UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA MATEMÁTICA

# ANÁLISIS DEL ALGORITMO DE DESCENSO DE GRADIENTE ESTOCÁSTICO DE LANGEVIN Y APLICACIÓN EN EL PROBLEMA DE TOMOGRAFÍA SÍSMICA PASIVA

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN CIENCIAS DE LA INGENIERÍA, MENCIÓN MATEMÁTICAS APLICADAS

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL MATEMÁTICO

BRUNO NICOLÁS HERNÁNDEZ PEREIRA

PROFESOR GUÍA:
JOAQUÍN FONTBONA TORRES

MIEMBROS DE LA COMISIÓN:
AXEL OSSES ALVARADO
CRISTÓBAL GUZMÁN PAREDES
JORGE PRADO GUZMÁN

Este trabajo ha sido parcialmente financiado por:
CMM ANID BASAL FB210005

SANTIAGO DE CHILE
2023

RESUMEN DE LA TESIS PARA OPTAR AL GRADO
DE MAGÍSTER EN CIENCIAS DE LA INGENIERÍA,
MENCIÓN MATEMÁTICAS APLICADAS
RESUMEN DE LA MEMORIA PARA OPTAR AL TÍTULO
DE INGENIERO CIVIL MATEMÁTICO
POR: BRUNO NICOLÁS HERNÁNDEZ PEREIRA
FECHA: 2023
PROF. GUÍA: JOAQUÍN FONTBONA TORRES

# ANÁLISIS DEL ALGORITMO DE DESCENSO DE GRADIENTE ESTOCÁSTICO DE LANGEVIN Y APLICACIÓN EN EL PROBLEMA DE TOMOGRAFÍA SÍSMICA PASIVA

En el siguiente trabajo se presenta el algoritmo de Descenso de Gradiente Estocástico de Langevin (SGLD, sigla por su nombre en inglés), algoritmo de minimización basado en los algoritmos de descenso de gradiente estocástico con la característica de añadir ruido Gaussiano exógeno con el fin de esquivar óptimos locales. Se recopila y desarrolla el marco teórico necesario para justificar el uso del algoritmo SGLD para el problema de riesgo esperado empírico en una base de datos finita, basado en los resultados principales de Borkar y Mitter, 1999 y Raginsky y col., 2017, en busca de las condiciones y parámetros que permitan la convergencia del algoritmo hacia el punto óptimo de la función objetivo del problema. Se muestra que el marco de condiciones y parámetros propuestos no asegura convergencia hacia el óptimo, pero, gracias al comportamiento asintótico del error medio encontrado en la Sección 4.6, es posible controlar el sesgo del valor de la función objetivo encontrada con el algoritmo SGLD.

En la segunda parte de este trabajo, se presenta el problema de tomografía sísmica pasiva y la formulación y resolución presentada en Delplancke y col., 2020. Junto con los resultados de Delplancke y col., 2023, se muestra cómo el problema de tomografía puede enmarcarse en un problema de riesgo esperado empírico con el fin de usar la teoría mostrada previamente y encontrar solución mediante el algoritmo SGLD. Se muestran las diferencias entre los rendimientos del algoritmo SGLD y el algoritmo de descenso de gradiente estocástico usual para la tomografía. Entre estas diferencias se puede destacar la mejora en resolución y detalles para los campos de velocidades de ondas sísmicas estimados mediante el algoritmo SGLD. Finalmente se discute la elección de los parámetros y bajo qué contexto el uso del algoritmo SGLD es más beneficioso que el método de gradiente usual.

*"No puedes ser diferente,*
*las matemáticas no mienten."*
***Baldor - Chancho En Piedra***

# Agradecimientos

Por mi contexto personal y la etapa de mi vida en que se enmarca este trabajo, por todo el tiempo invertido y el cierre del proceso que representa me gustaría agradecer principalmente a Paula, quien ha sido mi gran soporte emocional a través de los años y durante mi carrera, quien ha acompañado mis penas y celebrado mis triunfos. Así mismo a mi madre que, mediante su experiencia, me ha aconsejado en todo lo que ha podido, brindándome su apoyo incondicional y plena confianza. A la Nana, quién me dio la seguridad y fuerza para cumplir cada objetivo que me he propuesto, a quién le dedico este trabajo.

Agradecer a mi padre y a mi hermano, con quienes comparto mi fascinación en las ciencias, tecnología y las artes. A Miguel y Jaime, y los amigos del colegio e infancia que siguen dándome buenos momentos y alegrías.

Quisiera agradecer, de todo corazón a Joaquín Fontbona, mi profesor guía, quien me dio la oportunidad de participar en su proyecto y depositó su confianza en mí, y a Jorge Prado, quien compartió conmigo su tiempo y conocimientos, tanto de matemáticas como de computación. También agradecer al Centro de Modelamiento Matemático y al Departamento de Ingeniería Matemática de la Universidad de Chile junto a todos sus profesores e investigadores, por apostar siempre por el gran potencial de sus estudiantes.

Quiero dar las gracias a quienes me acompañaron en mi carrera universitaria. A mis compañeros y amigos de Bachillerato, quienes son mi representación de verdaderos alumnos Uchile. A mis compañeros de la carrera de Ingeniería en Matemáticas, quienes sufrieron las mismas penas a través de los años y a quienes estoy en obligación de nombrar, como a Sebastián López, Pablo Araya, Javier Santibañez, Nicolás Valenzuela y Benjamín Barrientos, a María José Alfaro, Matías Azocar, Benjamín Madariaga, Vicente Salinas, Javier Madariaga, Vicente Saavedra, al Gitano y, por último, por tanto menos importante, al Pelela.

A todos aquellos que estuvieron presentes y participaron con su aporte a mi formación y vida durante los últimos años, muchísimas gracias.

_____ 21 de mayo, 2023

Bruno Nicolás Hernández Pereira

# Tabla de Contenido

# Índice de Ilustraciones

# Introducción

## El algoritmo de descenso de gradiente estocástico

El estudio de los algoritmos de optimización ha tenido un crecimiento importante en los últimos años; más aún, dado el gran interés en el análisis masivo de datos. Entre ellos, el algoritmo de descenso de gradiente se ha convertido en la herramienta favorita para la búsqueda de óptimos de una función, desde su origen, atribuido al matemático francés, Augustin Louis Cauchy, en su artículo *Méthode générale pour la résolution des systémes d'équations simultanées* en octubre de 1847, hasta el día de hoy. Sin embargo, este algoritmo posee múltiples limitaciones en su uso: la regularidad de la función objetivo, la temprana detención de sus iteraciones en mínimos locales, el difícil ajuste del parámetro de aprendizaje y dirección de decrecimiento y, con el aumento en la cantidad de datos, el costo computacional del cálculo del gradiente de la función objetivo. Para el último de estos problemas, han surgido numerosas variaciones del algoritmo, siendo (quizás) el más importante el *algoritmo de gradiente estocástico* (SGD[1]), originado por las ideas de Herbert Robbins y Sutton Monro, en la publicación *A stochastic approximation method* de 1951. Este algoritmo está encargado de resolver el problema del cálculo del gradiente de la función objetivo, intercambiándolo por un estimador insesgado de este con menor costo computacional. En Ma y col., 2015, se propuso un método general de construcción de cierta familia de algoritmos SGD. Alguno de los algoritmos que se describen mediante este método son: Monte Carlo Hamiltoniano (HMC), Gradiente Estocástico Monte Carlo Hamiltoniano (SGHMC), Gradiente Estocástico con Termostato de Nosé-Hoover (SGNHT), Gradiente Estocástico de Dinámica de Langevine Riemanniana (SGRLD) y Gradiente Estocástico de Dinámica de Langevin (SGLD).

Hasta ahora, las investigaciones se han limitado a observar cómo el algoritmo SGD permite reducir la complejidad del cálculo de gradiente, esquivando el problema de los mínimos locales, restringiendo los casos a aquellos problemas de optimización cuya función objetivo es convexa. Una de estas investigaciones se puede encontrar en Bottou, 1998, donde se realiza un análisis profundo sobre la convergencia del algoritmo tradicional de gradiente estocástico bajo aproximaciones mediante martingalas, para funciones objetivos convexas. El algoritmo que sí intenta resolver el inconveniente de los mínimos locales es el algoritmo SGLD, cuya particularidad es el ruido aleatorio exógeno en cada paso del descenso, evitando que el algoritmo se estanque en puntos críticos de la función, dando la oportunidad de observar otros puntos en caso de caer en un óptimo local. Además, en Borkar y Mitter, 1999 se probó que la ley de la recursión del algoritmo SGLD converge, cuando el paso tiende a 0, a la ley de transición de un proceso de difusión, dado por una ecuación diferencial estocástica particular o proceso de difusión de Langevin, dando origen al nombre de la variante de descenso estocástico.

Diversos estudios se han centrado en el análisis del algoritmo SGLD y su relación con los

[1] Cada sigla de los algoritmos será originada por los nombres respectivos en inglés.

procesos de difusión. En Chiang y col., 1987, se probó que la ley de transición del proceso de difusión de Langevin converge débilmente a la distribución de Gibbs asociada a la función objetivo del problema de minimización, ley que, bajo valores altos de su parámetro de temperatura, distribuye uniformemente sobre el conjunto de óptimos globales de la función, incluso cuando la función objetivo resulta no ser convexa. En Li y col., 2017, se ocupó teoría de control óptimo para describir el paso y el ruido necesario para lograr acercarse a esta distribución límite. A pesar de que el conocimiento sobre el proceso de difusión da a entender que existe un acercamiento entre el algoritmo SGLD y el óptimo de la función objetivo, hasta recientemente no se había realizado un análisis completo que vinculara las realizaciones del algoritmo SGLD con el punto óptimo. El trabajo que más se ha acercado a esto último es el realizado en Raginsky y col., 2017, que impone condiciones necesarias sobre el contexto del problema, para concluir que el valor esperado de las realizaciones del algoritmo SGLD está (en cierto sentido) cerca del valor óptimo de la función. Lo cierto es que, con el resultado mostrado en Raginsky y col., 2017, no es posible observar que la cercanía concluida garantice una verdadera convergencia, con el fin de generar una heurística de parámetros para usar en la práctica.

El objetivo general de la primera parte de este trabajo es estudiar un marco de condiciones necesarias en el contexto de los problemas de optimización, para describir la proximidad del valor esperado de las realizaciones del algoritmo SGLD al valor del óptimo global de la función objetivo, utilizando herramientas matemáticas vinculadas a los procesos estocásticos, estadística y análisis funcional, con el fin de garantizar, bajo ciertas condiciones verificables, la proximidad entre estos valores, cuantificar su sesgo y cuestionar el uso de la herramienta. Para ello, se estudiaron y complementaron principalmente los resultados de los trabajos de Borkar y Mitter, 1999 y Raginsky y col., 2017, que comparten el enfoque deseado.

# Tomografía sísmica

La tierra está en constante deformación debido a esfuerzos internos o externos. Si consideramos un bloque de tierra bajo pequeños esfuerzos, existirá deformación, pero en caso de acumulaciones, eventualmente ocurrirá una fractura en el terreno. Una fractura implica una liberación repentina de energía y genera ondas sísmicas de naturaleza elástica, que viajan a través de la tierra (Lee y col., 1981). Para el sector minero, el estudio de la sismicidad inducida por la propia minería es de particular interés, pues tal conocimiento permite generar e implementar nuevas políticas de seguridad para sus trabajadores, que se sitúan constantemente en medios con mucha acumulación de tensión producida por las faenas. En Lee y col., 1981, capítulo 4, se presenta un modelo de rayos sísmicos bajo la regla del tiempo mínimo de viaje de las ondas sísmicas en un medio, ley física que se representa por la *ecuación eikonal*, ecuación diferencial que vincula el tiempo mínimo de viaje con el campo de lentitud (propio del medio).

Como el campo de lentitud es una función desconocida en la práctica, el cálculo de los tiempo mínimos mediante la resolución de la ecuación eikonal no es un procedimiento factible, lo que crea la necesidad de buscar métodos equivalentes para estimar los campos de lentitud, como lo descrito en Tarantola, 2004, Nolet, 2008 y Nowack y Li, 2009, utilizando el principio de Fermat para expresar el tiempo total de viaje entre dos puntos. En Delplancke y col., 2020, se desarrolló un algoritmo de estimación del campo de lentitud basado en el estimador máximo a posteriori (MAP), cuya justificación teórica se abordaría con más profundidad posteriormente en Delplancke y col., 2023. El método para producir un estimador

MAP a partir de una base de datos tiempo-espacial de sismos se llevó a cabo mediante la maximización de la densidad posteriori del modelo, ocupando el algoritmo SGD, dada la gran cantidad de dimensiones que poseer el parámetro, que se adapta al contexto de muchos sistemas de la gran minería.

Como objetivo particular, este trabajo tiene el propósito de ajustar el modelo de estimación del campo de lentitud con el fin de verificar, completa, nula o parcialmente, los criterios de convergencia para el uso del algoritmo SGLD, comparando ambos algoritmos y enumerando los puntos a favor y en contra del uso de cada método. Para cumplir este objetivo, se trabajará con el modelo de datos simulado, descrito en Delplancke y col., 2020, en el lenguaje de programación `python`.

