatizando la importancia teórica de la *ley de los grandes números* así como la importancia práctica de la *minimización del riesgo estructural*.

Para simplificar, continuamos enfocándonos en los problemas que surgen en el contexto de la *clasificación supervisada*; es decir, nos enfocamos en la optimización de las funciones de predicción para etiquetar datos no vistos en base a la información contenida en un conjunto de datos de entrenamiento etiquetados. Tal enfoque es razonable ya que muchas técnicas de aprendizaje no supervisadas y otras técnicas de aprendizaje se reducen a problemas de optimización de forma comparable; ver, por ejemplo, [27].

1.2.1 *Fundamentos*

Nuestro objetivo es determinar una función de predicción $h : \mathcal{X} \to \mathcal{Y}$ desde un espacio de entrada $\mathcal{X}$ a un espacio de salida $\mathcal{Y}$ tal que, dado $x \in \mathcal{X}$, el valor $h(x)$ ofrece una predicción precisa sobre el valor verdadero de salida y. Es decir, 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. 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 $P(x, y)$ que simultaneamente representa la distribución $P(x)$ de entradas así como la probabilidad condicionada $P(y|x)$ del valor y para un dado valor de entrada x. (Desde este punto de vista, uno a menudo se refiere

a los ejemplos como *muestras*, usaremos ambos términos en el resto del trabajo.) En lugar de algo que simplemente minimiza el riesgo empírico (1.1), 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

$$R_n(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left[h(x_i) \neq y\right], \quad \mathbb{1}\left[A\right] = \begin{cases} 1 & \text{si A es verdadero} \\ 0 & \text{sino} \end{cases} \tag{1.1}$$

$$R(h) = \mathbb{P}\left[h(x) \neq y\right] = \mathbb{E}\left[\mathbb{1}\left[h(x) \neq y\right]\right] \tag{1.2}$$

donde $\mathbb{P}\left[A\right]$ y $\mathbb{E}\left[A\right]$ respectivamente denotan la probabilidad y la esperanza de A. Tal contexto es *variacional* ya que estamos optimizando sobre un conjunto de funciones, y es *estocástico* ya que la función objetivo implica una esperanza.

Si bien se puede desear minimizar el riesgo esperado (1.2), en la práctica uno debe intentar hacerlo sin un conocimiento explícito de P. En cambio, la única opción manejable es construir un problema sustituto que dependa únicamente de los ejemplos $\{(x_i, y_i)\}_{i=1}^{n}$. En general, hay 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.

1.2.2 *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 que puedan lograr un riesgo empírico bajo sobre el conjunto de entrenamiento, 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 en 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 uno usa familias de funciones más ricas (ver a continuación). Este último hecho pone al segundo objetivo en desacuerdo con el primero. 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 al recordar la *ley de los grandes números*. Por ejemplo, cuando el riesgo esperado representa una probabilidad de

clasificación errónea como en (1.2), la desigualdad Hoeffding [6] garantiza que, con probabilidad de al menos $1 - \eta$, uno tiene

$$|R(h) - R_n(h)| \leq \sqrt{\frac{1}{2n} \log\left(\frac{2}{\eta}\right)} \quad \text{para un dado } h \in \mathcal{H} \tag{1.3}$$

Este límite ofrece la explicación intuitiva de que la brecha disminuye a medida que uno usa 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! Más bien, h es la variable sobre la cual uno 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]. Para la intuición detrás de este concepto, considere, por ejemplo, un esquema de clasificación binario en $\mathbb{R}^2$ donde uno asigna una etiqueta de 1 para puntos sobre un polinomio y -1 para puntos debajo. El conjunto de polinomios lineales tiene una baja capacidad en el sentido de que solo es capaz de clasificar con precisión los puntos de entrenamiento que pueden separarse por una línea; por ejemplo, en dos variables, un clasificador lineal tiene una dimensión VC de tres. Un conjunto de polinomios de alto grado, por otro lado, tiene una gran capacidad, ya que puede separar con precisión los puntos de entrenamiento que se intercalan; la dimensión VC de un polinomio de grado D en d variables es:

$$\begin{pmatrix} d+D \\ d \end{pmatrix}$$

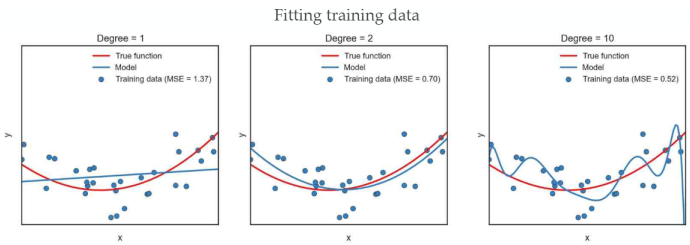

Figure 1: Fenómeno de sobreajuste en polinomios de alto grado

Dicho esto, la brecha entre el riesgo empírico y el riesgo esperado puede ser mayor para un conjunto de polinomios de alto grado ya que su alta capacidad les permite sobreajustar un conjunto dado de datos de entrenamiento. (Ver **??**)

Matemáticamente, con la capacidad de medición de la dimensión VC, se puede establecer uno de los resultados más importantes en Machine Learning

Proposición 1.2.1 (Cota en la complejidad algoritmica de la estimación riesgo) *Sea $d_{\mathcal{H}}$ la dimensión VC de $\mathcal{H}$ una familia de funciones generalizadora, luego se tiene una probabilidad de al menos $1 - \eta$ que:*

$$\sup_{h\in\mathcal{H}} |R(h) - R_n(h)| \leq \mathcal{O}\left(\sqrt{\frac{1}{2n}\log\left(\frac{2}{\eta}\right) + \frac{d_{\mathcal{H}}}{n}\log\left(\frac{n}{d_{\mathcal{H}}}\right)}\right) \tag{1.4}$$

Este límite proporciona una imagen más precisa de la dependencia de la brecha en la elección de $\mathcal{H}$. Por ejemplo, muestra que para una $d_{\mathcal{H}}$ fija, se obtiene una convergencia uniforme aumentando el número de puntos de entrenamiento n. Sin embargo, también muestra que, para una n fija, la brecha puede ensancharse si aumento $d_{\mathcal{H}}$. De hecho, para mantener la misma brecha, se debe aumentar n a la misma tasa si $d_{\mathcal{H}}$ se incrementa. La convergencia uniforme incorporada en este resultado es crucial en el machine learning, ya que uno quiere asegurarse de que el sistema de predicción funciona bien con cualquier dato que se le proporcione.

Curiosamente, una cantidad que no ingresa en (1.4) es el número de parámetros que distinguen a una función miembro particular h de la familia $\mathcal{H}$. En algunos entornos, como la regresión logística, este número es esencialmente el mismo que $d_{\mathcal{H}}$, lo que podría sugerir que la tarea de optimizar sobre $h \in \mathcal{H}$ es más engorrosa a medida que $d_{\mathcal{H}}$ aumenta. Sin embargo, este no es siempre el caso. Ciertas familias de funciones son más amenas para minimizar a pesar de tener un número muy grande o incluso infinito de parámetros; en [26] se diseñaron para aprovechar este hecho [Ver Teorema 10.3].

En general, mientras que los límites tales como (1.4) son teóricamente interesantes y proporcionan una visión útil, raramente se usan directamente en la práctica ya que generalmente es más fácil estimar la brecha entre riesgo empírico y riesgo esperado con experimentos de *validación cruzada*. Ahora presentaremos ideas subyacentes a un marco práctico que respeta las concesiones mencionadas anteriormente.

1.2.3 *Minimización de riesgos estructurales*

Un enfoque para elegir una función de predicción que ha demostrado ser ampliamente exitosa 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, elegir 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 **??**.

Otras formas de introducir estructuras son considerar un riesgo empírico regularizado $R_n(h) + \lambda\Omega(h)$ que puede verse como el Lagrangiano para minimizar $R_n(h)$ sujeto a $\Omega(h) \leq C$).