# Capítulo 1

# Preliminares

## 1.1. Elementos de probabilidades y estadística

Durante todo este trabajo, y cada vez que se presente alguna variable aleatoria, se asumirá que cada variable está definida sobre un espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$, donde $\Omega$ es el conjunto de eventos, $\mathcal{F}$ es una $\sigma$-álgebra sobre $\Omega$ y $\mathbb{P}$ es una medida de probabilidad sobre el espacio medible $(\Omega, \mathcal{F})$. Si se dice que esta variable aleatoria, digamos $Y$, toma valores en algún conjunto $A$, entonces se asumirá que la aplicación $Y : \Omega \longmapsto A$ es una función $\mathcal{F} - \mathcal{G}$-medible, para $\mathcal{G}$ una $\sigma$-álgebra sobre $A$ y se obviará la dependencia de $\omega \in \Omega$, por lo que escribiremos $Y := Y(\omega)$ a menos que sea necesario su uso. En caso de no especificar $\mathcal{G}$, se asumirá que esta corresponde a la $\sigma$-álgebra boreliana de $A$ (ver Definición 1.1.9 de San Martín, 2015) y se denotará como $\mathscr{B}(A)$, en tal situación se marcarán dos casos particulares; el primero será aquel que $A \subseteq \mathbb{R}^d$, para algún $d \in \mathbb{N}$, donde se utilizará la topología traza en $A$ heredada por la topología de la norma Euclideana en $\mathbb{R}^d$ (ver Capítulo 3 de Simmons, 1983), el segundo caso será cuando $A$ corresponda a un conjunto discreto, donde se usará la topología discreta (ver Simmons, 1983, Capítulo 3, Ejemplo 2).

Si $Y$ es una variable aleatoria sobre $(\Omega, \mathcal{F}, \mathbb{P})$ a valores en $\mathbb{R}^d$, con $d \in \mathbb{N}$, entonces podemos definir la ley de $Y$, $\mathcal{L}(Y)$ como (ver Sección 3 de Billingsley y col., 1999)

$$\mathcal{L}(Y)(A) = \mathbb{P}\left(Y^{-1}(A)\right),$$

para cualquier conjunto $A \in \mathscr{B}(\mathbb{R}^d)$. El caso más recurrente para este estudio será aquel en que $\mathcal{L}(Y)$ posea una densidad de probabilidad con respecto a la medida de Lebesgue en $\mathbb{R}^d$, $\lambda^d(\cdot)$. La siguiente definición y notación será utilizada posteriormente.

**Definición 1.1** (Definición 4.1.2 en San Martín, 2015) *Sean $\mu$ y $\nu$, medidas de probabilidad sobre el mismo espacio medible $(X, \mathcal{T})$. Se dirá que $\nu$ es* ***absolutamente continua*** *con respecto a $\mu$ (lo que notaremos como $\nu \ll \mu$), si para todo $A \in \mathcal{T}$ se tiene que $\mu(A) = 0 \Rightarrow \nu(A) = 0$.*

Notando que, tanto $\lambda^d(\cdot)$ como $\mathcal{L}(Y)$ son medidas de probabilidad sobre $\mathscr{B}(\mathbb{R}^d)$, si $\mathcal{L}(Y) \ll \lambda^d$, por el teorema de Radon-Nikodým (Teorema 4.1.3 en San Martín, 2015), existe una función $p_Y : \mathbb{R}^d \longmapsto \mathbb{R}_+$ tal que

$$\mathcal{L}(Y)(A) = \int_A p_Y(x)dx,$$

la cual será llamada la **función de densidad de probabilidad** (o tan sólo la densidad) de $Y$. Cabe destacar que la Definición 1.1 se puede extender al caso cuando $\nu$ es una medida con signo, e incluso cuando las medidas son medidas con signo no necesariamente finitas, bajo ciertas condiciones de $\mathcal{T}$. Sin embargo, estos casos no serán mencionados ya que no tendrán cabida en los contextos posteriores.

A una variable aleatoria $Y$, cuya realizaciones sean registrables, será llamada **observable**. Si $Y$ es v.a. observable y $\{y_i\}_{i=1}^n$ es un conjunto de observaciones de $Y$, se dirá que las realizaciones son **independientes e idénticamente distribuidas** (desde ahora i.i.d.) si cada una de ellas son independientes a pares de cualquier otra del conjunto y cada una posee ley $\mathcal{L}(Y)$. Si la medida de probabilidad no es especificada previamente, será posible denotar por $\mathcal{L}(Y)$-i.i.d. para señalar lo anterior con respecto a la ley $\mathcal{L}(Y)$. Un conjunto $\mathcal{D}$ será llamado **muestra aleatoria simple** (desde ahora M.A.S.) de $Y$ (de tamaño $n \in \mathbb{N}$) si está compuesto por $n$ realizaciones $\mathcal{L}(Y)$-i.i.d..

La siguiente definición es una modificación de las definiciones 2.4.4 y 2.4.5, en conjunto con el Teorema 2.4.6 de San Martín, 2015.

**Definición 1.2** *Sea $(X, \mathcal{T}, \mu)$ un espacio de medida, y sea $f : \Omega \longmapsto \mathbb{R}^d$ una función $\mathcal{T} - \mathscr{B}(\mathbb{R}^d)$-medible. Diremos que $f$ es una* ***función integrable*** *si*

$$\int |f| d\mu < \infty.$$

Esta definición es extensible a variables aleatorias. Si la medida $\mu$ es de probabilidad, e $Y$ una variable aleatoria en $X$, entonces decimos que $Y$ es una v.a. integrable si cumple la Definición 1.2 como función $\mathcal{T} - \mathscr{B}(\mathbb{R}^d)$-medible y denotamos por

$$\mathbb{E}(Y) = \int Y d\mu,$$

al valor de su integral.

El siguiente teorema clásico justifica los métodos de aproximación de problemas de riesgos que se mencionarán en los capítulos siguientes y es conocido como la **Ley de los Grandes Números** (desde ahora LGN), en sus versiones débil y fuerte.

**Teorema 1.1** (Teoremas 9.9.1 y 9.9.3 de San Martín, 2015) *Supongamos que $\{y_k\,,\, k \geq 1\}$ es una colección de realizaciones $\mathcal{L}(Y)$-i.i.d. de una variable aleatoria integrable $Y$, con $\mathbb{E}(Y) = u$. Entonces*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} y_k = u,$$

*donde la convergencia se cumple en probabilidad (versión débil), casi seguramente y en $L^1$ (versión fuerte).*

## 1.2. Elementos procesos markovianos y desigualdades logarítmicas de Sobolev

En este capítulo se revisarán algunos aspectos importantes sobre los procesos markovianos de difusión y desigualdades logarítmicas de Sobolev que serán mencionados en el análisis de los capítulos posteriores.

### 1.2.1. Semigrupo markoviano

Consideremos una familia de operadores $\mathbf{P} = (P_t)_{t\geq 0}$ definidos sobre un conjunto de funciones medibles, a valores reales, sobre un espacio medible $(X, \mathcal{T})$, y que cumplen las siguientes propiedades:

i) Para cada $t \geq 0$, $P_t$ es un operador lineal que manda cada función medible y acotada de $(X, \mathcal{T})$ en otra función medible y acotada.

ii) $P_0 = Id$, el operador identidad. (*condición inicial*)

iii) $P_t(\mathbf{1}) = \mathbf{1}$, donde $\mathbf{1}$ es la función constante, igual a 1. (*conservación de masa*).

iv) Si $f \geq 0$, entonces $P_t(f) \geq 0$. (*preserva positividad*)

v) Para cada $t, s \geq 0$, $P_{t+s} = P_t \circ P_s$. (*propiedad de semigrupo*)

**Definición 1.3** (Definición 1.2.1 de Bakry y col., 2013) *(Medida Invariante) Dada una familia de operadores* $\mathbf{P} = (P_t)_{t\geq 0}$ *que cumplan las condiciones* $(i) - (v)$ *anteriores, sobre un espacio medible* $(X, \mathcal{T})$*, una medida* $\sigma$*-finita* $\mu$ *sobre* $(X, \mathcal{T})$ *se dirá invariante para* $\mathbf{P}$ *si para cada función medible y acotada* $f : X \longmapsto \mathbb{R}$ *y cada* $t \geq 0$*,*

$$\int_X P_t f d\mu = \int_X f d\mu.$$

Sea $\mathbf{P} = (P_t)_{t\geq 0}$, una familia de operadores que satisfaga las condiciones $(i) - (v)$ anteriores. Supongamos que ahora posee una medida invariante $\mu$. Entonces,

vi) Para cada $f \in L^2(X, \mathcal{T}, \mu)$, $P_t\, f$ converge a $f$, en $L^2(X, \mathcal{T}, \mu)$, cuando $t \to 0$. (*propiedad de continuidad*).

vii) Para cada $1 \leq p < \infty$, el operador $P_t$ con $t \geq 0$, es una contracción en cada $L^p(X, \mathcal{T}, \mu)$.

(Ver Capítulo 1 de Bakry y col., 2013)

**Definición 1.4** (Definición 1.2.2 de Bakry y col., 2013) *Una familia de operadores* $\mathbf{P} = (P_t)_{t\geq 0}$ *definidas sobre funciones medibles y acotadas de* $(X, \mathcal{T})$ *con medida invariante* $\mu$*,* $\sigma$*-finita, que cumpla las condiciones desde* $(i)$ *hasta* $(vi)$ *se dirá* ***semigrupo markoviano****.*

### 1.2.2. Desigualdades logarítmicas de Sobolev y entropía

Para una medida $\mu$, no necesariamente finita sobre un espacio medible $(X, \mathcal{T})$, se define para todas las funciones positivas integrables $f$ tal que $\int_X f|\log f|d\mu < \infty$, la **entropía** de la función $f$, bajo la medida $\mu$, como

$$Ent_\mu(f) = \int_X f \log f \, d\mu - \left(\int_X f \, d\mu\right) \log \left(\int_X f \, d\mu\right). \tag{1.1}$$

Si $\mu$ es una medida de probabilidad, decimos que satisface la **desigualdad logarítmica de Sobolev**, de constante $c < \infty$, si

$$Ent_\mu(f^2) \leq c \int_X \|\nabla f\|^2 d\mu, \tag{1.2}$$

para toda función diferenciable $f : X \longmapsto \mathbb{R}$. Denotamos por $c_{LS} := c_{LS}(\mu)$ a la constante más pequeña que cumple esta desigualdad (ver Barthe y Strzelecki, 2021, Cattiaux y col., 2008 y Gross, 1993).

Consideremos la siguiente ecuación diferencial estocástica en $\mathbb{R}^n$, con $n \in \mathbb{N}$,

$$dX_t^a = b(X_t^a)dt + \sigma(X_t^a)dB_t, \quad X_0^a = a, \tag{1.3}$$

donde el super índice $a$ indica el punto inicial del proceso continuo $X_t$ y $(B_t)_{t\geq 0}$ es un movimiento Browniano estándar. La función $b$, conocida como *drift*, es una función a valores en $\mathbb{R}^n$, y $\sigma$, conocida como la *volatilidad*, una matriz de dimensiones $n \times n$. Ambas funciones son diferenciables.

Sea $\mathcal{B}_b(\mathbb{R}^n)$, el conjunto de funciones medibles y acotadas a valores en $\mathbb{R}^n$. Se define $(P_t)_{t\geq 0}$ como

$$(P_t\, f)(x) := \mathbb{E}\left[f(X_t^x)\right], \quad t \geq 0, \quad f \in \mathcal{B}_b(\mathbb{R}^n), \tag{1.4}$$

a la familia $(P_t)_{t\geq 0}$ es conocida como el semigrupo Markoviano asociado al proceso $X_t$.

En (1.4) la esperanza $\mathbb{E}$ es tomada con respecto a la ley de probabilidad de $X_t$. En el caso de que interactúen más variables aleatorias en las expresiones se agregará un super índice para indicar la ley de la variable aleatoria por la cual se está tomando esperanza, en el caso anterior sería $\mathbb{E}^{X_t}(\cdot)$. En caso de que la esperanza sea tomada con respecto a una ley conjunta que involucre a más de una variable, tan solo se agregarán las variables respectivas al super índice. Por ejemplo, si $h : \mathbb{R}^2 \longmapsto \mathbb{R}$ es un función medible, escribimos

$$\mathbb{E}^{X,Y}\left[h(X,Y)\right],$$

si $(X, Y)$ es un par aleatorio.

Los siguientes resultados intentan caracterizar la convergencia de las distribuciones de un subgrupo markoviano, cuyo caso aplica también para la distribución de un proceso de difusión como el presentado en la ecuación estocástica (1.3).

Sea $\mathbf{P} = (P_t)_{t\geq 0}$ un semigrupo markoviano, tal que $\mu$ sea su medida de probabilidad invariante.

**Teorema 1.2** (Teorema 5.2.1 en Bakry y col., 2013) *(Decaimiento exponencial en entropía) La desigualdad logarítmica de Sobolev (con constante $c_{LS}$) para la medida $\mu$ es equivalente a decir que para cada función positiva $f \in L^1(X, \mathcal{T}, \mu)$ (con entropía finita) cumple que*

$$Ent_\mu(P_t\, f) \leq e^{-2t/c_{LS}} Ent_\mu(f) \tag{1.5}$$

*para cada $t \geq 0$.*

Para dos medidas de probabilidad $\mu$ y $\nu$ se puede definir la **entropía relativa** o **divergencia de Kullback-Lieber** (ver Raginsky y col., 2017, Borkar y Mitter, 1999 y Cattiaux y col., 2008).

**Definición 1.5** *Se define la divergencia de Kullback-Liebler o entropía relativa (divergencia*

*KL) de una medida de probabilidad $\mu$ con respecto a otra, $\nu$, como*

$$D(\nu \,\|\, \mu) = \begin{cases} \int \log\left(\frac{d\nu}{d\mu}\right) d\nu, & \nu \ll \mu \\ \\ +\infty & \text{otro caso.} \end{cases} \tag{1.6}$$

*Si $\mu$ y $\nu$ son medidas en $\mathbb{R}^d$ y, suponiendo que existen funciones, $q(x)$ y $p(x)$, densidades de probabilidad respecto a la medida de Lebesgue en $\mathbb{R}^d$ de las medidas $\mu$ y $\nu$, respectivamente, entonces podemos escribir a divergencia KL de la forma*

$$D(\nu \,\|\, \mu) = \int_{\mathbb{R}^d} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx \tag{1.7}$$

Como $\nu \ll \mu$, tomando $f = \frac{d\nu}{d\mu}$, la derivada de Radon-Nikodym, tenemos que $D(\nu \,\|\, \mu) = Ent_\mu(f)$. Entonces, el Teorema 1.2 puede ser interpretado como

$$D(\nu_t \,\|\, \mu) \leq e^{-2t/c_{LS}} D(\nu_0 \,\|\, \mu), \tag{1.8}$$

donde $d\nu_t = P_t\, f\, d\mu$, $t \geq 0$.

Sea $\mathcal{P}(X) := \mathcal{P}(X, \mathcal{T})$ el espacio de medidas de probabilidad del espacio medible $(X, \mathcal{T})$, entonces la variación total entre medidas de probabilidad, definida como

$$\|\mu - \nu\|_{TV} := \sup_{A \in \mathcal{T}} |\mu(A) - \nu(A)|,$$

es una distancia en $\mathcal{P}(X)$. Además, se cumple la desigualdad de *Pinsker-Csizsár-Kullback*,

$$\|\mu - \nu\|_{TV}^2 \leq \frac{1}{2} D(\nu \,\|\, \mu).$$

Por lo tanto, bajo control de $D(\nu_0 \,\|\, \mu)$, el Teorema 1.2 (junto a (1.8)) implica una convergencia fuerte de $\nu_t$ hacia $\mu$ en variación total. (Bakry y col., 2013, capítulo 5.2)

## 1.3. Elementos de transporte óptimo

**Definición 1.6** *(Acoplamiento) Un acoplamiento (o Coupling) entre dos medidas de probabilidad en un espacio medible $(\mathcal{X}, \mathcal{T})$, $\mu$ y $\nu$, es otra medida $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{X})$, tal que sus leyes marginales sean $\mu$ y $\nu$ (ver Bakry y col., 2013, capítulo 9).*

La forma principal de medir la convergencia en leyes de los procesos analizados en este trabajo será mediante la distancia de 2-*Wasserstein*. Para dos medidas de probabilidad $\mu, \nu \in \mathcal{P}(\mathcal{X})$, se define la distancia 2-*Wasserstein* como

$$\mathcal{W}(\mu, \nu) = \inf\left\{ \left( \mathbb{E}^{X,Y}\left[ \|X - Y\|^2 \right] \right)^{1/2} \,:\, \mu = \mathcal{L}(X),\, \nu = \mathcal{L}(Y) \right\}, \tag{1.9}$$

donde el ínfimo se toma sobre todos los "acoplamientos" entre $\mu$ y $\nu$, ínfimo que siempre se puede alcanzar por el hecho de que el conjunto de los acoplamientos resulta ser un conjunto relativamente compacto al interior de las medidas conjuntas $\mathcal{P}(\mathcal{X} \times \mathcal{X})$. La Proposición 9.1.2 de Villani, 2003 nos asegura la existencia de acoplamiento óptimo para cuales quiera dos leyes

en $\mathcal{P}(\mathcal{X})$. Cada vez que sólo dos leyes, digamos $\mathcal{L}(X)$ y $\mathcal{L}(Y)$ para $X$ e $Y$ v.a.'s, participen por separados en una expresión, como por ejemplo en

$$\mathbb{E}^X(g(X)) - \mathbb{E}^Y(f(Y)),$$

para $g$ y $f$ funciónes medibles en el espacio respectivo, entonces, sin pérdida de generalidad, podemos generar el espacio adecuado para que cada una de esas esperanzas están tomadas mediante la ley marginal del acoplamiento óptimo que genera la distancia de Wasserstein. Es decir,

$$\mathbb{E}^X(g(X)) - \mathbb{E}^Y(f(Y)) = \mathbb{E}^{X,Y}(g(X) - f(Y)), \tag{1.10}$$

con

$$\sqrt{\mathbb{E}^{X,Y}(\|X - Y\|^2)} = \mathcal{W}(\mathcal{L}(X), \mathcal{L}(Y)). \tag{1.11}$$

Aunque la expresión (1.10) se cumpla para cualquier ley conjunta entre $X$ e $Y$, podemos asegurar la expresión (1.11), lo que representa una herramienta de gran utilidad para los desarrollos futuros.

En el Teorema 7.3 de Villani, 2003, la distancia de Wasserstein representa una métrica en el espacio de las medidas de probabilidad, $\mathcal{P}(X)$. Más aún, esta distancia caracteriza la convergencia débil de medidas de probabilidad en el espacio $\mathcal{P}_2(X)$, el espacio de las medidas de probabilidad con segundo momento finito, mediante el siguiente teorema.

**Teorema 1.3** (Teorema 7.12 de Villani, 2003) *Sea $(\mu_k)_{k\in\mathbb{N}}$ una sucesión de medidas de probabilidad en $\mathcal{P}_2(X)$, y sea $\mu \in \mathcal{P}_2(X)$. Entonces, las siguientes proposiciones son equivalentes:*

*i)* $\mathcal{W}(\mu_k, \mu) \to 0$, *cuando* $k \to \infty$.

*ii)* $\mu_k \to \mu$ *cuando* $k \to \infty$, *en el sentido débil, y* $(\mu_k)_{k\in\mathbb{N}}$ *satisface la siguiente condición de* ***tensión****: Para todo* $x_0 \in X$

$$\lim_{R\to 0} \limsup_{k\to\infty} \int_{\|x-x_0\|\geq R} \|x - x_0\|^2 d\mu_k(x) = 0. \tag{1.12}$$

*iii)* $\mu_k \to \mu$ *cuando* $k \to \infty$, *en el sentido débil, y existe convergencia de los momentos de orden 2: Para cualquier* $x_0 \in X$,

$$\int \|x - x_0\|^2 d\mu_k(x) \longrightarrow \int \|x - x_0\|^2 d\mu(x), \tag{1.13}$$

*cuando* $k \to \infty$.

*iv) Cuando una función* $\varphi$ *definida en* $X$ *satisface la condición de crecimiento,* $|\varphi(x)| \leq C(1 + \|x - x_0\|^2)$, *para algún* $x_0 \in X$, $C \in \mathbb{R}$. *Entonces*

$$\int \varphi d\mu_k \longrightarrow \int \varphi d\mu, \tag{1.14}$$

*cuando* $k \to \infty$.

Finalmente, se enuncia la desigualdad de trasporte de costo cuadrático que vincula la distancia de Wasserstein con la Definición 1.5, de la Sección 1.2.2.

**Definición 1.7** (Definición 9.2.2 de Bakry y col., 2013) *Decimos que $\mu \in \mathcal{P}_2(X)$ satisface la desigualdad de trasporte de costo cuadrático (desde ahora tan sólo desigualdad de trasporte) con constante $C > 0$, si para cada $\nu \in \mathcal{P}(X)$,*

$$\mathcal{W}(\mu, \nu)^2 \leq 2C\, D(\nu \,\|\, \mu). \tag{1.15}$$

**Ejemplo** (Desigualdad de Talagrand) Sea $d\mu(x) = (2\pi)^{-n/2} e^{-|x|^2/2} dx$, la medida Gaussiana estándar en los conjuntos borelianos de $\mathbb{R}^n$. Entonces, por el Teorema 9.2.1 de Bakry y col., 2013, para cada $\nu \in \mathcal{P}(\mathbb{R}^n)$, tenemos que

$$\mathcal{W}(\mu, \nu)^2 \leq 2\, D(\nu \,\|\, \mu),$$

es decir, la ley gaussiana cumple la desigualdad de transporte de la Definición 1.7 con constante igual a 1.

# Capítulo 2

# Problema de riesgo esperado y de máxima verosimilitud

## 2.1. Función de riesgo y riesgo empírico

Consideremos el siguiente problema de optimización

$$\underset{x\in\mathbb{R}^d}{\textbf{minimize}}\ \ F(x) := \mathbb{E}^Z[f(x,Z)] = \int f(x,z)\mathbb{P}(dz), \tag{2.1}$$

donde $x \in \mathbb{R}^d$ y $Z$ es una variable aleatoria fija definida en algún espacio de probabilidades $(\Omega, \mathcal{F}, \mathbb{P})$ y que toma valores en un espacio $\mathcal{Z}$ al que se llamará *espacio muestral*. La función $f : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ es $\mathscr{B}(\mathbb{R}\times\mathcal{Z})$-medible y es conocida como la *función de riesgo* del problema.

En el supuesto de que la medida de probabilidad $\mathbb{P}$ es desconocida, el problema (2.1) se vuelve infactible, o su tratamiento estará sujeto a aproximaciones o estimaciones de la función $F(\cdot)$. Si se supone que la variable $Z$ es observable y se considera $\mathbf{z} = (z_i)_{i=1}^n \in \mathcal{Z}^n$, una muestra aleatoria simple de tamaño $n$ de esta variable, entonces gracias a la LGN (Teorema 1.1), la función

$$F_{\mathbf{z}}(x) = \frac{1}{n}\sum_{i=1}^{n} f(x, z_i)$$

llamada *riesgo empírico*, es un aproximador (o "estimador") de la función $F(\cdot)$, en el sentido de que converge a ella cuando $n \to \infty$.

Así, el problema a resolver se suele reemplazar por

$$\underset{x\in\mathbb{R}^d}{\textbf{minimize}}\ \ F_{\mathbf{z}}(x). \tag{2.2}$$

Investigaciones y resultados indican que la resolución del problema (2.2) entrega usualmente un buen estimador del valor óptimo del problema (2.1) (ver capítulo 3 de Vapnik, 2000). Esta metodología de resolución de problemas estocásticos explica cómo algunos sistemas están dotados de la capacidad del aprendizaje a base de una muestra finita observaciones. (ver Bottou, 1998,Vapnik, 1979)

## 2.2. Problema de máxima verosimilitud

Un ejemplo simple de un problema de riesgo empírico esperado es el problema de máxima verosimilitud. En la siguiente sección se definen los elementos de este problema. (Ver Held y Bové, 2013, Capítulo 2)

### 2.2.1. Verosimilitud y Log-Verosimilitud

Sea $\mathcal{D}$ un conjunto de realizaciones del vector aleatorio $Z$ con masa de probabilidad o función de densidad conjunta $p_\theta(\mathcal{D})$. La función $p_\theta(\mathcal{D})$ depende de las realizaciones $z \in \mathcal{D}$ y de un parámetro, usualmente desconocido, $\theta$. El conjunto de todas las posibles realizaciones de la variable $Z$ es el *espacio muestral* $\mathcal{Z}$, que en este caso se considera un subconjunto de $\mathbb{R}^l$ para $l \in \mathbb{N}$, mientras que el conjunto de todos los valores que $\theta$ puede tomar, será llamado *espacio de parámetros*, denotado por $\Theta$. Típicamente, $\Theta$ será un subconjunto de algún espacio de dimensión finita, en este caso se considerará a $\mathbb{R}^n$ con $n \in \mathbb{N}$.

El objetivo de la inferencia estadística es inferir $\theta$ a partir de un conjunto de valores observados $\mathcal{D}$. De esta forma se define la *función de verosimilitud*

$$L(\theta; \mathcal{D}) = p_\theta(\mathcal{D}), \quad \theta \in \Theta, \tag{2.3}$$

visto como función de $\theta$ para $\mathcal{D}$ fijo. Usualmente escribimos $L(\theta)$ para la verosimilitud si no cabe duda de la dependencia de $\mathcal{D}$.

**Definición 2.1** (Definición 2.1 en Held y Bové, 2013) *La función de verosimilitud $L(\theta)$ es la masa de probabilidad conjunta o función de densidad conjunta del conjunto de datos observados $\mathcal{D}$, visto como una función del parámetro desconocido $\theta$.*

### 2.2.2. Estimador de máxima verosimilitud

**Definición 2.2** (Definición 2.2 de Held y Bové, 2013) *El estimador de máxima verosimilitud o estimador máximo verosímil (EMV) $\hat{\theta}_{EMV}$ del parámetro $\theta$ es aquel que maximiza la función de verosimilitud*

$$\hat{\theta}_{EMV} = \arg\max_{\theta \in \Theta} L(\theta) \tag{2.4}$$

Para calcular el estimador EMV, es posible ignorar valores constantes que multipliquen a la función $L(\theta)$ sin cambiar el punto óptimo, sin embargo, se verá que esta facilidad de ponderar la función de verosimilitud sin cambiar el valor del estimador EMV será de utilidad para reformular el problema de optimización respectivo a un problema de riesgo esperado.

En ocasiones, es numéricamente conveniente el uso de la *función de log-verosimilitud*

$$l(\theta) = \log L(\theta),$$

el logaritmo natural de la verosimilitud, para el cálculo el estimador EMV. Como el logaritmo es una función creciente, entonces

$$\hat{\theta}_{EMV} = \arg\max_{\theta \in \Theta} l(\theta).$$

Además, se puede observar que el estimador sigue siendo el mismo cuando se tienen funciones

iguales salvo ponderación por valores constantes positivos. Si el valor ponderante es negativo el problema es equivalente a uno de minimización, sin embargo, el punto óptimo se mantiene.