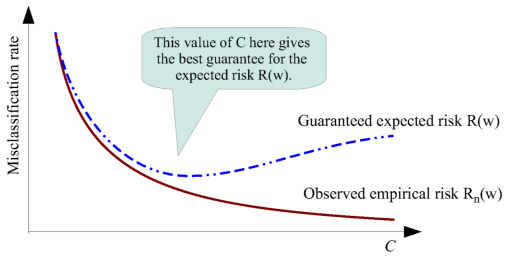

Figure 2: Fenómeno de distancia de los riesgos en funcion de la evolución de hiperparámetros

Dada tal configuración, uno puede evitar estimar la brecha entre el riesgo empírico y esperado dividiendo los datos disponibles en subconjuntos: un *conjunto de entrenamiento* utilizado para producir un subconjunto de soluciones candidatas, un *conjunto de validación* utilizado para estimar el riesgo esperado para cada candidato, y un *conjunto de prueba* utilizado para estimar el riesgo esperado para el candidato que finalmente se elige. Específicamente, sobre el conjunto de entrenamiento, uno minimiza una medida de riesgo empírico R_n sobre $\mathcal{H}_C$ para varios valores de C. Esto da como resultado un puñado de funciones candidatas. El conjunto de validación se usa para estimar el riesgo esperado correspondiente a cada solución candidata, luego de lo cual se elige la función que arroje el menor valor de riesgo estimado. 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 **??**.

Otro camino, aunque indirecto, hacia la minimización del riesgo es emplear un algoritmo para minimizar R_n, pero terminar el algoritmo *antes* de que se encuentre un minimizador real de R_n. De esta manera, el papel del hiperparámetro lo asume el tiempo de entrenamiento permitido, según el cual uno encuentra típicamente las relaciones ilustradas en la Figura **??**. Los análisis teóricos relacionados con la idea de la detención temprana son mucho más desafiantes que los de otras formas de minimización del riesgo estructural. Sin embargo, vale la pena mencionar estos efectos ya que la detención temprana es una técnica popular en la práctica y, a menudo, es *esencial* debido a las limitaciones del presupuesto computacional.

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.

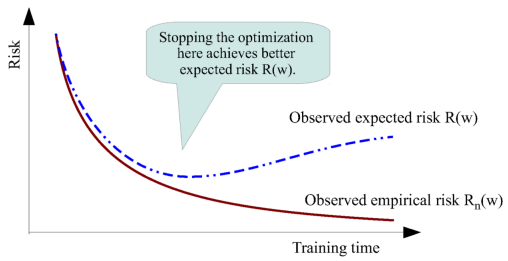

Figure 3: Ilustración de la detención temprana. Detener prematuramente la optimización del riesgo empírico R_n a menudo resulta en un mejor riesgo esperado R. De esta manera, el tiempo de parada juega un papel similar al del hiperparámetro C en la ilustración de minimización de riesgo estructural

1.3 ENUNCIADOS DE PROBLEMAS DE OPTIMIZACIÓN FORMAL

Los problemas de optimización en el machine learning surgen a través de la definición de las funciones de predicción y pérdida que aparecen en las mediciones de riesgo esperado y empírico que se pretenden minimizar. Nuestras discusiones giran en torno a las siguientes definiciones.

1.3.1 *Funciones de Predicción y Pérdida*

En lugar de considerar un problema de optimización variacional sobre una familia genérica de funciones de predicción, suponemos que la función de predicción h tiene una forma fija y está parametrizada por un vector real $w \in \mathbb{R}^d$ d sobre el cual se realizará la optimización. Formalmente, para algunos $h(\cdot;\cdot) : \mathbb{R}^{d_x} \times \mathbb{R}^d \to \mathbb{R}^{d_y}$ dados y, considerandos la familia de funciones de predicción

$$\mathcal{H} := \{h(\cdot; w) : w \in \mathbb{R}^d\} \tag{1.5}$$

Nuestro objetivo es encontrar la función de predicción en esta familia que minimice las pérdidas incurridas por predicciones inexactas. Para este propósito, asumimos una función de pérdida dada $\ell : \mathbb{R}^{d_y} \times \mathbb{R}^{d_y} \to \mathbb{R}$ como aquella que, dado un par de entrada-salida (x, y), produce la pérdida $\ell(h(x; w), y)$ cuando $h(x; w)$ e y son las salidas predichas y verdaderas, respectivamente.

1.3.2 *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. Es decir, suponemos que el espacio de entrada-salida $\mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$ está dotado con $P : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to [0, 1]$ y la función objetivo que deseamos minimizar es

$$R(w) = \int_{\mathbb{R}^{d_x} \times \mathbb{R}^{d_y}} \ell\left(h(x; w), y\right) dP(x, y) = \mathbb{E}\left[\ell\left(h(x; w), y\right)\right] \tag{1.6}$$

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

1.3.3 *Riesgo empírico*

Si bien puede ser deseable minimizar (1.6), 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 involucra una estimación del riesgo esperado R. 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 $R_n : \mathbb{R}^d \to \mathbb{R}$ dada por la ecuación

$$R_n(\omega) = \frac{1}{n} \sum_{i=1}^{n} l\left(h(x; w), y\right) \tag{1.7}$$

En términos generales, la minimización de R_n puede considerarse el problema práctico de optimización de interés. Por ahora, consideramos la medida no regularizada (1.7), pero notemos que los métodos de optimización que discutiremos en las secciones siguientes pueden aplicarse fácilmente cuando se incluye un término suave de regularización.

1.3.4 *Notación simplificada*

Las expresiones (1.7) y (1.6) muestran explícitamente cómo los riesgos esperado y empírico dependen de la función de pérdida, espacio de muestra o conjunto de muestras, etc. Sin embargo, cuando hablamos de métodos de optimización, a menudo emplearemos una notación simplificada que también ofrece algunos caminos para generalizar ciertas ideas algorítmicas. En particular, 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 $g(w; \xi)$, es decir, g es la composición de la función de pérdida y la función de predicción h.

De esta manera, el riesgo esperado para una w dada es el valor esperado de esta función compuesta tomada con respecto a la distribución de ξ:

$$R(w) = \mathbb{E}\left[g(w; \xi)\right] \tag{1.8}$$

De manera similar, dado un conjunto de realizaciones $\left\{\xi_{[i]}\right\}_{i=1}^n$ de ξ correspondientes a un conjunto de muestras $\{(x_i, y_i)\}_{i=1}^n$, definamos

la pérdida incurrida por el vector de parámetro w con respecto a la i-ésima muestra como

$$g_i(w) := g(w, \xi_{[i]}) \tag{1.9}$$

y luego escribamos el riesgo empírico como el promedio de las pérdidas de la muestra:

$$R_n(w) = \frac{1}{n} \sum_{i=1}^{n} g_i(w) \tag{1.10}$$

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 .

1.3.5 *Métodos Estocásticos vs. Optimización por Batch*

Ahora vamos a introducir algunos algoritmos de optimización fundamentales para minimizar el riesgo. Por el momento, dado que es la configuración típica en la práctica, presentamos dos clases de algoritmo en el contexto de minimizar el riesgo empírico medido R_n en (1.10). Hay que tener en cuenta, sin embargo, que gran parte de nuestra discusión posterior se centrará en el rendimiento de los algoritmos al considerar la verdadera medida de interés, es decir, el riesgo esperado R en (1.8).

Los métodos de optimización para el machine learning se dividen en dos grandes categorías. Nos referimos a ellos como estocásticos y por Batch. El método prototípico de optimización estocástica es el método del gradiente estocástico (DE) [22], que, en el contexto de minimizar R_n y con $w_1 \in \mathbb{R}^d$ dado, se define por

$$w_{k+1} \leftarrow w_k - \alpha_k \nabla g_{i_k}(w_k) \tag{1.11}$$

Aquí, para todo $k \in \mathbb{N} := \{1, 2, ...\}$, el índice i_k (correspondiente a la variable aleatoria $\xi_{[i_k]}$, es decir, el par de muestras (x_{i_k}, y_{i_k})) se elige *al azar* de $\{1, ..., n\}$ y α_k es un incremento positivo. Cada iteración de este método es, por lo tanto, muy barata, y solo incluye el cálculo del gradiente $\nabla g_{i_k}(w_k)$ correspondiente a una muestra. El método es notable porque la secuencia de iteración no está determinada únicamente por el la función R_n, 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 $\{i_k\}$. Aún así, como veremos en nuestro análisis en el capítulo 8, mientras que cada dirección $-\nabla g_{i_k}(w_k)$ podría no ser una de descendencia de w_k (en el sentido de producir una derivada direccional negativa para R_n de w_k), sí es una dirección de descenso *en esperanza*, luego la secuencia $\{w_k\}$ puede guiarse hacia un minimizador de R_n.