### 2.2.3. Formulación del problema

Sea $\mathcal{D} = \{z_i\}_{i=1}^n$ una M.A.S. de la variable $Z$. Además se asume que existe $\theta^* \in \Theta$ tal que la distribución de $Z$ corresponde a una ley **con densidad** $p_{\theta^*}$, perteneciente a una familia paramétrica de densidades $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$. Dado que la muestra es independiente, se puede escribir la densidad conjunta de la muestra como

$$p_\theta(\mathcal{D}) = \prod_{i=1}^{n} p_\theta(z_i), \tag{2.5}$$

por lo tanto, la función de log-verosimilitud es

$$l(\theta) = \sum_{i=1}^{n} \log p_\theta(z_i). \tag{2.6}$$

Como la ponderación de la función $l(\theta)$ no afecta al punto máximo, el problema de encontrar el estimador de máxima verosimilitud $\hat{\theta}_{EMV}$ puede escribirse como

$$\underset{\theta\in\Theta}{\textbf{minimize}} \;\; -\frac{1}{n}\sum_{i=1}^{n} \log p_\theta(z_i). \tag{2.7}$$

Este problema es equivalente al problema de riesgo empírico (2.2) tomando $f(\theta, z_i) = -\log p_\theta(z_i)$.[2] De la misma forma, tomando $n$ suficientemente grande, gracias a la ley de los grandes números se tiene que la función objetivo converge a una esperanza dependiente de la *verdadera distribución de la muestra*, caracterizada por el parámetro $\theta^*$.

$$\frac{1}{n}\sum_{i=1}^{n} \log p_\theta(z_i) \longrightarrow \mathbb{E}^Z_{\theta^*}\left(\log p_\theta(Z)\right) \;\text{ cuando }\; n \to \infty,$$

donde $\mathbb{E}^Z_{\theta^*}$ denota la esperanza de la función bajo la ley de probabilidad de $Z$ con densidad $p_{\theta^*}$. De esta forma, también se puede definir el problema del estimador de máxima verosimilitud como

$$\underset{\theta\in\Theta}{\textbf{minimize}} \;\; \mathbb{E}^Z_{\theta^*}\left(-\log p_\theta(Z)\right). \tag{2.8}$$

Esta última formulación equivale al planteamiento de problema de optimización estocástica presentado en el problema (2.1). El siguiente lema asegura que el problema (2.8) efectivamente se resuelve en el parámetro buscado.

**Lema 2.1** *Si $\theta^* \in \Theta$, entonces el mínimo del problema (2.8) se alcanza, al menos, en $\theta^*$. (Donde $\theta^*$ es el parámetros real, es decir $Z \sim p_{\theta^*}$)*

Demostración. El objetivo de esta demostración será probar que

$$\mathbb{E}^Z_{\theta^*}\left[\log p_\theta(Z)\right] \leq \mathbb{E}^Z_{\theta^*}\left[\log p_{\theta^*}(Z)\right] \quad \forall\, \theta \in \Theta. \tag{2.9}$$

[2] Acá se destaca la diferencia con respecto a la notación, donde la variable a optimizar $x$ es reemplazada por $\theta$ en la nueva formulación.

De esta forma, si $\theta^* \in \Theta$, entonces resuelve el problema (2.8) escrito en forma de maximización.

Notemos que

$$\mathbb{E}_{\theta^*}^Z \left[\log p_\theta(Z)\right] - \mathbb{E}_{\theta^*}^Z \left[\log p_{\theta^*}(Z)\right] = \mathbb{E}_{\theta^*}^Z \left[\log\left(\frac{p_\theta(Z)}{p_{\theta^*}(Z)}\right)\right] \leq \log\left(\mathbb{E}_{\theta^*}^Z \left[\frac{p_\theta(Z)}{p_{\theta^*}(Z)}\right]\right), \qquad (2.10)$$

donde la última desigualdad corresponde a la desigualdad de Jensen, utilizando la concavidad de la función logaritmo. Sin embargo, como $p_{\theta^*}$ es la densidad real de la distribución de $Z$, tenemos que

$$\mathbb{E}_{\theta^*}^Z \left[\frac{p_\theta(Z)}{p_{\theta^*}(Z)}\right] = \int \frac{p_\theta(z)}{p_{\theta^*}(z)} p_{\theta^*}(z)dz = \int p_\theta(z)dz \leq 1, \quad \forall\, \theta \in \Theta, \qquad (2.11)$$

dado que cada una de las funciones $p_\theta$ son densidades de probabilidad. De la expresión (2.11), cabe destacar que la integral puede estar definida sobre el soporte de la función $p_{\theta^*}$, es decir $\{z : p_{\theta^*}(z) > 0\}$, puesto que el valor esperado se toma sobre esta ley. Sin embargo, este conjunto no necesariamente corresponde al soporte de la función $p_\theta$, por lo que podría ser que la integral no alcanzara a completar la masa total de esta densidad y, por esta razón, la última parte corresponde a una desigualdad. Finalmente, juntando la desigualdad (2.10) y la expresión (2.11), tenemos que

$$\mathbb{E}_{\theta^*}^Z \left[\log p_\theta(Z)\right] - \mathbb{E}_{\theta^*}^Z \left[\log p_{\theta^*}(Z)\right] \leq 0, \quad \forall\, \theta \in \Theta, \qquad (2.12)$$

completando la demostración. □

No precisamente existe un único parámetro que alcance el óptimo, pues dos densidades que se diferencien tan sólo en un conjunto de medida nula podrían llegar de la misma manera al óptimo. Sin embargo, esto puede restringirse pidiendo que la familia $\mathcal{P}$ sea **identificable**, esto es, que el mapeo $\theta \longmapsto p_\theta$ sea inyectivo en $\Theta$.

Otra observación importante es que el problema (2.8) es independiente de cualquier muestra, por lo tanto (en caso de resolverse) una solución debiese ser el parámetro mismo y no tan sólo un estimador de éste. Finalmente, sólo queda aclarar de qué forma, o en qué nivel, el estimador de máxima verosimilitud $\hat{\theta}_{EMV}$ aproxima al valor real $\theta^*$.

**Definición 2.3** *Sea $\hat{\theta}_n$ un estimador de $\theta$, que depende de una muestra aleatoria $(Z_1, Z_2, \cdots, Z_n)$. Se dirá que $\hat{\theta}_n$ es un estimador* ***consistente*** *si $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta$, es decir*

$$\mathbb{P}\left(|\hat{\theta}_n - \theta| > \epsilon\right) \longrightarrow 0, \quad \forall\, \epsilon > 0,$$

*cuando $n \to \infty$.*

El siguiente lema conecta la idea de que, para muestras suficientemente grandes, el EMV entrega (con gran probabilidad) la densidad real que rige la ley de probabilidad de la muestra, bajo condiciones de regularidad recurrentemente usuales en el contexto de modelos paramétricos de estimación.

**Lema 2.2** (Lema 2.5 en Newey y McFadden, 1994) *Suponga que $z_i$, con $i \in \{1, 2, \cdots\}$ son muestras i.i.d. con densidad $p_{\theta^*}$, $\theta^* \in \Theta$, tal que se cumplan las siguientes condiciones:*

1. *La familia* $\mathcal{P} = \{p_\theta,\, \theta \in \Theta\}$ *es identificable.*

2. $\Theta$ *es compacto.*

3. *El mapeo* $\theta \mapsto \log p_\theta(z_i)$ *es continuo c.s. para todo* $z_i$*.*

4. $\mathbb{E}\left(\sup_{\theta\in\Theta} \log p_\theta(z_i)\right) < \infty$.

*Entonces el estimador máximo verosímil de* $\theta$ *es consistente.*

Esto asegura que los problemas (2.7) y (2.8) corresponden a los problemas propuestos en (2.2) y (2.1) (respectivamente), y por lo tanto; el problema del estimador de máxima verosimilitud corresponde a un problema de riesgo esperado (tanto teóricamente como en su versión empírica).

# Capítulo 3

# Algoritmo SGLD

Resolver (2.2) usando el método usual de descenso de gradiente requiere el cálculo de $n$ gradientes en cada actualización del nuevo valor (uno por cada $z_i$) y resulta un proceso muy caro cuando $n$ es un valor demasiado alto. Alternativamente a este enfoque, surge la formulación estocástica del algoritmo de gradiente, llamado descenso de gradiente estocástico (SGD, por su nombre en inglés). Para su formulación, es conveniente definir el *oráculo de gradiente estocástico* u *oráculo estocástico de primer orden*, siguiendo, por ejemplo, a Raginsky y col., 2017, Ghadimi y Lan, 2012 y Ghadimi y Lan, 2013.

**Definición 3.1** *Sea $g : \mathbb{R}^d \times \mathcal{U} \to \mathbb{R}^d$, una función medible, donde $\mathcal{U}$ es cierto espacio medible conveniente, y $U_{\mathbf{z}}$ una variable aleatoria en $\mathcal{U}$ que sólo depende de la muestra $\mathbf{z}$, bajo el contexto del problema (2.2). Se dirá que $g$ y $U_{\mathbf{z}}$ conforman un mecanismo de oráculo de gradiente estocástico de primer orden para $F_{\mathbf{z}}$ si*

$$\mathbb{E}^{U_{\mathbf{z}}}\left(g(x, U_{\mathbf{z}})\right) = \nabla F_{\mathbf{z}}(x), \quad \forall\, x \in \mathbb{R}^d \tag{3.1}$$

De esta forma, dado el vector de observaciones $\mathbf{z}$, el algoritmo SGD puede ser escrito de la forma

$$x_{k+1} = x_k - \eta g(x_k, U_{\mathbf{z},k}), \tag{3.2}$$

donde $\eta > 0$ es el parámetro de aprendizaje o tasa de aprendizaje y $g(x_k, U_{\mathbf{z},k})$ es un estimador insesgado de $\nabla F_{\mathbf{z}}(x_k)$ mediante el mecanismo de oráculo con $U_{\mathbf{z},k}$, realizaciones i.i.d. de $U_{\mathbf{z}}$. Esto puede escribirse en función del oráculo de gradiente estocástico como $g_k = g(x_k, U_{\mathbf{z},k})$ cuando no quepa duda de la dependencia de las variables $U_{\mathbf{z},k}$. Generalmente, la variable $U_{\mathbf{z}}$ aparecerá implícitamente en la práctica, pero su existencia engloba un amplio espectro de estimadores insesgados de $\nabla F_{\mathbf{z}}$.

Cuando la función de riesgo $f(x, Z)$ no es convexa en la variable $x$, el problema se dificulta considerablemente, por esta razón se introduce una variante del algoritmo SGD que incluya un término de error que permita al algoritmo evitar caer en (o poder salir de) mínimos locales. Tal variante es llamada Gradiente Estocástico de dinámica de Langevin (SGLD, por su nombre en inglés) y está dada por la recursión

$$x_{k+1} = x_k - \eta g_k + \sqrt{2\eta\beta^{-1}}\xi_k, \tag{3.3}$$

donde $\xi_k$ es una variable gaussiana estándar en $\mathbb{R}^d$ y $\beta > 0$ es el parámetro de temperatura.

El nombre de este algoritmo viene dado del hecho de que el proceso Markoviano discreto definido por (3.3) puede verse como la discretización de un proceso de difusión de Langevin a tiempo continuo, o una aproximación de este, dado por la ecuación diferencial estocástica de Itô

$$dX_t = -\nabla F_{\mathbf{z}}\left(X_t\right) dt + \sqrt{2\beta^{-1}} dB_t, \;\; t \geq 0, \tag{3.4}$$

con $\{B_t\}_{t\geq 0}$ un movimiento browniano estándar en $\mathbb{R}^d$. Según Raginsky y col., 2017,Chiang y col., 1987 y Ma y col., 2015, bajo ciertas condiciones de $f$, es posible demostrar que, para una realización $\mathbf{z}$ dada, la medida de Gibbs

$$\pi_{\mathbf{z}}(dx) \propto \exp\left(-\beta F_{\mathbf{z}}(x)\right)$$

es la única distribución invariante de (3.4) y que la distribución de $X_t$ converge a $\pi_{\mathbf{z}}$ cuando $t \to \infty$. Más aún, para valores suficientemente altos de $\beta$, la distribución $\pi_{\mathbf{z}}$ se tiende a concentrar en el conjunto de óptimos globales de $F_{\mathbf{z}}(x)$. (ver Raginsky y col., 2017, Chiang y col., 1987 y Hwang, 1980)

Denotaremos por $\mu_{\mathbf{z},k} = \mathcal{L}\left(x_k|\mathbf{Z} = \mathbf{z}\right)$ a la ley del proceso definido en (3.3) y por $\nu_{\mathbf{z},t} = \mathcal{L}\left(X_t|\mathbf{Z} = \mathbf{z}\right)$ a la ley de (3.4). La aproximación que tiene el algoritmo SGLD para lograr óptimos de los problemas (2.1) y (2.2) son diferentes en el sentido que todas las distribuciones de los procesos definidos ($\mu_{\mathbf{z},k}$, $\nu_{\mathbf{z},t}$ y $\pi_{\mathbf{z}}$) están construidos a partir de la función $F_{\mathbf{z}}$ y de la muestra $\mathbf{z}$, muy presentes en el problema (2.2), a diferencia del problema (2.1), que es independiente de la muestra. En la realidad, las aplicaciones se restringen directamente a una muestra dada, por lo que es más realista la formulación aproximada del problema que tomar el valor esperado de la ley desconocida. Por último, la diferencia entre los sesgos de los problemas (2.1) y (2.2) difieren tan sólo en un término proporcional a $1/n$, donde $n$ es la cantidad de la muestra (ver Raginsky y col., 2017), por lo que el enfoque completo se recupera con un tamaño suficientemente alto de la muestra.

# Capítulo 4

# Aproximación al óptimo

El objetivo de esta sección es evidenciar la aproximación del valor de salida del algoritmo $x_k$ y el punto óptimo de la función de riesgo esperado empírico $F_{\mathbf{z}}$, cuando $k \in \mathbb{N}$ es un valor suficientemente alto. El valor de $x_k$ es una variable aleatoria con ley $\mu_{\mathbf{z},k}$, la pregunta recae en: si $x^*$ es el valor mínimo de la función $F_{\mathbf{z}}$, entonces ¿qué tan cerca está $\mathbb{E}\left(F_{\mathbf{z}}(x_k)\right)$ de $F_{\mathbf{z}}(x^*)$, cuando la esperanza es tomada con respecto a la ley $\mu_{\mathbf{z},k}$?, ¿qué tan grande debe ser $k$ para que la aproximación sea buena?

Para ser realista con el contexto en que se desenvuelve este problema, se debe aclarar que muchos de los enfoques que se consideran para este análisis en general en la literatura son del tipo asintótico, cuando $t \to \infty$, puesto que la distribución de Gibbs, $\pi_{\mathbf{z}}$, es el límite débil de la variable $X_t$ definida en (3.4). Sin embargo, en la aplicación es imposible simular esta variable hasta lograr la convergencia, por lo tanto, se parte reconociendo que siempre existirá un error de estimación (el cual se busca cuantificar o justificar) y que toda aproximación se hará en un intervalo de tiempo finito, $[0, T_0]$, donde $T_0$ es de la forma $\eta K$, con $K \in \mathbb{N}$, la cantidad final de iteraciones del algoritmo SGLD.

Se impondrán las siguientes hipótesis sobre el conjunto de elementos del problema (2.2).

## 4.1. Condiciones requeridas

La primera condición tiene que ver con la suavidad de la función $F_{\mathbf{z}}$. Además de ser diferenciable para poder asegurar la existencia de su gradiente, pediremos que este gradiente sea una función acotada y de clase Lipschitz para cierta constante que no dependa de la realización $\mathbf{z}$.

**Condición 1** (**A.1**-**A.2** en Raginsky y col., 2017, **A.1** en Borkar y Mitter, 1999) Para cada $z \in \mathcal{Z}$, la función $\nabla f(\cdot, z)$ es $L$-Lipschitz. Es decir, existe una constante $L > 0$ fija, tal que

$$\|\nabla f(x, z) - \nabla f(v, z)\| \leq L\|x - v\|, \quad \forall\, x, v \in \mathbb{R}^d. \tag{4.1}$$

Además, la función $\nabla f(\,\cdot\,, z)$ es acotada en algún punto de $\mathbb{R}^d$. Es decir, existe $\tilde{x} \in \mathbb{R}^d$, tal que existe una constante $R > 0$ finita que cumpla que

$$\|\nabla f(\tilde{x}, z)\| \leq R, \quad \forall\, z \in \mathcal{Z}. \tag{4.2}$$

La siguiente condición es tomada por Raginsky y col., 2017 y juega un rol importante al momento de acotar la varianza de $x_k$.

**Condición 2** (**A.3** en Raginsky y col., 2017) La función $f(\cdot, z)$ es $(m, b)$-disipativa, para cada $z \in \mathcal{Z}$. Esto es; para algún $m > 0$ y $b \geq 0$,

$$\langle x, \nabla f(x, z) \rangle \geq m\|x\|^2 - b, \quad \forall\, x \in \mathbb{R}^d \forall\, z \in \mathcal{Z}. \tag{4.3}$$

Las siguientes condiciones propuestas en Raginsky y col., 2017 y Borkar y Mitter, 1999, son impuestas para restringir la elección del mecanismo de oráculo estocástico (definición 3.1).

**Condición 3** (**A.4** en Raginsky y col., 2017, **A.2 - (i)** en Borkar y Mitter, 1999) Existe una constante $\delta \in [0, 1)$, y dos constantes $G, N > 0$ finitas, tal que para todo $\mathbf{z} \in \mathcal{Z}^n$,

$$\mathbb{E}^{U_{\mathbf{z}}} \left( \|g(x, U_{\mathbf{z}}) - \nabla F_{\mathbf{z}}(x)\|^2 \right) \leq 2\delta \left( G^2\|x\|^2 + N^2 \right), \quad \forall\, x \in \mathbb{R}^d. \tag{4.4}$$

Finalmente, imponemos condiciones sobre la distribución del punto inicial del algoritmo y la *brecha espectral uniforme*.

**Condición 4** (**A.5** en Raginsky y col., 2017) La ley de probabilidad $\mu_0$ del punto inicial del algoritmo $x_0$, tiene densidad $p_0$ acotada y

$$\kappa_0 := \log \int_{\mathbb{R}^d} e^{\|x\|^2} p_0(x) dx < \infty. \tag{4.5}$$

**Definición 4.1** *Se define la brecha espectral uniforme como*

$$\lambda_* = \inf_{\mathbf{z} \in \mathcal{Z}^n} \inf \left\{ \frac{\int_{\mathbb{R}^d} \|\nabla g\|^2 d\pi_{\mathbf{z}}}{\int_{R^d} g^2 d\pi_{\mathbf{z}}} \,:\, g \in \mathcal{C}^1(\mathbb{R}^d) \cap L^2(\pi_{\mathbf{z}}),\, g \neq 0,\, \int_{\mathbb{R}^d} g d\pi_{\mathbf{z}} = 0 \right\} \tag{4.6}$$

Bajo ciertas condiciones es posible asegurar que $\lambda_* > 0$ (ver Raginsky y col., 2017, Sección 4).

## 4.2. Lemas útiles

Se introduce la siguiente notación: $\mathbb{E}_{\mathbf{z}}(\cdot) = \mathbb{E}(\cdot|\mathbf{Z} = \mathbf{z})$, ya que en el análisis general, se considera que la muestra aleatoria simple representa otra variable con la opción a cambiar en cada experimento independientemente de este algoritmo.

Los siguientes lemas serán de utilidad para la conclusión de esta sección.

**Lema 4.1** *Bajo la Condición 1, existe una constante finita $B$ tal que*

$$\|\nabla f(x, z)\| \leq L\|x\| + B, \tag{4.7}$$

*para todo $x \in \mathbb{R}^d$, y todo $z \in \mathcal{Z}$.*

Demostración. Simple consecuencia de la Condición 1 y de la desigualdad triangular de la norma en $\mathbb{R}^d$. En efecto, sea $\tilde{x} \in \mathbb{R}^d$, el valor mencionado en la Condición 1, en (4.2),

supongamos que $\|\tilde{x}\| = M$, se cumple que $\|\nabla f(\tilde{x}, z)\| \leq R$ para todo $z \in \mathcal{Z}$, entonces

$$\begin{aligned}\|\nabla f(x,z)\| &\leq \|\nabla f(x,z) - \nabla f(\tilde{x},z)\| + \|\nabla f(\tilde{x},z)\| \\ &\leq L\|x - \tilde{x}\| + R \\ &\leq L\|x\| + LM + R := L\|x\| + B,\end{aligned}$$

donde se define a $B = LM + R < \infty$. □

Recordemos que $X_t$ era el proceso definido por la ecuación diferencial estocástica

$$dX_t = -\nabla F_{\mathbf{z}}(X_t)dt + \sqrt{2\beta^{-1}}dB_t, \quad t \geq 0,$$

que, gracias a la Condición 1, nos asegura la existencia de una única solución fuerte para cada condición inicial, en el sentido de los procesos adaptados a la filtración canónica del movimiento Browniano $B_t$.

El siguiente lema es una modificación del Lema 3 en Raginsky y col., 2017.

**Lema 4.2** (Cota uniforme $L^2$ del algoritmo SGLD) *Bajo las Condiciones 1, 2, 3 y 4, para todo* $0 < \eta < 1 \wedge \frac{m}{2(L^2+G^2)}$ *y todo* $\mathbf{z} \in \mathcal{Z}^n$,

$$\sup_{k\geq 0} \mathbb{E}_{\mathbf{z}}^{x_k}\|x_k\|^2 \leq \kappa_0 + 2\left(1 \wedge \frac{1}{m}\right)\left(b + B^2 + N^2 + \frac{d}{\beta}\right) \tag{4.8}$$

*y*

$$\begin{aligned}\mathbb{E}_{\mathbf{z}}^{X_t}\|X_t\|^2 &\leq \kappa_0 e^{-2mt} + \tfrac{b+d/\beta}{m}(1 - e^{-2mt}) \\ &\leq \kappa_0 + \tfrac{b+d/\beta}{m}\end{aligned} \tag{4.9}$$

*donde B corresponde a la constante que nace del Lema 4.1 y la primera desigualdad de (4.9) se cumple para todo* $t > 0$.

Gracias a este último lema, podemos definir (Sección 3 en Borkar y Mitter, 1999)

$$\mathcal{K} = \sup_{k\geq 0} \mathbb{E}_{\mathbf{z}}^{x_k}\left[\|x_k\|^2\right], \tag{4.10}$$

y asegurar que $\mathcal{K} < \infty$, cuya utilidad será visualizada en los lemas posteriores.

**Lema 4.3** (Integrabilidad exponencial de la difusión, Lema 4 en Raginsky y col., 2017) *para todo* $\beta > 2/m$, *tenemos*

$$\log \mathbb{E}_{\mathbf{z}}^{X_t}\left[e^{\|X_t\|^2}\right] \leq \kappa_0 + 2\left(b + \frac{d}{\beta}\right)t \tag{4.11}$$

El siguiente lema es una consecuencia de la Condición 3 y el Lema 4.2.

**Lema 4.4** *Existe una constante positiva* $C < \infty$, *tal que*

$$\sup_k \mathbb{E}_{\mathbf{z}}^{x_k, U_{\mathbf{z},k}}\|g(x_k, U_{\mathbf{z},k}) - \nabla F_{\mathbf{z}}(x_k)\|^2 \leq C, \tag{4.12}$$

*donde, recordemos que, $x_k$ y $U_{\mathbf{z},k}$ con variables independientes.*

Demostración. Notemos que, condicionando al valor de $x_k$ tenemos que

$$\mathbb{E}_{\mathbf{z}}^{x_k,U_{\mathbf{z},k}}\|g(x_k,U_{\mathbf{z},k})-\nabla F_{\mathbf{z}}(x_k)\|^2=\mathbb{E}_{\mathbf{z}}^{x_k}\left(\mathbb{E}_{\mathbf{z}}^{U_{\mathbf{z},k}}\left(\|g(x_k,U_{\mathbf{z},k})-\nabla F_{\mathbf{z}}(x_k)\|^2\,|x_k\right)\right),$$

usando que la Condición 3 es para todo valor de $x\in\mathbb{R}^d$, y por la monotonía de la esperanza,

$$E_{\mathbf{z}}^{U_{\mathbf{z},k}}\left(\|g(x_k,U_{\mathbf{z},k})-\nabla F_{\mathbf{z}}(x_k)\|^2\,|x_k\right)\leq 2\delta(G^2\|x_k\|^2+N^2) \tag{4.13}$$

y, por lo tanto,

$$\mathbb{E}_{\mathbf{z}}^{x_k,U_{\mathbf{z},k}}\|g(x_k,U_{\mathbf{z},k})-\nabla F_{\mathbf{z}}(x_k)\|^2\leq\mathbb{E}_{\mathbf{z}}^{x_k}\left(2\delta(G^2\|x_k\|^2+N^2)\right). \tag{4.14}$$

Definiendo

$$C:=2\delta(G^2\mathcal{K}+N^2),$$

con $\mathcal{K}$ dada por (4.10), se concluye el resultado. □

**Observación:** Un aspecto a destacar del lema anterior es que la constante encontrada es independiente de $\eta$, implicando una cota uniforme en este parámetro.