Para muchos en la comunidad de investigación de optimización, un enfoque por *Batch* es una idea más natural y conocida. El método más

simple de esta clase es el algoritmo de descenso más pronunciado - también conocido como gradiente, descenso de gradiente por Batch o método de gradiente completo- (GD)que se define mediante la siguiente iteración:

$$w_{k+1} \leftarrow w_k - \alpha_k \nabla R_n(w_k) = w_k - \frac{\alpha_k}{n} \sum_{i=1}^{n} \nabla g_i(w_k) \tag{1.12}$$

Calcular el paso $-\alpha_k \nabla R_n(w_k)$ en tal enfoque es más costoso que calcular el paso $-\alpha_k \nabla g_{i_k}(w_k)$ en DE, aunque se puede esperar que se calcule un mejor paso cuando todas las muestras se consideran en una iteración.

Los enfoques estocástico y por Batch ofrecen diferentes compensaciones en términos de costos de iteración y mejora esperada de la iteración para minimizar el riesgo empírico. ¿Por qué, entonces, DE ha alcanzado tal prominencia en el contexto del machine learning a gran escala? Comprender el razonamiento detrás de esto requiere una consideración cuidadosa de las compensaciones computacionales entre los métodos estocástico y por Batch, así como una mirada más profunda en sus habilidades para garantizar la mejora en el subyacente riesgo esperado R. Comenzaremos a investigar estos temas en la siguiente subsección.

1.4 MOTIVACIÓN PARA LOS MÉTODOS ESTOCÁSTICOS

Antes de analizar las fortalezas de los métodos estocásticos, como 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 basados en gradiente. 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 (1.12), 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 ∇R_n. 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. Permítanos motivarlos contrastando la iteración DE característica (1.11) con la iteración de gradiente de Batch completo (1.12).

1.4.1 *Motivación intuitiva*

En un nivel intuitivo, DE emplea información de manera más eficiente que un método por batch. Para ver esto, considere una situación en la cual un conjunto de entrenamiento, llámelo 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 uno tenía una copia de S_{sub}. Por otro lado, DE realiza los mismos cálculos en ambos escenarios, en el sentido de que los cálculos de gradientes estocásticos implican la elección de elementos de S_{sub} con las mismas probabilidades. 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.

Se puede llegar a una conclusión similar con el uso de conjuntos de entrenamiento, validación y prueba. Si uno cree que trabajar con solo, por ejemplo, la mitad de los datos en el conjunto de entrenamiento es suficiente para hacer buenas predicciones sobre datos no vistos, entonces uno puede argumentar en contra de trabajar con todo el conjunto de entrenamiento en cada iteración de optimización. Repitiendo este argumento, trabajar con solo una cuarta parte del conjunto de entrenamiento puede ser útil al inicio, o incluso con solo un octavo de los datos, y así sucesivamente. De esta manera, llegamos a la motivación de la idea de que trabajar con muestras pequeñas, al menos inicialmente, puede ser bastante atractivo.

1.4.2 *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 **??** compara el rendimiento de un método **L-BFGS** por batch [10] [16] y el método DE 1.11 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 RCV1. 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 DE realiza n pasos por *epoch*. La trama muestra el comportamiento en los primeros 10 *epochs*. La ventaja de DE 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 DE 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)

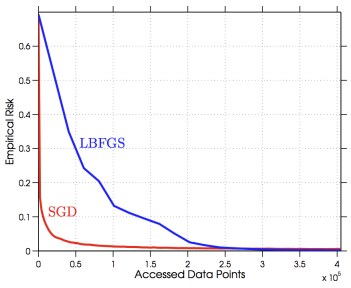

Figure 4: Riesgo empírico R_n en función del número de puntos de datos accedidos (ADP) para un método **L-BFGS** por batch y el método de gradiente estocástico (DE) 1.11 en un problema de clasificación binaria con un objetivo de pérdida logística y el conjunto de datos RCV1. SG se ejecutó con un incremento fijo de $\alpha = 4$

1.4.3 *Motivación Teórica*

También se pueden citar argumentos teóricos para una preferencia de DE sobre un enfoque por batch. Vamos a dar una vista previa de estos argumentos ahora, que se estudian con más profundidad y más detalles en la parte III:

- 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:

$$R_n(w_k) - R_n^* \leq \mathcal{O}\left(\rho^k\right) \tag{1.13}$$

  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 1.13, 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 $\log\left(\frac{1}{\epsilon}\right)$. Esto significa que, con un costo por iteración proporcional a n (debido a la necesidad de calcular $\nabla R_n(w_k)$ para todo $k \in \mathbb{N}$), el trabajo total requerido para obtener $\epsilon-$optimalidad para un método de gradiente por batch es proporcional a $n \log\left(\frac{1}{\epsilon}\right)$.

- 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 $\{1, \ldots, n\}$, entonces, para todo $k \in \mathbb{N}$, las iteraciones DE satisfacen la propiedad de convergencia sublineal (ver 8.3.3):

$$\mathbb{E}\left[R_n(w_k) - R_n^*\right] = \mathcal{O}\left(\frac{1}{k}\right) \tag{1.14}$$

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 $\epsilon-$optimalidad para DE es proporcional a $\frac{1}{\epsilon}$. Es cierto que esto puede ser mayor que $n \log\left(\frac{1}{\epsilon}\right)$ para valores chicos de n, ϵ, pero la comparación favorece a DE cuando se pasa al régimen de *big data* donde n es grande y uno es simplemente limitado por un presupuesto de tiempo computacional.

- Otra característica importante de DE es que produce la misma velocidad de convergencia que en 1.14 para el error en el riesgo esperado, $R - R^* *$, donde R^* es el valor mínimo de R. Específicamente, aplicando la iteración DE, pero con $g(wk, \xi_k)$ reemplazado por $\nabla f(w_k; \xi_k)$ con cada ξ_k tomado independientemente de acuerdo con la distribución P, vale:

$$\mathbb{E}\left[R(w_k) - R^*\right] = \mathcal{O}\left(\frac{1}{k}\right) \tag{1.15}$$

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 por batches en métodos de optimización para el machine learning a gran escala. Sin embargo, no pretendemos que los métodos por batch no tengan lugar en la práctica, si la Figura **??** uno considerara un mayor número de *epochs*, entonces uno vería que el algoritmo de batch eventualmente mejora al método estocástico y produce un menor error de entrenamiento. Esto motiva por qué muchos métodos propuestos recientemente intentan 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] [29]

1.5 ORGANIZACIÓN DE LA TESIS

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

La primer parte (I) refiere a la motivación tanto matemática como algoritmica y del área para analizar la convergencia de los algoritmos presentados, como a su vez los contenidos preliminares usados a lo largo del documento. Al final, en el capítulo 3, se incluye un resumen de los resultados vistos, pensando mayoritariamente en el aplicante del Machine Learning que quiere rápidamente saber condiciones de convergencia para sus algoritmos.

La segunda parte (II) trata exclusivamente los algoritmos deterministicos, comunmente denominados de *tipo batch*. En el Capítulo 4 , utilizando la gran referencia [15], analizamos la convergencia puntual del descenso de gradiente con condiciones de convexidad débil y luego la convergencia *lineal* con convexidad mas fuerte. Luego en el capítulo 5 nos basamos en [7] para ver a estos algoritmos como discretizaciones de sistemas dinámicos y con elt eorema de la variedad estable concluímos un método práctico para analizar la convergencia *casi todo punto*; esta forma de analizar los algortitmos es aplicada en el capítulo 6 con varios algoritmos estándar en el área. Finalmente en el capítulo 7 nos basamos en [4] para ver que aunque se tiene convergencia *casi todo punto*, la complijdad algoritmica del descenso de gradiente es exponencial.

Esto nos motiva pasar a analizar los algoritmos estocásticos en la parte III, donde nos basamos en [2] y [3] para analizar en el capítulo 8 la convergencia en *norma* $L1$ al mínimo (o a un entorno de éste) y finalmente en el capítulo 9 la convergencia *casi todo punto* tanto en los casos convexo como no convexo, ya que el algoritmo induce un proceso estocástico que induce una *cuasi-martingala* convergente.