Por último, se define una variación del proceso utilizado en Borkar y Mitter, 1999, el proceso continuo $x(t)$, dado para $t>0$ por

$$x(t):=x_0-\int_0^t g(x(\lfloor s/\eta\rfloor\eta),U_{\mathbf{z},\lfloor s/\eta\rfloor})ds+\sqrt{2\beta^{-1}}B_t, \tag{4.15}$$

donde $(B_t)_{t\geq 0}$ es el mismo movimiento browniano estándar en $\mathbb{R}^d$ que aparece en la definición de $X_t$, en la Ecuación (3.4). Gracias a los crecimientos Brownianos y a la composición de $x(t)$, notamos que $x(k\eta)$ y $x_k$ son iguales en ley para todo $k\in\mathbb{N}$. En efecto, para $\eta>0$ y $k\in\mathbb{N}$;

$$x\left((k+1)\eta\right)=x_0-\int_0^{(k+1)\eta}g(x(\lfloor s/\eta\rfloor\eta),U_{\mathbf{z},\lfloor s/\eta\rfloor})ds+\sqrt{2\beta^{-1}}B_{(k+1)\eta}, \tag{4.16}$$

$$=x_0-\int_0^{k\eta}g(x(\lfloor s/\eta\rfloor\eta),U_{\mathbf{z},\lfloor s/\eta\rfloor})ds-\int_{k\eta}^{(k+1)\eta}g(x(\lfloor s/\eta\rfloor\eta),U_{\mathbf{z},\lfloor s/\eta\rfloor})ds+\sqrt{2\beta^{-1}}B_{(k+1)\eta}, \tag{4.17}$$

como $x(\lfloor s/\eta\rfloor\eta)=x(k\eta)$ para todo $s\in[k\eta,(k+1)\eta)$, y reescribiendo

$$B_{(k+1)\eta}=B_{k\eta}+\left(B_{(k+1)\eta}-B_{k\eta}\right),$$

tenemos que

$$\begin{aligned}x\left((k+1)\eta\right)&=x_0-\int_0^{k\eta}g(x(\lfloor s\rfloor),U_{\mathbf{z},\lfloor s/\eta\rfloor})ds+\sqrt{2\beta^{-1}}B_{k\eta}\\&\quad+\left(-\eta g(x(k\eta),U_{\mathbf{z},k})+\sqrt{2\beta^{-1}}(B_{(k+1)\eta}-B_{k\eta})\right),\end{aligned}$$

es decir;

$$x((k+1)\eta)=x(k\eta)-\eta g(x(k\eta),U_{\mathbf{z},k})+\sqrt{2\beta^{-1}}(B_{(k+1)\eta}-B_{k\eta}). \tag{4.18}$$

Finalmente, gracias a las propiedades de saltos independientes y reescalamiento del movimiento Browniano, tenemos que $B_{(k+1)\eta}-B_{k\eta}\overset{d}{=}B_\eta\overset{d}{=}\sqrt{\eta}\hat{\xi}$, donde $\hat{\xi}$ corresponde a una

variable normal estándar multivariada $d$-dimensional e independiente al pasado del movimiento Browniano hasta el tiempo $k\eta$, y el símbolo $\stackrel{d}{=}$ indica que la igualdad corresponde a una igualdad en distribución. Entonces, para los tiempos de la forma $k\eta$ tenemos la siguiente recurrencia

$$x((k+1)\eta) \stackrel{d}{=} x(k\eta) - \eta g(x(k\eta), U_{\mathbf{z},k}) + \sqrt{2\eta\beta^{-1}}\hat{\xi}, \tag{4.19}$$

que corresponde a la misma recurrencia de la definición de $x_k$ en (3.3). Añadiendo la misma condición inicial a ambos procesos y dado que los $\xi_k$ son independientes podemos concluir que tanto $x(k\eta)$ como $x_k$ son iguales como proceso, es decir $(x(k\eta) : k \in \mathbb{N}) \stackrel{d}{=} (x_k : k \in \mathbb{N})$. El proceso $x(t)$ por lo tanto corresponde a una extensión continua, en todo punto, de las realizaciones del algoritmo SGLD. Por lo tanto, debería existir convergencia en alguna norma cuando la discretización del tiempo sea lo suficientemente fina. Esto se prueba en el siguiente lema.

**Lema 4.5** $\mathbb{E}^{x(t),x_{\lfloor t/\eta \rfloor}}\left(\|x(t) - x_{\lfloor t/\eta \rfloor}\|^2\right) \to 0$ *cuando* $\eta \to 0$*, para todo* $t > 0$.

Demostración. Sea $t > 0$, $0 < \eta$ y $k \in \mathbb{N}$ tal que $k\eta \leq t < (k+1)\eta$, tenemos que

$$\begin{aligned} x(t) &= x_0 - \int_0^t g(x(\lfloor s/\eta \rfloor \eta), U_{z,\lfloor s/\eta \rfloor})ds + \sqrt{2\beta^{-1}}B_t, \\ &= x_0 - \int_0^{k\eta} g(x(\lfloor s/\eta \rfloor \eta), U_{z,\lfloor s/\eta \rfloor})ds - \int_{k\eta}^t g(x(\lfloor s/\eta \rfloor \eta), U_{z,\lfloor s/\eta \rfloor})ds + \sqrt{2\beta^{-1}}\left(B_{k\eta} + (B_t - B_{k\eta})\right), \\ &= x(k\eta) - \int_{k\eta}^t g(x(\lfloor s/\eta \rfloor \eta), U_{\mathbf{z},\lfloor s/\eta \rfloor})ds + \sqrt{2\beta^{-1}}(B_t - B_{k\eta}). \end{aligned}$$

Nuevamente, ocupando las propiedades del movimiento Browniano, tenemos que $B_t - B_{k\eta} \stackrel{d}{=} \sqrt{t - k\eta}\ \hat{\xi}_k$, con $\hat{\xi}_k \sim \mathcal{N}(\vec{0}, \mathbb{I})$. Así, recordando que $x(k\eta) \stackrel{d}{=} x_k = x_{\lfloor t/\eta \rfloor}$, tenemos que

$$x(t) - x_{\lfloor t/\eta \rfloor} \stackrel{d}{=} -\int_{k\eta}^t g(x(\lfloor s \rfloor), U_{\mathbf{z},\lfloor s/\eta \rfloor})ds + \sqrt{2\beta^{-1}(t - k\eta)}\hat{\xi}_k. \tag{4.20}$$

Como $k\eta \leq t < (k+1)\eta$, entonces

$$\int_{k\eta}^t g(x(\lfloor s/\eta \rfloor \eta), U_{\mathbf{z},\lfloor s/\eta \rfloor})ds = (t - k\eta)g(x(k\eta), U_{\mathbf{z},k}). \tag{4.21}$$

Reemplazando en la ecuación (4.20) tenemos que

$$\|x(t) - x_{\lfloor t/\eta \rfloor}\|^2 \leq (t - k\eta)^2\|g(x(k\eta), U_{\mathbf{z},k})\|^2 + |2\beta^{-1}(t - k\eta)|\|\hat{\xi}_k\|^2. \tag{4.22}$$

Por la definición de $k$ tenemos que $t - k\eta \leq \eta$. Además, gracias a los lemas 4.1, 4.2 (definición 4.10) y 4.4, tenemos la siguiente cota

$$\mathbb{E}_{\mathbf{z}}^{x_k,U_{\mathbf{z},k}}\|g(x(k\eta), U_{\mathbf{z},k})\|^2 \leq 2(C + 2L^2\mathcal{K}^2 + 2B^2). \tag{4.23}$$

Así, como $\mathbb{E}\left[\|\hat{\xi}_k\|^2\right] = d$, tomando esperanza a ambos lados de (4.22) y acotando, tenemos que

$$\mathbb{E}_{\mathbf{z}}^{x(t),x_{\lfloor t/\eta \rfloor}}\left(\|x(t) - x_{\lfloor t/\eta \rfloor}\|^2\right) \leq 2(C + 2L^2\mathcal{K}^2 + 2B^2)\eta^2 + 2\beta^{-1}d\eta. \tag{4.24}$$

Basta con notar que el lado derecho de (4.24) sólo depende de $\eta$, y tiende a cero cuando $\eta \to 0$. El resultado se mantiene cuando la esperanza es tomada sobre la distribución conjunta que incluye a la variable $\mathbf{Z}$, que fue condicionada en esta demostración. □

También es posible ver al proceso $x(t)$ como el proceso continuo que aproxima la parte de variación acotada del proceso (3.4) (integral) mediante un proceso simple y replica la parte correspondiente a la martingala (Browniano). De esta forma, $x(t)$ toma el rol de proceso intermedio entre el algoritmo SGLD $x_{\lfloor t/\eta \rfloor}$ y el proceso de difusión $X_t$.

**Lema 4.6** (Lema 3.1 en Borkar y Mitter, 1999) $\forall T_0 > 0, \forall\, t \in (0, T_0]$, $\mathbb{E}_{\mathbf{z}}^{x(t),X_t}\left[\|x(t) - X_t\|^2\right] \to 0$, *cuando* $\eta \to 0$. *Donde* $x(t)$ *el proceso descrito en (4.15).*

En Borkar y Mitter, 1999 además es posible encontrar la siguiente desigualdad para la diferencia cuadrática entre $x(t)$ y $X_t$, para $t \in (0, T_0)$,

$$\mathbb{E}_{\mathbf{z}}^{x(t),X_t}\left[\|x(t) - X_t\|^2\right] \leq L \int_0^{T_0} \mathbb{E}_{\mathbf{z}}^{x(s),X_t}\left[\|x(s) - X_s\|^2\right] ds + \eta^2 T_0 \left(L\mathcal{K} + C + \frac{2}{\beta}\right) L + \eta(T_0 + 1)\left(C + \frac{2b}{\beta}\right).$$

Usando el lema de Gronwall se encuentra la siguiente cota:

$$\mathbb{E}_{\mathbf{z}}^{x(t),X_t}\left[\|x(t) - X_t\|^2\right] \leq \left(\eta^2 T_0 \left(L\mathcal{K} + C + \frac{2}{\beta}\right) L + \eta(T_0 + 1)\left(C + \frac{2b}{\beta}\right)\right) e^{LT_0}. \tag{4.25}$$

Por último, es posible notar que, tanto el Lema 4.5 como el Lema 4.6, no dependen de los acoplamientos entre $x_{\lfloor t/\eta \rfloor}$ y $x(t)$, o entre $x(t)$ y $X_t$, por lo tanto, esta conclusión también puede expresarse en términos de la métrica de Wasserstein. En particular, se tiene que cada tiempo de la dinámica $X_t$, de ley $\nu_{\mathbf{z},t}$, se aproxima en distribución por algún paso del algoritmo SGLD de la forma $x_{\lfloor t/\eta \rfloor}$, de ley $\mu_{\mathbf{z},\lfloor t/\eta \rfloor}$, cuando $\eta \to 0$, si $t \in (0, T_0)$.

Esto se refleja en el siguiente corolario, que surge como consecuencia de los lemas 4.5 y 4.6.

**Corolario 4.1** $\forall\, t \in (0, T_0)$, $\mathcal{W}(\mu_{\mathbf{z},\lfloor t/\eta \rfloor}, \nu_{\mathbf{z},t}) \to 0$, *cuando* $\eta \to 0$. *Más aún, tenemos la siguiente cota para la distancia de Wasserstein entre* $\mu_{\mathbf{z},\lfloor t/\eta \rfloor}$ *y* $\nu_{\mathbf{z},t}$*:*

$$\mathcal{W}(\mu_{\mathbf{z},\lfloor t/\eta \rfloor}, \nu_{\mathbf{z},t})^2 \leq \mathcal{Q}_1 \eta^2 + \mathcal{Q}_2 \eta, \tag{4.26}$$

*donde*

$$\mathcal{Q}_1 = 2\left(T_0 \left(L\mathcal{K} + C + \frac{2}{\beta}\right) L e^{LT_0} + 2(C + 2L^2\mathcal{K}^2 + 2B^2)\right), \tag{4.27}$$

$$\mathcal{Q}_2 = 2\left((T_0 + 1)\left(C + \frac{2b}{\beta}\right) e^{LT_0} + 2\beta^{-1} d\right). \tag{4.28}$$

Demostración. Sea $\mathbb{E}^{x_{\lfloor t/\eta \rfloor}, X_t}(\,\cdot\,)$ el valor esperado bajo la ley conjunta entre $x_{t/\eta \rfloor}$ y $X_t$ que induce la distancia de Wasserstein. Ocupando desigualdad triangular de la norma Euclideana

tenemos que

$$\mathcal{W}(\mu_{\mathbf{z},\lfloor t/\eta \rfloor}, \nu_{\mathbf{z},t})^2 = \mathbb{E}^{x_{\lfloor t/\eta \rfloor}, X_t}\left(\|x_{\lfloor t/\eta \rfloor} - X_t\|^2\right)$$

$$\leq 2\left(\mathbb{E}^{x_{\lfloor t/\eta \rfloor}, x(t)}\left(\|x_{\lfloor t/\eta \rfloor} - x(t)\|^2\right) + \mathbb{E}^{x(t), X_t}\left(\|x(t) - X_t\|^2\right)\right).$$

Ocupando las cotas (4.24) y (4.25), y reagrupando términos, se concluye el resultado. □

**Observación**: Es necesario mencionar la importancia de considerar la dinámica en un intervalo finito de tiempo, puesto que el factor exponencial $e^{LT_0}$, en (4.28), no le permite a $\eta$ disminuir el valor de $\mathbb{E}_{\mathbf{z}}^{x(t),X_t}[\|x(t) - X_t\|^2]$ si $T_0 \to \infty$, como se aprecia también en (4.25). Este punto será de importancia en el análisis final de la aproximación de la solución, puesto que uno de los errores relacionados al error cuadrático del algoritmo depende de horizonte temporal en que se busca discretizar el proceso continuo. Por último, otro punto importante de mencionar es la dependencia del parámetro $\beta$, cuyo valor usualmente es alto, sin embargo, por muy grande que sea, tanto la cota como la conclusión del lema anterior siguen siendo válidas.

Para comparar el valor de $F_{\mathbf{z}}$ obtenido con el algoritmo SGLD y su valor óptimo ocuparemos la siguiente descomposición: para $t \in (0, T_0)$

$$\mathbb{E}^{x_{\lfloor t/\eta \rfloor}}\left(F_{\mathbf{z}}(x_{\lfloor t/\eta \rfloor})\right) - F_{\mathbf{z}}^* = \mathbb{E}^{x_{\lfloor t/\eta \rfloor}, X_t}\left(F_{\mathbf{z}}(x_{\lfloor t/\eta \rfloor}) - F_{\mathbf{z}}(X_t)\right) +$$

$$\mathbb{E}^{X_t, X^*}\left(F_{\mathbf{z}}(X_t) - F_{\mathbf{z}}(X^*)\right) + \mathbb{E}^{X^*}\left(F_{\mathbf{z}}(X^*) - F_{\mathbf{z}}^*\right),$$

donde $F_{\mathbf{z}}^*$ es el valor óptimo de la función $F_{\mathbf{z}}(x)$, $x_{\lfloor t/\eta \rfloor}$ iteración del algoritmo SGLD hasta el último entero menor a $t/\eta$, distribuido como $\mu_{\mathbf{z},x_{\lfloor t/\eta \rfloor}}$, $X_t$ es la variable aleatoria resultante de la dinámica (3.4) en el tiempo $t$, distribuido como $\nu_{\mathbf{z},t}$ y $X^*$ es una variable aleatoria distribuida por la medida de Gibbs, $\pi_{\mathbf{z}}$. La descomposición anterior analiza tres términos; el primero, $\mathbb{E}^{x_{\lfloor t/\eta \rfloor}, X_t}\left(F_{\mathbf{z}}(x_{\lfloor t/\eta \rfloor}) - F_{\mathbf{z}}(X_t)\right)$, corresponde al error de aproximación desde el algoritmo SGLD al proceso de difusión y será principalmente controlado por la convergencia débil que existe desde $\mu_{\mathbf{z},\lfloor t/\eta \rfloor}$ hacia $\nu_{\mathbf{z},t}$. El segundo término, $\mathbb{E}^{X_t, X^*}\left(F_{\mathbf{z}}(X_t) - F_{\mathbf{z}}(X^*)\right)$, corresponde al error de aproximación desde la dinámica $X_t$ a su distribución estacionaria $\pi_{\mathbf{z}}$. Finalmente, $\mathbb{E}^{X^*}\left(F_{\mathbf{z}}(X^*) - F_{\mathbf{z}}^*\right)$ corresponde al error de estimación por parte de la medida de Gibbs al óptimo de la función.