2

PRELIMINARES

> Mathematics is the most beautiful and most powerful creation of the human spirit
>
> Stefan Banach

2.1 ESPECTRO

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

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

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

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

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

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

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

2.2 VARIEDADES DIFERENCIABLES Y TEOREMA DE LA VARIEDAD ESTABLE

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

2.2.1 *Un repaso por Variedades*

Definición 2.2.1 *[Capítulo 1 de [9]] Dado un espacio topológico X decimos que es una variedad diferenciable de dimensión d si:*

- *X es Haussdorf*
- *X es $2-$contable*
- *Existe un atlas suave para X, o sea existe un conjunto de pares $\{(U_i, \phi_i)\}$ tales que:*

1. *Para todo $x \in X$ existe (U, ϕ) con $x \in U$ y $\phi : U \to \phi U$ homeomorfismo*
2. *Si existen dos cartas $(U, \phi), (V, \psi)$ en el entorno de x con $U \cap V \neq \emptyset$ entonces $\phi \circ \psi^{-1} : \psi(U \cap V) \to \phi(U \cap V)$ es difeomorfismo*

Definición 2.2.2 *[Capítulo 6 de [9]] Dada una variedad de dimensión d χ y el espacio de medida $\left(\mathbb{R}^d, \mathcal{B}, \mu\right)$, decimos que $E \subset \chi$ tiene medida cero si existe un atlas $\mathcal{A} = \left\{U_i, \phi^i\right\}_{i \in \mathbb{N}}$ tal que $\mu\left(\phi^i\left(E \cap U_i\right)\right) = 0$. En este caso usamos el abuso de notación $\mu(E) = 0$.*

Definición 2.2.3 *[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 $D_g(x)v = \dfrac{d(g \circ \gamma)}{dt}(0) \in T_{g(x)}$. El determinante del operador lineal $det(D_g(x))$ es el determinante de la matriz que representa $D_g(x)$ con respecto a una base arbitraria.*

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

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

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

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

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

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

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

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

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

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

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

2.2.2 *Teorema de la variedad estable*

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

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

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

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

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

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

Demostración Ver Teorema III.1 de [23]

2.3 PRELIMINARES DE PROCESOS ESTOCÁSTICOS

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

2.3.1 *Esperanza condicional*

Dado un espacio de probabilidad $(\Omega, \mathcal{F}, P)$ definimos una *variable aleatoria* como una función $X : \Omega \mapsto \mathbb{R}$ tal que $X^{-1}(B) \in \mathcal{F}$ para todo $\mathcal{B} \in \mathcal{B}$ 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 *sigma algebra generada por las X_γ* $\mathcal{F} = \sigma\left(X_\gamma\right)$ como la menor sigma álgebra (en el sentido de la inclusión) tal que todas las X_γ son $\mathcal{F}$ medibles.

Recordemos además:

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

$$\int\limits_\Omega f_n d\mu \nearrow \int\limits_\Omega f d\mu$$

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

$$\int_\Omega |f_n - f| \, d\mu \to 0$$

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, $w_k := w_{k-1} - \alpha_k g(w_k, .) : \Omega \to \mathbb{R}$ y por hipótesis es $\mathcal{F}$ medible; es más, podemos ver a W como un proceso estocástico discreto.

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

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

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

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

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

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

Demostración Demostremos la existencia y unicidad:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Demostración En general con cuentas bastante similares a las vistas en cualquier curso de Análisis Real teniendo cuidado de las proyecciones. Para una mejor referencia ver [28].

2.3.2 *Martingalas y Cuasi-martingalas*

Definición 2.3.6 *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 $(\Omega, \mathcal{F}, P)$ dada por $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$

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

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

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

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

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

Observación El concepto de cuasi-martingalas es una generalización natural de las martingalas, submartingalas y supermartingalas. Fueron introducidos por primera vez por Fisk [5] para extender la descomposición de Doob-Meyer a una clase más grande de procesos. La forma en que las cuasimartingalas se relacionan con sub y súper martingalas es muy similar a cómo las funciones de variación finita se relacionan con funciones crecientes y decrecientes. En particular, mediante la descomposición de Jordan, cualquier función de variación finita en un intervalo se descompone como la suma de una función creciente y una función decreciente. De manera similar, un proceso estocástico es una cuasimartingala si y solo si puede escribirse como la suma de una submartingala y una supermartingala. Este importante resultado fue mostrado primero por Rao [20], y significó el inicio de la extensión de gran parte de la teoría de submartingalas a cuasimartingalas.

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

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

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

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

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

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

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

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

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

3 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 para utilidad próxima, 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 $F \in C^1$ la función objetivo, w^* su mínimo, asumamos que F es débilmente convexo 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 incrementos $\{\alpha_k\}$ cumplen la condición Robbins - Monro entonces:

$$w_k \to w^*$$

Teorema 3.1.2 *Sea $F \in C^1$ la función objetivo tal que existe F_{inf} valor mínimo, F es L-Lipshitz y PL-convexa; entonces el algortimo descenso de gradiente por batch con incremento fijo $\alpha_k = \frac{1}{L}$ cumple:*

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

3.1.2 *Objetivos no convexos*

Teorema 3.1.3 *Sea $F \in C^2$ 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:*

$$w_k \to w^* \quad ctp$$

Teorema 3.1.4 *Sea $F \in C^2$ 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:*

$$w_k \to w^* \quad ctp$$

Teorema 3.1.5 *Sea $F \in C^2$ 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:*

$$w_k \to w^* \quad ctp$$

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, $\mu-$Lipshitz en el Hessiano con $B, l, \mu \in poly(d)$ tal que si $\alpha_k = \alpha \leq \frac{1}{l}$ entonces w_k va a estar a $\Omega(1)$ de cualquier mínimo para todo $k \leq e^{\Omega(d)}$*

3.2 ALGORITMOS ESTOCÁSTICOS

3.2.1 *Objetivos convexos*

Teorema 3.2.1 *Sea $F \in C^1$ 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:*

$$\mathbb{E}\left[F(w_k) - F_{inf}\right] \to \frac{\alpha LM}{2c\mu}$$

Teorema 3.2.2 *Sea $F \in C^1$ 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:*

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

Luego el algoritmo descenso estocástico de gradiente generalizado cumple::

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

Teorema 3.2.3 *Sea $F \in C^1$ 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:*

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

Teorema 3.2.4 *Sea $F \in C^1$ 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:

$$w_k \to w^* \quad ctp$$
$$(w_k - w^*)\,\nabla F(w_k) \to 0 \quad ctp$$

3.2.2 *Objetivos no convexos*

Teorema 3.2.5 *Sea $F \in C^1$ 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:*

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K} \|\nabla F(w)_k\|_2^2\right] \to \frac{\alpha L M}{\mu}$$

Teorema 3.2.6 *Sea $F \in C^1$ 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:*

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

Teorema 3.2.7 *Sea $F \in C^2$ 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:*

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

Teorema 3.2.8 *Sea F 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. *Para $j = 2, 3, 4$ existen $A_j, B_j \geq 0$ tal que:*

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

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

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

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

$$F(w_k) \to F_\infty \quad ctp$$
$$\nabla F(w_k) \to 0 \quad ctp$$

Part II

Algoritmos de tipo Batch

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

4 CONVERGENCIA PUNTUAL

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

Galileo

Si existe un algoritmo que todo estudiante o practicante del Machine Learning conoce, es el descenso de gradiente clásico (o *Descenso de gradiente en batch*) [Ver algoritmo 4.1]. Un buen inicio es analizar la convergencia puntual del descenso de gradiente y bajo que condiciones se da.

Algorithmus 4.1 : Descenso de gradiente en batch

Input: $F \in C^1$, $\alpha_k > 0$, $w_1 \in \mathbb{R}^d$, $X = \{\xi_j\}_{j \leq N}$ muestra

for $k \in \mathbb{N}$ **do**

$$w_{k+1} \leftarrow w_k - \alpha_k \sum_{j=1}^{N} \nabla F(\xi_j)$$

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

Definición 4.0.1 *Decimos que $F \in C^1$ es débilmente convexo si cumple las siguientes dos propiedades:*

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

Observación Notemos que existen funciones no convexas tal que cumplen 4.0.1.

4.1 INTUICIÓN

Ganemos intuición acerca del proceso como probar la convergencia del algoritmo 4.1 en 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.

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

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

1. Vamos a definir una *función de Lyapunov*
2. Vamos a verificar computando su derivada temporal que es una función monótona decreciente y acotada, por lo que converge
3. Vamos a probar que converge a 0.

Proposición 4.1.1 (Objetivo débilmente convexo, Versión continua) *Sea $F \in C^1$ 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:*

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

Demostración Vayamos con los pasos que definimos:

Paso 1 Definamos la función de Lyapunov:

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

Paso 2 Notemos que:

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

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

Paso 3 Como $h(t) \searrow h_{inf}$ entonces $\frac{dh}{dt} \to 0$, supongamos por el absurdo que $h_{inf} > 0$, luego existe $\widetilde{\epsilon} > 0$ y $T \in \mathbb{R}$ tal que para todo $t \geq T$ vale que $h(t) = (w(t) - w^*)^2 > \widetilde{\epsilon}$. Si juntamos entonces 4.2 y 4.0.1 llegamos a un absurdo, concluímos que $h_{inf} = 0$ por lo que:

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

∎

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:*

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

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

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

y (u_k) converge.

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

Demostración Ver A

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

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

Teorema 4.2.2 (Objetivo débilmente convexo, incrementos decrecientes) *Sea $F \in C^1$, asumamos 4.0.1 y que existen $A, B \geq 0$ tal que para todo $w \in \mathbb{R}^d$ vale que:*

$$(\nabla F(w))^2 \leq A + B\,(w - w^*)^2 \tag{4.4}$$

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

$$w_k \xrightarrow{k \to \infty} w^* \tag{4.5}$$

Demostración Hagamos los 3 pasos análogos a 4.1.1:

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

Paso 2 Análogo a 4.1.1 notemos que:

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

Notemos que a diferencia de antes la naturaleza discreta del algoritmo lleva a un término positivo de ruido en las variaciones. Notemos que si usamos 4.3 y 4.4 entonces:

$$h_{k+1} - \left(1 + \alpha_k^2 B\right) h_k \leq \underbrace{-2\alpha_k\,(w_k - w^*)\,\nabla F(w_k)}_{\leq 0 \text{ por } 4.0.1} + \alpha_k^2 A \leq \alpha_k^2 A$$

Definamos ahora las sucesiones auxiliares:

$$\mu_k = \prod_{j=1}^{k-1} \frac{1}{1 + \alpha_j^2 B} \tag{4.6a}$$

$$h'_k = \mu_k h_k \tag{4.6b}$$

Notemos que $\log(\mu_k) = -\sum_{j=1}^{k-1} \log\left(1 + \underbrace{\alpha_j^2 B}_{\geq 0}\right) \geq -B \sum_{j=1}^{k-1} \alpha_j^2 \geq -B \sum_{j=1}^{\infty} \alpha_j$, por lo que μ_k es una sucesión decreciente acotada inferiormente por $e^{-B \sum_{j=1}^{\infty} \alpha_j}$, luego $\mu_k \searrow \mu_\infty > 0$. Ahora si volvemos a 4.2 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 concluímos que $\{h'_k\}$ converge; como $\underbrace{\mu_k}_{\geq 0} \to \mu_\infty > 0$ entonces $\{h_k\}$ converge.

Paso 3 De 4.2 como ya vimos que $h_{k+1} - (1 + \alpha_k^2 B) h_k$ es sumable concluímos que:

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

Supongamos que $h_k \to h_{inf} \neq 0$, luego existiría $K \in \mathbb{N}$ y $\widetilde{\epsilon} > 0$ tal que $h_k = (w_k - w^*)^2 > \widetilde{\epsilon}$ para todo $k \geq K$; luego de 4.0.1 concluímos que existe $M > 0$ tal que $M \leq (w_k - w^*) \nabla F(w_k)$ para todo $k \geq K$. Por 4.3 eso implica que $\sum_{k=1}^{\infty} \alpha_k (w_k - w^*) \nabla F(w_k) = \infty$, concluímos que $w_k \xrightarrow{k \to \infty} w^*$. ∎

Corolario 4.2.3 *Sea $F \in C^2$, asumamos 4.0.1 y que $\left\|\nabla^2 F\right\|_2^2 \leq L$; si consideramos el algoritmo 4.1 tal que incrementos $\{\alpha_k\}$ cumplen 4.3 entonces:*

$$w_k \xrightarrow{k \to \infty} w^* \tag{4.7}$$

4.3 ACERCA DE CONVEXIDAD FUERTE Y FUNCIONES L-LIPSHITZ

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

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

$$f(y) \geq f(x) + \nabla F(x)^T (y - x) + \frac{\mu}{2} \|y - x\|_2^2 \tag{4.8}$$

Proposición 4.3.2 *Sea $f \in C^1$ una función $\mu-$convexa, entonces son equivalentes:*

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

Demostración Ver A

Definición 4.3.3 *Decimos que una función $f \in C^1$ 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:*

$$\frac{1}{2} \|\nabla f(x)\|_2^2 \geq \mu \left(f(x) - f_{inf}\right) \tag{4.9}$$

Proposición 4.3.4 *Sea $f \in C^1$ 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

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

Corolario 4.3.5 *Sea $h = f + g$ donde f es fuertemente convexa y g es convexa, entonces h es fuertemente convexa. En particular, si f es convexa entonces el problema regularizado en L2 de minimizar $h = f + \lambda \|x\|_2^2$ es fuertemente convexo.*

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

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

∎

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

Definición 4.3.6 *Sea $f \in C^1$, decimos que es L-Lipshitz si existe $L > 0$ tal que para todos $x, y \in \mathbb{R}^d$ vale:*

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L \|y - x\|_2 \tag{4.10}$$

Proposición 4.3.7 *Sea $f \in C^1$ una función L−Lipshitz, entonces para las siguientes propiedades:*

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

5. $f(\alpha x + (1-\alpha)y) \geq \alpha f(x) + (1-\alpha)f(y) - \frac{\alpha(1-\alpha)}{2L}\|y-x\|_2^2$
6. $f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{1}{2L}\|y-x\|_2^2$
7. $(\nabla f(x) - \nabla f(y))^T(x-y) \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|_2^2$
8. $f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) - \frac{\alpha(1-\alpha)}{2L}\|\nabla f(y) - \nabla f(x)\|_2^2$

Valen las siguientes cadenas de equivalencias:

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

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

Demostración Ver A

Con todas estas propiedades, probemos el resultado histórico de [19]

Teorema 4.3.8 (Convergencia lineal , Objetivos L-Lipshitz y PL-convexos) *Sea $F \in C^1$ tal que existe F_{inf} valor mínimo, F cumple 4.3.6 y 4.3.3; entonces el algortimo 4.1 con incremento fijo $\alpha_k = \frac{1}{L}$ cumple:*

$$f(w_k) - f_{inf} \leq \left(1 - \frac{\mu}{L}\right)^k (f(w_o) - f_{inf}) \tag{4.11}$$

Demostración Notemos que si usamos 4.3.7 entonces tenemos:

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

Luego por 4.3.3 tenemos:

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

Luego obtenemos:

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

∎

5

TEOREMA DE LA VARIEDAD ESTABLE Y LOS PUNTOS FIJOS INESTABLES

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

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 que casos el algoritmo 4.1 converge (en alguna forma) con objetivos no convexos?

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

Usemos un caso modelo para ejemplificar porque no es probable que los metodos de primer orden (entre ellos el algoritmo 4.1) convergan a puntos silla. Sea $f : \mathbb{R}^n \to \mathbb{R}^n$ dada por $f(x) = \frac{1}{2}x^T Hx$ con $H = \textbf{diag}(\lambda_1, \ldots, \lambda_n)$; supongamos además que $\lambda_1, \ldots, \lambda_k > 0$ y $\lambda_{k+1}, \ldots, \lambda_n < 0$.

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

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

Por lo tanto:

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

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

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

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

Por lo tanto por inducción es fácil probar que:

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

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

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

Con lo que concluímos que:

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

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

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

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

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

De lo que vemos que los puntos críticos son:

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

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

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

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

Demostración Del lema Sea $x_0 \in \mathbb{R}^n$ y g la iteración de *gradient descent* dada por 5.1, luego:

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

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

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

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

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

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

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

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

Definición 5.2.1 *Sea:*

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

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

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

Teorema 5.2.2 *Sea $g \in C^1(\chi)$ tal que $\det\left(Dg(x)\right) \neq 0$ para todo $x \in \chi$, luego el conjunto de puntos iniciales que convergen por g a un punto fijo inestable tiene medida cero, i. e.:*

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

Demostración Para cada $x^* \in \mathcal{A}_g^*$ por 5.2.2 existe B_{x^*} un entorno abierto; es más, $\bigcup_{x^* \in \mathcal{A}_g^*} B_{x^*}$ forma un cubrimiento abierto del cual existe un subcubrimiento numerable pues X es variedad, i. e.

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

Primero si $x_0 \in \chi$ sea:

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

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

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

olo: Hace falta mostrar esto??