Cabe mencionar que en otras investigaciones (como Raginsky y col., 2017 y Borkar y Mitter, 1999) se realiza un análisis parecido añadiendo un error extra proveniente de la aproximación del problema de riesgo esperado por el problema de riesgo empírico. Este análisis se evitó por dos simples motivos: el primero corresponde al trabajo ya realizado en investigaciones al respecto, donde se prueba que la convergencia se logra cuando la cantidad de datos es suficientemente grande (por ejemplo Vapnik, 1979 y Vapnik, 1991). La segunda razón es la importancia práctica, ya que el problema de riesgo empírico se relaciona directamente con aplicaciones como el problema de tomografía sísmica que se planteará posteriormente.

Recordamos que, sin pérdida de generalidad, tanto la esperanza $\mathbb{E}^{x_{\lfloor t/\eta \rfloor}, X_t}$ como $\mathbb{E}^{X_t, X^*}$, pueden ser tomadas bajo la ley que induce el acoplamiento óptimo que genera a la distancia 2-Wasserstein. Para las siguientes secciones se tomando a Raginsky y col., 2017 como principal referencia.

## 4.3. Error de discretización

Gracias a las condiciones anteriores y los lemas mencionados antes logramos obtener el siguiente resultado.

**Teorema 4.1** *Bajo las condiciones (1), (2), (3) y (4), para todo $t \in (0, T_0]$, existe una constante $\mathcal{C}_1 > 0$ que no depende de $\eta$, tal que*

$$\left|\mathbb{E}_{\mathbf{z}}^{x_{\lfloor t/\eta \rfloor}}\left(F_{\mathbf{z}}(x_{\lfloor t/\eta \rfloor})\right) - \mathbb{E}_{\mathbf{z}}^{X_t}\left(F_{\mathbf{z}}(X_t)\right)\right| \leq \mathcal{C}_1 \mathcal{W}(\mu_{\mathbf{z},\lfloor t/\eta \rfloor}, \nu_{\mathbf{z},t}), \tag{4.29}$$

*con*

$$\mathcal{C}_1 = \sqrt{2L^2\hat{\sigma}_1^2 + 2B^2}, \tag{4.30}$$

$\hat{\sigma}_1^2 \in \mathbb{R}_+$ *y*

$$\hat{\sigma}_1^2 \leq \kappa_0 + \frac{b + d/\beta}{m}. \tag{4.31}$$

*Además, por el Corolario 4.1, se tiene que*

$$\left|\mathbb{E}_{\mathbf{z}}^{x_{\lfloor t/\eta \rfloor}}\left(F_{\mathbf{z}}(x_{\lfloor t/\eta \rfloor})\right) - \mathbb{E}_{\mathbf{z}}^{X_t}\left(F_{\mathbf{z}}(X_t)\right)\right| \to 0, \tag{4.32}$$

*cuando $\eta \to 0$.*

**Observación:** Tanto este teorema, como los lemas anteriores utilizan la esperanza condicional a la observación $\mathbf{z}$, $\mathbb{E}_{\mathbf{z}}$, pero dado que las cotas encontradas no dependen de esta variable, notamos que estos resultados también logran ser válidos bajo la ley conjunta total de $F_{\mathbf{Z}}(x_{\lfloor t/\eta \rfloor})$ y $F_{\mathbf{Z}}(X_t)$, donde $\mathbf{Z}$ ahora es una variable aleatoria en $\mathcal{Z}^n$, independiente del algoritmo.

Demostración. (Teorema 4.1) Gracias a la condición (1), sabemos que la función $f(\cdot, z)$ es continuamente diferenciable para todo $z \in \mathcal{Z}$. Sean $x, w \in \mathbb{R}^d$ y $z \in \mathcal{Z}$, entonces

$$f(x,z) = f(w,z) + \int_0^1 \langle x - w, \nabla f(tx + (1-t)w, z)\rangle dt, \tag{4.33}$$

entonces, utilizando la desigualdad de Cauchy-Schwarz para el producto interno en $\mathbb{R}^d$,

$$\begin{aligned} |f(x,z) - f(w,z)| &\leq \int_0^1 |\langle x - w, \nabla f(tx + (1-t)w, z)\rangle| \, dt \\ &\leq \int_0^1 \|x - w\| \cdot \|\nabla f(tx + (1-t)w, z)\| dt. \end{aligned} \tag{4.34}$$

Juntando la desigualdad (4.34) y el Lema 4.1, tenemos que

$$\begin{aligned} |f(x,z) - f(w,z)| &\leq \int_0^1 \|x - w\| \cdot \|\nabla f(tx + (1-t)w, z)\| dt. \\ &\leq \|x - w\| \left(L \int_0^1 t\|x\| + (1-t)\|w\| dt + B\right) \\ &\leq \|x - w\| \left(\tfrac{L}{2}(\|x\| + \|w\|) + B\right). \end{aligned} \tag{4.35}$$

Sea $\mathbb{E}_{\mathbf{z}}^{x_{\lfloor t/\eta \rfloor}, X_t}(\,\cdot\,)$, el valor esperado para el acoplamiento óptimo que genera la distancia de

Wasserstein, es decir

$$\sqrt{\mathbb{E}_{\mathbf{z}}^{x_{\lfloor t/\eta \rfloor}, X_t}\left[\|x_{\lfloor t/\eta \rfloor} - X_t\|^2\right]} = \mathcal{W}(\mu_{\mathbf{z},\lfloor t/\eta \rfloor}, \nu_{\mathbf{z},t}). \tag{4.36}$$

Entonces, dado un vector de observaciones $\mathbf{z} \in \mathcal{Z}^n$ y $t \in (0, T_0]$, tenemos que

$$\left|\mathbb{E}_{\mathbf{z}}^{x_{\lfloor t/\eta \rfloor}}\left(F_{\mathbf{z}}(x_{\lfloor t/\eta \rfloor})\right) - \mathbb{E}_{\mathbf{z}}^{X_t}\left(F_{\mathbf{z}}(X_t)\right)\right| = \left|\mathbb{E}_{\mathbf{z}}^{x_{\lfloor t/\eta \rfloor}, X_t}\left(F_{\mathbf{z}}(x_{\lfloor t/\eta \rfloor}) - F_{\mathbf{z}}(X_t)\right)\right|$$

$$\leq \mathbb{E}_{\mathbf{z}}^{x_{\lfloor t/\eta \rfloor}, X_t}\left(\left|F_{\mathbf{z}}(x_{\lfloor t/\eta \rfloor}) - F_{\mathbf{z}}(X_t)\right|\right) \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\mathbf{z}}^{x_{\lfloor t/\eta \rfloor}, X_t}\left(|f(x_{\lfloor t/\eta \rfloor}, z_i) - f(X_t, z_i)|\right)$$

$$\leq \mathbb{E}_{\mathbf{z}}^{x_{\lfloor t/\eta \rfloor}, X_t}\left(\|x_{\lfloor t/\eta \rfloor} - X_t\|\left(\frac{L}{2}(\|x_{\lfloor t/\eta \rfloor}\| + \|X_t\|) + B\right)\right)$$

Finalmente, usando la desigualdad de Cuachy-Schwarz, esta vez para la esperanza condicional,

$$\mathbb{E}_{\mathbf{z}}^{x_{\lfloor t/\eta \rfloor}, X_t}\left(\|x_{\lfloor t/\eta \rfloor} - X_t\|\left(\frac{L}{2}(\|x_{\lfloor t/\eta \rfloor}\| + \|X_t\|) + B\right)\right)$$

$$\leq \sqrt{L^2\left(\mathbb{E}_{\mathbf{z}}^{x_{\lfloor t/\eta \rfloor}}(\|x_{\lfloor t/\eta \rfloor}\|^2) + \mathbb{E}_{\mathbf{z}}^{X_t}(\|X_t\|^2)\right) + 2B^2}\sqrt{\mathbb{E}_{\mathbf{z}}^{x_{\lfloor t/\eta \rfloor}, X_t}\left[\|x_{\lfloor t/\eta \rfloor} - X_t\|^2\right]}.$$

Recordando que $\lfloor t/\eta \rfloor$ es un valor $k \in \mathbb{N}$ de iteración, es posible definir $\hat{\sigma}_1^2 = \mathbb{E}_{\mathbf{z}}^{x_{\lfloor t/\eta \rfloor}}(\|x_{\lfloor t/\eta \rfloor}\|^2) \vee \mathbb{E}_{\mathbf{z}}^{X_t}(\|X_t\|^2)$. Gracias al lema 4.2, $\hat{\sigma}_1^2$ es finito y no depende de $\eta$. Así,

$$\left|\mathbb{E}_{\mathbf{z}}^{x_{\lfloor t/\eta \rfloor}}\left(F_{\mathbf{z}}(x_{\lfloor t/\eta \rfloor})\right) - \mathbb{E}_{\mathbf{z}}^{X_t}\left(F_{\mathbf{z}}(X_t)\right)\right| \leq \sqrt{2L^2\hat{\sigma}_1^2 + 2B^2}\sqrt{\mathbb{E}_{\mathbf{z}}^{x_{\lfloor t/\eta \rfloor}, X_t}\left[\|x_{\lfloor t/\eta \rfloor} - X_t\|^2\right]}$$

$$\leq \sqrt{2L^2\hat{\sigma}_1^2 + 2B^2}\,\mathcal{W}(\mu_{\mathbf{z},\lfloor t/\eta \rfloor}, \nu_{\mathbf{z},t}).$$

Ocupando el Corolario 4.1, concluimos que

$$\left|\mathbb{E}_{\mathbf{z}}^{x_{\lfloor t/\eta \rfloor}}\left(F_{\mathbf{z}}(x_{\lfloor t/\eta \rfloor})\right) - \mathbb{E}_{\mathbf{z}}^{X_t}\left(F_{\mathbf{z}}(X_t)\right)\right| \to 0$$

cuando $\eta \to 0$, para todo $t \in (0, T_0]$. □

Se vuelve a recalcar lo mencionado en el Lema 4.6, donde el peso de la importancia de los supuestos cae en que esta aproximación sólo se asegura bajo horizontes temporales finitos.

## 4.4. Error de distribución estacionaria

En esta sección se propone una forma de cuantificar el sesgo existente entre la estimación hecha por la distribución de la dinámica hasta un tiempo finito, $T_0 > 0$, y la estimación hecha por la distribución de Gibbs, ley estacionaria de la dinámica anterior.

En el caso del contexto de la dinámica (3.4), el siguiente resultado de Raginsky y col., 2017, nos asegura que la medida $\pi_{\mathbf{z}}$ cumple la desigualdad logarítmica de Sobolev (1.2) y nos facilita una cota para la constante $c_{LS}$.

**Lema 4.7** (Proposición 9 en Raginsky y col., 2017) *Bajo las condiciones (1), (2), (3) y (4), para $\beta > 2/m$, todas las distribuciones de Gibbs $\pi_{\mathbf{z}}$ cumplen la desigualdad logarítmica de*

*Sobolev, con constante*

$$c_{LS} \leq \frac{2m^2 + 8L^2}{m^2 L\beta} + \frac{1}{\lambda_*}\left(\frac{6L(d+\beta)}{m} + 2\right). \tag{4.37}$$

*Además, según Raginsky y col., 2017, Cattiaux y col., 2008 y Bakry y col., 2013, por el teorema de Otto y Villani, tenemos que la distribución de Gibbs cumple con la desigualdad de transporte-entropía de orden 2 o costo cuadrático de transporte (Definición 1.7):*

$$\mathcal{W}(\nu_{\mathbf{z},t}, m)^2 \leq 2c_{LS} D(m||\pi_{\mathbf{z}}), \tag{4.38}$$

*donde $m \in \mathcal{P}(\mathbb{R}^d)$, arbitraria.*

El siguiente Lema nos entrega una cota para la discrepancia que existe entre la ley inicial del algoritmo SGLD, $\mu_{\mathbf{z},0}$, y la distribución de Gibbs, cuando $\mu_{\mathbf{z},0}$ cumple la condición (4). Sin embargo, hace uso una pequeña condición extra.