Lema 5.2.3 *$x_t \in \bigcap_{k \in \mathbb{N}} g^{-k}(B_{x_i^*})$ para todo $t \geqslant T$*

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

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

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

∎

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 $\mu(W_g) \leqslant \mu(W) \overset{5.2.2}{=} 0$. ∎

6

CONVERGENCIA CTP A MÍNIMOS : CASO GENERAL

6.1 DESCENSO DE GRADIENTE EN BATCH

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

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

$$x_{k+1} = g(x_k) \stackrel{\triangle}{=} x_k - \alpha \nabla f(x_k) \tag{6.1}$$

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

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

Demostración Es claro que un punto crítico de f es punto fijo de g; si $x^* \in \chi^*$ entonces $Dg(x^*) = Id - \alpha \nabla^2 f(x^*)$ y entonces por 2.1.3 los autovalores de Dg son $\{1 - \alpha\lambda_i \ : \ \lambda_i \in \{\mu \ : \ \nabla^2 f(x^*)v = \mu v \quad \text{para algún } v \neq 0\}\}$. Como $x^* \in \chi^*$ existe $\lambda_{j^*} < 0$ por lo que $1 - \alpha\lambda_{j^*} > 1$; concluímos que $x^* \in \mathcal{A}_g^*$. ∎

Proposición 6.1.3 *Bajo* 6.1.1 *y* $\alpha < \frac{1}{L}$ *entonces* $\det\left(Dg(x)\right) \neq 0$.

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

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

Luego por 6.1.1 tenemos que $\alpha < \frac{1}{|\lambda_i|}$ y entonces $1 - \alpha\lambda_i > 0$ para todo $i \in \{1, \dots, d\}$; concluímos que $\det\left(Dg(x)\right) > 0$. ∎

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

Demostración Por 6.1.2 y 6.1.3 tenemos que vale 5.2.4 y concluímos que $\mu\left(W_g\right) = 0$. ∎

6.2 PUNTO PRÓXIMO

El algoritmo de punto próximo esta dado por la iteración:

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

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

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

mos esto? Me un poco claro

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 $g \in \mathcal{C}^1(\chi)$*

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

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

Finalmente por diferenciación implicita obtenemos:

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

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

Corolario 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 concluímos 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.

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 $y_k^{S_i} = \left(x_{k+1}^{S_1}, \dots, x_{k+1}^{S_i}, x_k^{S_{i+1}}, \dots, x_k^{S_b}\right)$
6 $x_{k+1}^j \leftarrow x_k^j - \alpha \dfrac{\partial f}{\partial x_j}\left(y_k^{S_{i-1}}\right)$

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

Lema 6.3.1 *La iteración de Descenso por coordenadas esta dada por:*

$$x_{k+1} = g(x_k) \stackrel{\triangle}{=} g_d \circ g_{d-1} \circ \cdots \circ g_1(x) \tag{6.3}$$

Lema 6.3.2 *Si g está dada por 6.3 entonces si notamos $P_S = \sum_{i \in S} e_i e_i^T$ entonces:*

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

Demostración Notemos primero que:

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

Por lo tanto:

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

∎

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

Proposición 6.3.3 *Bajo 6 y $\alpha < \frac{1}{L_b}$ se tiene que* $\det(Dg(x)) \neq 0$

Demostración Basta probar que cada término de 6.4 es invertible, para eso:

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

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

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

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

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

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

Esta demo es tenda, hay que r una mejor y la en el Anexo

Afirmación 6.3.5 *Sea $y_t \in Ran(H)$, luego existe $i \in \{1,\dots,b\}$ y $\delta > 0$ tal que $\alpha\sum_{j\in S_i}\left|e_j^T H z_i\right| \geq \delta\left\|z_i\right\|_2$*

Lema 6.3.6 *Existe $\epsilon > 0$ tal que para todo $t \in \mathbb{N}$:*

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

Demostración Manteniendo la notación previa a la afirmación:

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

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

$$z_{i+1}^T H z_{i+1} < z_i^T H z_i - \alpha \left\| \sum_{j \in S_i} \left(e_j^T H z_i \right) e_j \right\|_2^2 \tag{6.5}$$

Por otro lado sabemos que para todo w vale:

$$w^T H w \geq \lambda_{min}(H) \|w\|_2^2 \geq -L_b \|w\|_2^2 \tag{6.6}$$

Luego si usamos 6.3.5, 6.6 y Cauchy-Schwartz existe $i \in \{1, \ldots, b\}$ y $\delta > 0$ tal que:

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

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

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

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

∎

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

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

Luego:

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

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

Corolario 6.3.7 *Sea g dado por el algoritmo de descenso por coordenadas con ecuación 6.3, bajo 6 y $\alpha < \frac{1}{L_b}$ se tiene que $\mu(W_g) = 0$.*

Demostración Por 6.3.3 y 6.3.4 tenemos que vale 5.2.4 y concluímos que $\mu(W_g) = 0$. ∎

7

RESULTADOS NEGATIVOS

Ya vimos de 6.1.4, que el descenso de gradiente, con cualquier inicialización aleatoria razonable, siempre escapará de los puntos de silla estrictos *eventualmente*, pero sin ninguna garantía sobre el número de pasos requeridos. Esto motiva a la siguiente pregunta:

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

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:

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

Luego si tomamos $\epsilon > 0$ y $w_0 = (x_1^0, x_2^0)$ uniformemente en $w_0 \in \widetilde{U} = [-1,1] \times \left[-\frac{3}{2}^{-e^{\frac{1}{\epsilon}}}, \frac{3}{2}^{-e^{\frac{1}{\epsilon}}}\right]$ entonces el algoritmo 4.1 necesita $k \geq e^{\frac{1}{\epsilon}}$ pasos para que $w_k \notin U$. Concluímos que el algoritmo es exponencial en converger a cualquier mínimo si $w_0 \in \widetilde{U}$. ∎

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

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

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

Luego si para el algoritmo 4.1 tomamos $\alpha_k = \alpha = \frac{1}{4}$ tendríamos la siguiente dinámica:

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

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

Teorema 7.1.1 (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, $l-$Lipshitz, $\mu-$Lipshitz en el Hessiano con $B, l, \mu \in poly(d)$ tal que si $\alpha_k = \alpha \leq \frac{1}{l}$ entonces w_k va a estar a $\Omega(1)$ de cualquier mínimo para todo $k \leq e^{\Omega(d)}$*

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

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

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

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 $w_0 = \left(x_1^0, x_2^0\right) \in U$ y definamos:

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

Notemos que como la dirección de escape en $(0, 0)$ es por x_1 y *luego* por x_2 (por el cambio de comportamiento de f) podemos concluir que k_1, k_2 estan bien definidos y que $k_2 \geq k_1 \geq 0$; la observación clave va a ser que $k_2 = Ck_1$ con $C > 1$, es decir que el tiempo en pasar el siguiente punto silla es exponencialmente mayor que los anteriores. En pos de esto, veamos como va a ser la iteración del algoritmo 4.1:

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

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

Luego evaluando en k_1 y k_2:

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

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

Concluímos que:

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

Esta f que presentamos tiene varios problemas:

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

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

Demostración Vayamos de a pasos

Paso 1 - Definiciones

Fijemos 4 constantes: $L = e, \gamma = 1, \tau = e, \eta$ a definir proximamente; inspirados en el ejemplo anterior vamos a construir una f definida en un cerrado D_0 tal que tenga $d-1$ puntos silla estrictos y la complejidas del algoritmo 4.1 sea exponencial. Sea D_0 dado por:

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

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:

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

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

Para $i = d$ definamos:

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

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

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

Lema 7.1.2 *Sea $g(x_i, x_{i+1}) = g_1(x_i) + g_2(x_i)x_{i+1}^2$, existen g_1, g_2 polinomios y $\eta = -g_1(2\tau) + 4L\tau^2$ tal que para todo $1 \leq i \leq d$ si $x_i = \tau$ vale:*

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

Y si $x_i = 2\tau$ entonces:

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

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

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

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

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

Demostración Ver A

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

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

Observación Notemos que f queda C^2 y que sus $d+1$ puntos críticos son $z_i = \left(4\tau, \ldots, \underbrace{4\tau}_{i}, 0, \ldots, 0\right)$ donde todos son puntos silla estrictos menos $z_d = (4\tau, \ldots, 4\tau)$ 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 $T_k = \min_{x_k^t \geq 2\tau}$ el tiempo de escape de $D_{k,2}$; notemos que como $x^0 \in D_{1,1}$ vale que $T_k \geq 0$ para todo k y esta bien definido. Definamos además T_k^τ como la cantidad de iteraciones que x^k esta en $D_{k,2}$ antes de escapar; como del lema $\frac{\partial g}{\partial x_k}(x_k, x_{k+1}) \leq -2\gamma\tau$ tenemos que $\left|x^k - x^{k+1}\right| \geq 2\alpha\gamma\tau$ por lo que:

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

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:

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

Paso 4 - El algoritmo 4.1 se queda confinado a D_0:

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

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

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

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

Separemos el caso $j = k+1$, donde el lema 7.1.2 nos dice que:

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

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

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

Y concluímos que $x^t \in D_0$

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

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

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

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

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

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

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

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

Inductivamente:

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

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:

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

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

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:

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

Luego definimos:

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

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 $(4\tau, \ldots, 4\tau)$.

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

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

∎

Part III

Algoritmos Estocásticos

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

8

CONVERGENCIA EN L1

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

Jerry Bona

8.1 CONTEXTO

En esta parte vamos a analizar la convergencia en L1 de algoritmos estocásticos para optimizar una $F : \mathbb{R}^d \mapsto \mathbb{R}$ que puede representar tanto el costo esperado como el empírico. Recordemos que F lo asumimos parametrizado por$w \in \mathbb{R}^d$ e imaginamos a los datos (x, y) como extraídos de una variable aleatoria ξ, cuya distribución desconocida es P, luego F se representa como:

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

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 ξ_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:

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

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

Para iniciar el analisis de la convergencia, lo mínimo que necesitamos es que el gradiente se mantenga controlado, por lo tanto recordemos la condición 4.3.6:

Hipótesis 8.1.1 (F es l-Lipshitz) *La función a optimizar $F \in C^1(\mathbb{R}^d)$ y existe $L > 0$ tal que para todos $w, z \in \mathbb{R}^d$:*

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

Observación Sea F bajo 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 4.3.7

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 $\mathbb{E}_{\xi_k}[.] := \mathbb{E}_{P_k}[.|w_k]$ 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 $k \in N$:*

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

Demostración Notemos que por 8.1.1 vale que:

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

Luego tomando esperanza de ambos lados y recordando 2.3.5:

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

∎

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

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

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

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

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 $k \in N$ valen:*

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

Y

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

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

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

Observación Notemos que si g es un estimador insesgado de ∇F entonces 8.6a y 8.6b valen con $\mu_G = \mu = 1$. Dejamos de ejercicio al lector notar que si H_k es simétrica positiva definida tal que H_k es independiente de ξ_k entonces tanto 8.6a como 8.6b valen.

Observación Bajo 8.2.2 y por 8.5 tenemos que:

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

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

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

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

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

Demostración Por 8.2.1 y 8.6a vale que:

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

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 $w \in \mathbb{R}^d$

Recordemos que de 4.3.1 y 6 vale que $c \leq L$

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

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

Se puede verificar que $z_* := w - \frac{1}{c}\nabla F(w)$ cumple que $q(z_*) = F(w) - \frac{1}{2c}\|\nabla F(w)\|_2^2 \leq q(z)$ para todo $z \in \mathbb{R}^d$; luego por 4.3.4 se tiene:

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

∎

Ya estamos en condiciones de demostrar nuestro primer resultado de convergencia para 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 $\xi_1, \ldots, \xi_{k-1}$ definamos:

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

Teorema 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:*

$$0 < \alpha \leq \frac{\mu}{LM_G} \tag{8.9}$$

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

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

Demostración Usando 8.2.3 con 8.9 y **??** tenemos para todo $k \in \mathbb{N}$ que:

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

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

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

Por otro lado notemos que:

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

Luego deducimos inductivamente que:

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

∎

Observación Notemos que si g es un estimador insesgado de ∇F entonces $\mu = M_G = 1$ por lo que $\alpha \in [0, \frac{1}{L})$ que es la condición que pedimos en 6.1.4.

Observación Notemos además que si $M = 0$ (o sea el algortimo 8.1 no tiene ruido) entonces la convergencia es lineal, recuperando el resultado de 4.3.8.

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 volúmen.

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

Teorema 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:*

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

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

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

Donde:

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

Demostración Notemos primero que por 8.10 para todo $k \in \mathbb{N}$ vale:

$$\alpha_k L M_G \leq \alpha_1 L M_G \leq \mu$$

Luego por 8.2.3 y **??** uno tiene para todo $k \in \mathbb{N}$:

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

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

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

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

rque vale k=1?

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

∎

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

Teorema 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 $\xi \in (0, 1)$ tal que para todo $k \in \mathbb{N}$:*

$$\mathbb{V}_{\xi_k}\left[g(w_k, \xi_k)\right] \leq M\xi^{k-1} \tag{8.11}$$

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

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

Luego vale:

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

Donde:

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

$$\rho := \max\left\{1 - \frac{\alpha c\mu}{2}, \xi\right\} < 1 \tag{8.13b}$$

Demostración Por 8.8a vale que:

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

Luego si juntamos 8.5, 8.6b, 8.12 y 8.11 entonces:

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

Por lo tanto:

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

Por que vale k

Probemos ahora la identidad por inducción. Para esto, notemos que el caso $k = 1$ vale por la definición de ω; luego si asumimos que vale para algún $k \geq 1$ entonces de 8.14, 8.13a y 8.13b:

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

∎

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:*

$$0 < \alpha \leq \frac{\mu}{LM_G} \tag{8.15}$$

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

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

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

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

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

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

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

∎

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

Para el de incrementos decrecientes, asumamos que $\{\alpha_k\}$ cumple la condición de *Robbins - Monro* 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 $k \in \mathbb{N}$::*

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

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

Demostración Como $\alpha_k \to 0$ por 4.3 entonces podemos asumir sin pérdida de generalidad que $\alpha_k L M_G \leq \mu$ para todo $k \in \mathbb{N}$, luego:

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

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

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

Luego:

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

Por lo que 8.17a esta probado. Finalmente como por 4.3 tenemos que $A_K \to \infty$ se tiene 8.17b. ∎

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

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 :*

$$\liminf_{k\to\infty} \mathbb{E}\left[\|\nabla F(w_k)\|_2^2\right] = 0 \tag{8.18}$$

Corolario 8.4.4 *Bajo las mismas hipótesis de 8.4.2 sea $k(K) \in \{1, \ldots, K\}$ un índice aleatorio elegido con probabilidades respectivas $\{\alpha_k\}_{k=1}^{K}$; luego $\|\nabla F(w_k)\|_2 \to 0$ en probabilidad.*

Demostración Sea $\epsilon > 0$, luego de 8.17a y la desigualdad de Markov:

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

∎

Porque aca es esperanza de la esperanza condicional?

.

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:*

$$\lim_{k\to\infty} \mathbb{E}\left[\|\nabla F(w_k)\|_2^2\right] = 0 \tag{8.19}$$

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:

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

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

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

Luego obtenemos tomando esperanza total:

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

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:

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

Sean:

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

Luego como en 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 $0 \leq \mathbb{E}\left[G(w_k)\right] = \mathbb{E}\left[G(w_0)\right] + S_k^+ - S_k^-$ tenemos que $\left\{S_K^-\right\}$ también es monótona y acotada superiormente, por lo que es convergente; concluímos que $\mathbb{E}\left[G(w_k)\right]$ debe ser convergente, y por 8.4.3 tenemos $\mathbb{E}\left[\|\nabla F(w_k)\|_2^2\right] = \mathbb{E}\left[G(w_k)\right] \to 0$. ∎

9 CONVERGENCIA CTP

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

Ahora que ya analizamos la convergencia en L1 de 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 simpleza 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:

$$w_k \xrightarrow[ctp]{k \to \infty} w^* \tag{9.1a}$$

$$(w_k - w^*)\,\nabla F(w_k) \xrightarrow[ctp]{k \to \infty} 0 \tag{9.1b}$$

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:

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

definimos bien filtracion? Es saber toda esa informacion a upo k, como lo notamos?

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

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

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

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

Ahora si incorporamos 9.1.1 obtenemos:

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

Definamos ahora las sucesiones auxiliares:

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

$$h'_k = \mu_k h_k \tag{9.4b}$$

Por lo que replicando las operaciones en 4.2.2 llegamos a:

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

Luego, recordando 2.3.10 notemos que:

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

Por el mismo motivo que en 4.2.2 concluímos que:

$$h'_k \geq 0 \tag{9.5a}$$

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

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

Paso 3 Como h_k converge ctp, de 9.3 concluímos que:

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

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

$$w_k \xrightarrow[\text{ctp}]{k \to \infty} w^* \tag{9.6a}$$

$$\left(w_k - w^*\right) \nabla F(w_k) \xrightarrow[\text{ctp}]{k \to \infty} 0 \tag{9.6b}$$

∎

9.2 CASO NO CONVEXO

Vamos a tomar las siguientes hipótesis para probar la convergencia ctp 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: $\mathbb{E}\left[g(w_k, \xi_k)\right] = \nabla F(w_k)$, luego tomemos las siguientes hipótesis:*

1. $F \in C^3$
2. *Existe $w^* \in \chi$ tal que $F_{inf} = F(w^*) \leq F(w)$, aunque notemos que no necesariamente es único*
3. *$F(w) \geq 0$ para todo $w \in \chi$ (Notemos que como hay un mínimo global, podemos redefinir $\widetilde{F} = F - F_{inf} \geq 0$)*
4. *Sean $\{\alpha_k\}$ los incrementos del algoritmo 8.1, entonces estos cumplen 4.3*
5. *Para $j = 2, 3, 4$ existen $A_j, B_j \geq 0$ tal que:*

$$\mathbb{E}\left[\left\|g(w_k, \xi_k)\right\|_2^j\right] \leq A_j + B_j \left\|w\right\|_2^j \tag{9.7}$$

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

$$\inf_{(w)^2 > D} w \nabla F(w) > 0 \tag{9.8}$$

9.2.1 *Acotación global del algoritmo*

Usemos 9.8 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 $w_1 \in U \subset \chi$ 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 9.8 y definamos $\phi : \chi \mapsto \mathbb{R}$ dada por:

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

Luego, sea:

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

Paso 2 Notemos que por la definición vale:

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

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

$$\begin{aligned} f_{k+1} - f_k \quad \leq \quad & \left(-2\alpha_k w_k g(w_k, \xi_k) + \alpha_k^2 g(w_k, \xi_k)^2\right) \phi'(w_k^2) \\ & +4\alpha_k^2 \left(w_k g(w_k, \xi_k)\right)^2 - 4\alpha_k^3 w_k g(w_k, \xi_k)^3 \\ & +4\alpha_k^4 g(w_k, \xi_k)^4 \end{aligned} \tag{9.9}$$

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

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

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

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

Porque pablo? esto no me cie que tantas acotaciones ha

Si ahora incluímos 9.7 entonces esto implica que existen $A, B \geq 0$ tal que:

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

Notemos ahora que si $w_k^2 < D$ entonces $\phi'(w_k^2)=0$, y si $w_k^2 \geq D$ entonces $-2\alpha_k w_k \nabla F(w_k)\phi'(w_k^2) < 0$ por 9.8 por lo que deducimos:

$$\mathbb{E}\left[f_{k+1}-f_k|\mathcal{P}_k\right] \leq \alpha_k^2\left(A+Bf_k\right) \tag{9.10}$$

Dado que hici mismo 3 veces valdría que se lema?

Ahora siguiendo los mismo pasos que al demostrar 9.1.2 definiendo μ_k, f_k' y usando 2.3.11 concluímos que $\{f_k\}$ converge ctp.

Paso 3 Supongamos que $f_{inf} > 0$, entonces existe $T \in \mathbb{N}$ tal que $w_k^2, w_{k+1}^2 > D$ para todo $k \geq T$ por lo que 9.9 es una igualdad y deducimos que:

$$\sum_{k=1}^{\infty}\alpha_k w_k \nabla F(w_k)\phi'(w_k^2) < \infty \quad ctp \tag{9.11}$$

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

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

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

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:*

$$F(w_k) \xrightarrow[ctp]{k\to\infty} F_\infty \tag{9.12a}$$

$$\nabla F(w_k) \xrightarrow[ctp]{k\to\infty} 0 \tag{9.12b}$$

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

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

Paso 2 Por 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:

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

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

$$\mathbb{E}\left[h_{k+1} - h_k | \mathcal{P}_k\right] \leq \underbrace{-2\alpha_k \left(\nabla F(w_k)\right)^2}_{\leq 0} + \alpha_k^2 K_1 \underbrace{\mathbb{E}\left[g(w_k, \xi_k)\right]}_{\leq K_2 \text{ por } 9.2.2} \tag{9.13}$$

Lo que implica:

$$\mathbb{E}\left[h_{k+1} - h_k | \mathcal{P}_k\right] \leq \alpha_k^2 K_1 K_2 \tag{9.14}$$

Luego como tenemos de esto que:

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

Por lo que por 2.3.11 obtenemos:

$$F(w_k) \xrightarrow[\text{ctp}]{k \to \infty} F_\infty \tag{9.15}$$

Paso 3 Si retomamos 9.13, reordenamos, sumamos hasta k y tomamos esperanza tenemos:

$$\begin{aligned} 2\sum_{k=1}^{K} \alpha_k \left(\nabla F(w_k)\right)^2 \leq & \quad \sum_{k=1}^{K} \mathbb{E}\left[h_{k+1} - h_k\right] + K_2 K_1 \sum_{k=1}^{K} \alpha_k^2 \\ = & \quad \mathbb{E}\left[\sum_{k=1}^{K} h_{k+1} - h_k\right] + K_2 K_1 \sum_{k=1}^{K} \alpha_k^2 \\ = & \quad \mathbb{E}\left[h_{k+1}\right] + K_2 K_1 \sum_{k=1}^{K} \alpha_k^2 \\ \xrightarrow[2.3.5]{\text{ctp}} & \quad F_\infty + K_2 K_1 \sum_{k=1}^{\infty} \alpha_k^2 \end{aligned} \tag{9.16}$$

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

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

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

$$\mathbb{E}\left[g_{k+1} - g_k | \mathcal{P}_k\right] \leq 2\alpha_k K_4 \left(\nabla F(w_k)\right)^2 + K_2 K_3 \alpha_k^2 \tag{9.17}$$

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

$$\begin{aligned} \sum_{k=1}^{\infty} \mathbb{E}\left[\delta_k^g\,(g_{k+1} - g_k)\right] &= \sum_{k=1}^{\infty} \mathbb{E}\left[\delta_k^g \mathbb{E}\left[(g_{k+1} - g_k)|\mathcal{P}_k\right]\right] \\ &\leq K_4 \underbrace{\sum_{k=1}^{\infty} \alpha_k \left(\nabla F(w_k)\right)^2}_{9.16} + K_2 K_3 \underbrace{\sum_{k=1}^{\infty} \alpha_k^2}_{4.3} \\ &< \infty \end{aligned} \tag{9.18}$$

Nuevamente por 2.3.11concluímos que $\{g_k\}$ converge *ctp*, y por 9.16 este límite es 0; luego:

$$g_k \xrightarrow[\text{ctp}]{k \to \infty} 0 \tag{9.19a}$$

$$\nabla F(w_k) \xrightarrow[\text{ctp}]{k \to \infty} 0 \tag{9.19b}$$

∎

Part IV

Apéndice

A

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:*

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

Donde:

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

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

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

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

Demostración Ver [21]

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

Demostración Ver [1]

A.2 DEMOSTRACIONES

Demostración [De 4.2.1] Definamos:

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

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

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

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

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

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

Demostración [De 4.3.2]

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

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

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:

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

me queda claro este paso final...

∎

Demostración [De 4.3.4]

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

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

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:

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

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 $\mu-$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:

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

4. Notemos que por el punto anterior:

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

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 4.3.7]

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

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

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]

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

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

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

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

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

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

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

   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:

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

∎

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

[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. URL: http://dx.doi.org/10.1007/BF01589116.

[11] S. Lojasiewicz. *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.

[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. 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] Vladimir Vapnik. *Estimation of Dependences Based on Empirical Data: Springer Series in Statistics (Springer Series in Statistics)*. Berlin, Heidelberg: Springer-Verlag, 1982. ISBN: 0387907335.

[28] David Williams. *Probability with Martingales*. Cambridge University Press, 1991. DOI: 10.1017/CBO9780511813658.

[29] Gersende Fort Yves F Atchade and Eric Moulines. «On stochastic proximal gradient algorithms.» In: (2014).