**Condición 5** Existe una constante $A > 0$ tal que

$$|f(0, z)| \leq A, \quad \forall\, z \in \mathcal{Z}. \tag{4.39}$$

**Lema 4.8** (Lema 5 en Raginsky y col., 2017) *Suponiendo las condiciones (1), (2), (3), (4) y (5) Se tiene que para todo $\mathbf{z} \in \mathcal{Z}^n$,*

$$D(\mu_{\mathbf{z},0}||\pi_{\mathbf{z}}) \leq \log \|p_0\|_\infty + \frac{d}{2}\log\frac{3\pi}{m\beta} + \beta\left(\frac{L\kappa_0}{3} + B\sqrt{\kappa_0} + A + \frac{b}{3}\log 3\right), \tag{4.40}$$

*donde $p_0$ es la densidad de $\mu_{\mathbf{z},0}$ con respecto a la medida de Lebesgue.*

Con este último lema, el Lema 4.7 (desigualdad 4.38), junto al teorema de decaimiento exponencial (Teorema 1.2), que implica la relación

$$D(\nu_{\mathbf{z},t} \,\|\, \pi_{\mathbf{z}}) \leq D(\nu_{\mathbf{z},0} \,\|\, \pi_{\mathbf{z}}) e^{-2t/\beta c_{LS}},$$

y recordando que en el instante $t = 0$ tenemos que $\mu_{\mathbf{z},0} = \nu_{\mathbf{z},0}$, nos permite enunciar el siguiente resultado, que corresponde a una leve modificación del Lema 9 de Raginsky y col., 2017.

**Lema 4.9** $\mathcal{W}(\nu_{\mathbf{z},t}, \pi_{\mathbf{z}})^2 \to 0$ *cuando $t \to \infty$. Más aún, se tiene la siguiente cota de convergencia*

$$\mathcal{W}(\nu_{\mathbf{z},t}, \pi_{\mathbf{z}}) \leq \mathcal{C}_2 e^{-t/\beta c_{LS}}, \tag{4.41}$$

*con*

$$\mathcal{C}_2 = \sqrt{2c_{LS}\left(\log\|p_0\|_\infty + \frac{d}{2}\log\frac{3\pi}{m\beta} + \beta\left(\frac{L\kappa_0}{3} + B\sqrt{\kappa_0} + A + \frac{b}{2}\log 3\right)\right)}. \tag{4.42}$$

La importancia de la constante $\mathcal{C}_2$ yace en la independencia de $t$ y de $\mathbf{z}$, aspectos que serán de gran importancia en el resultado principal de este apartado. Sin embargo, para un análisis en conjunto del algoritmo, escribiremos $\mathcal{C}_2 = \mathcal{C}_2(\beta)$ o $\mathcal{C}_2 := \mathcal{C}_2(-\log(\beta), \beta)$ para recalcar la dependencia de $\beta$. Por otro lado, lo que se encuentra al interior de la raíz cuadrada en la

definición de $\mathcal{C}_2$ es positivo, y define de buena forma la constante, para todo valor de $\beta$ si se cumple que

$$\frac{6\pi e}{dm}\left(\frac{L\kappa_0}{3}+B\sqrt{\kappa_0}+A+\frac{b}{2}\log 3\right)\|p_0\|_\infty^{2/d}>1. \tag{4.43}$$

De lo contrario, si estas constantes (positivas), no cumplen la última condición expuesta, de todas formas $\mathcal{C}_2$ estará bien definido como número real para valores suficientemente altos de $\beta$. En efecto, considerando que $2c_{LS}$ es un valor positivo, notamos que la expresión

$$\log\|p_0\|_\infty+\frac{d}{2}\log\frac{3\pi}{m\beta}+\beta\left(\frac{L\kappa_0}{3}+B\sqrt{\kappa_0}+A+\frac{b}{2}\log 3\right)$$

es convexa como función de $\beta$, por lo que posee un único punto mínimo, con valor

$$\log\|p_0\|_\infty+\frac{d}{2}\log\left[\frac{6\pi}{dm}\left(\frac{L\kappa_0}{3}+B\sqrt{\kappa_0}+A+\frac{b}{2}\log 3\right)\right]+\frac{d}{2}.$$

Imponiendo que este valor sea positivo y ocupando la monotonía de la función exponencial, se encuentra el valor expresado en (4.43). Finalmente notamos que la expresión tiende a infinito cuando $\beta\to\infty$, por lo tanto, encontrando valores suficientemente grandes podemos asegurar que la raíz esté bien definida en caso de que (4.43) no se cumpla.

**Teorema 4.2** *Bajo las condiciones (1), (2), (3), (4) y (5), se tiene que* $\forall\,\beta>0$,

$$\left|\mathbb{E}_{\mathbf{z}}^{X_t}\left(F_{\mathbf{z}}(X_t)\right)-\mathbb{E}_{\mathbf{z}}^{X^*}\left(F_{\mathbf{z}}(X^*)\right)\right|\leq\mathcal{C}_3\mathcal{W}(\nu_{\mathbf{z},t},\pi_{\mathbf{z}}),\quad\forall\,t>0, \tag{4.44}$$

*donde*

$$\mathcal{C}_3=\sqrt{2L^2\hat{\sigma}_2^2+2B^2}, \tag{4.45}$$

*con* $\hat{\sigma}_2^2\in\mathbb{R}_+$ *y*

$$\hat{\sigma}_2^2\leq\frac{b+d/\beta}{m}. \tag{4.46}$$

*Además, por el Lema 4.9 tenemos que*

$$\left|\mathbb{E}_{\mathbf{z}}^{X_t}\left(F_{\mathbf{z}}(X_t)\right)-\mathbb{E}_{\mathbf{z}}^{X^*}\left(F_{\mathbf{z}}(X^*)\right)\right|\leq\mathcal{C}_3\mathcal{C}_2e^{-t/\beta c_{LS}}, \tag{4.47}$$

*(donde* $\mathcal{C}_2$ *corresponde a la constante del Lema 4.9) y, por lo tanto,*

$$\left|\mathbb{E}_{\mathbf{z}}^{X_t}\left(F_{\mathbf{z}}(X_t)\right)-\mathbb{E}_{\mathbf{z}}^{X^*}\left(F_{\mathbf{z}}(X^*)\right)\right|\to 0, \tag{4.48}$$

*cuando* $t\to\infty$.

Demostración. Recordando la cota utilizada en el Teorema 4.1, desigualdad (4.35),

$$|f(x,z)-f(w,z)|\leq\|x-w\|\left(\frac{L}{2}(\|x\|+\|w\|)+B\right),\quad\forall\,z\in\mathcal{Z}. \tag{4.49}$$

Sea $\mathbb{E}_{\mathbf{z}}^{X_t,X^*}(\,\cdot\,)$, la esperanza tomada en el acoplamiento óptimo tal que $\mathbb{E}_{\mathbf{z}}^{X_t,X^*}(\|X_t-X^*\|^2)=$

$\mathcal{W}(\nu_{\mathbf{z},t}, \pi_{\mathbf{z}})^2$. Entonces, ocupando la desigualdad de Cauchy-Schwarz, se tiene que

$$\left|\mathbb{E}_{\mathbf{z}}^{X_t}\left(F_{\mathbf{z}}(X_t)\right) - \mathbb{E}_{\mathbf{z}}^{X^*}\left(F_{\mathbf{z}}(X^*)\right)\right| \tag{4.50}$$

$$\leq \sqrt{L^2\left(\mathbb{E}_{\mathbf{z}}^{X_t}(\|X_t\|^2) + \mathbb{E}_{\mathbf{z}}^{X^*}(\|X^*\|^2)\right) + 2B}\sqrt{\mathbb{E}_{\mathbf{z}}^{X_t,X^*}\left(\|X_t - X^*\|^2\right)} \tag{4.51}$$

Por el Lema 4.9, sabemos que $\mathcal{W}(\nu_{\mathbf{z},t}, \pi_{\mathbf{z}}) \to 0$. Combinado con el Teorema 1.3 podemos concluir que $X_t$ converge débilmente a $X^*$, cuando $t \to \infty$. Además, tomando el punto *iii)* del mismo teorema, tenemos que

$$\int \|w\|^2 d\pi_{\mathbf{z}} = \lim_{t\to\infty} \int \|w\|^2 d\nu_{\mathbf{z},t} \leq \frac{b + d/\beta}{m}, \tag{4.52}$$

donde la última desigualdad viene de la cota (4.9) del Lema 4.2. De esta forma, definiendo $\hat{\sigma}_2^2$ como

$$\hat{\sigma}_2^2 = \mathbb{E}_{\mathbf{z}}^{X_t}(\|X_t\|^2) \vee \mathbb{E}_{\mathbf{z}}^{X^*}(\|X^*\|^2) \leq \frac{b + d/\beta}{m}, \tag{4.53}$$

tenemos que

$$\left|\mathbb{E}_{\mathbf{z}}^{X_t}\left(F_{\mathbf{z}}(X_t)\right) - \mathbb{E}_{\mathbf{z}}^{X^*}\left(F_{\mathbf{z}}(X^*)\right)\right| \leq \sqrt{2L^2\hat{\sigma}_2^2 + 2B}\,\sqrt{\mathbb{E}_{\mathbf{z}}^{X_t,X^*}\left(\|X_t - X^*\|^2\right)}. \tag{4.54}$$

Recordando que $\mathbb{E}_{\mathbf{z}}^{X_t,X^*}$ definía la distancia de Wasserstein, se concluye la desigualdad (4.44). □

Se vuelve a destacar la presencia del parámetro $\beta$ en las constantes $\mathcal{C}_2$ y $\mathcal{C}_3$, cuya consideración será importante en los resultados posteriores. Por el momento sólo se dirá que, cualquiera sea el valor de $\beta$, ambas constantes mantienen su carácter de finitas, aun cuando $\beta$ pueda tomar valores positivos tan grandes como se quiera.

## 4.5. Error de estimación del óptimo

Es bien sabido desde Raginsky y col., 2017, Chiang y col., 1987 y Hwang, 1980, que la ley de probabilidad $\pi_{\mathbf{z}}$ converge débilmente a una distribución uniforme en el conjunto de puntos óptimos globales de la función $F_{\mathbf{z}}(x)$ cuando $\beta \to \infty$. De poderse generar muestras de esta distribución, el problema de minimización estaría resuelto, pues $F_{\mathbf{z}}(x) = F_{\mathbf{z}}^*$, casi seguramente, donde $F_{\mathbf{z}}^*$ representa el valor óptimo de la función. Sin embargo, al tratarse de un comportamiento asintótico, es infactible lograr esta convergencia, por lo tanto se buscará cuantificar el nivel de sesgo que tendrá el valor esperado de $F_{\mathbf{z}}(X^*)$ cuando $X^*$ distribuya como $\pi_{\mathbf{z}}$ para valores de $\beta$ suficientemente altos.

El siguiente resultado de Raginsky y col., 2017 nos entrega una cota para este error.

**Teorema 4.3** (Proposición 11 en Raginsky y col., 2017) *Para todo $\beta > 2/m$,*

$$\mathbb{E}_{\mathbf{z}}^{X^*}(F_{\mathbf{z}}(X^*)) - F_{\mathbf{z}}^* \leq \frac{d}{2\beta}\log\left(\frac{eL}{m}\left(\frac{b\beta}{d} + 1\right)\right). \tag{4.55}$$

*En particular, tenemos que*

$$\left|\mathbb{E}_{\mathbf{z}}^{X^*}(F_{\mathbf{z}}(X^*)) - F_{\mathbf{z}}^*\right| \to 0, \tag{4.56}$$

*cuando* $\beta \to \infty$.

## 4.6. Sesgo del algoritmo SGLD

Basta con juntar los resultados de los Teoremas 4.1, Teorema 4.2 y Teorema 4.3, acompañado del Lema 4.9, para enunciar la cota de error que podemos encontrar con el algoritmo SGLD.

**Teorema 4.4** *Sea* $T_0 > 0$, $\beta > 2/m$. *Bajo las condiciones (1), (2), (3), (4) y (5), tenemos que el error esperado entre la realización del algoritmo SGLD y el óptimo de la función, cumple que*

$$\left|\mathbb{E}^{x_{\lfloor T_0/\eta \rfloor}}\left(F_{\mathbf{Z}}(x_{\lfloor T_0/\eta \rfloor})\right) - F_{\mathbf{Z}}^*\right| \leq \mathcal{C}_1\,\varphi(\eta) + \mathcal{C}_3\mathcal{C}_2(\beta)e^{-T_0/\beta c_{LS}}$$
$$+\frac{d}{2\beta}\log\left(\frac{eL}{m}\left(\frac{b\beta}{d}+1\right)\right),$$

*donde* $\mathcal{C}_1$ *y* $\varphi(\eta)$ *corresponde a las constantes del Teorema 4.1 y el Corolario 4.1, es decir*

$$\mathcal{C}_1 = \sqrt{2L^2\hat{\sigma}_1^2 + 2B^2}, \tag{4.57}$$

$$\varphi(\eta) = \left(\mathcal{Q}_1(\beta)\eta^2 + \mathcal{Q}_2(\beta)\eta\right)^{1/2}, \tag{4.58}$$

*con*

$$\mathcal{Q}_1(\beta) = 2\left(T_0\left(L\mathcal{K} + C + \frac{2}{\beta}\right)Le^{LT_0} + 2(C + 2L^2\mathcal{K}^2 + 2B^2)\right), \quad y \tag{4.59}$$

$$\mathcal{Q}_2(\beta) = 2\left((T_0+1)\left(C + \frac{2b}{\beta}\right)e^{LT_0} + 2\beta^{-1}d\right). \tag{4.60}$$

$\mathcal{C}_2(\beta)$ *y* $\mathcal{C}_3$ *corresponden a las constantes del Lema 4.9 y Teorema 4.2, es decir*

$$\mathcal{C}_2(\beta) = \sqrt{2c_{LS}\left(\log\|p_0\|_\infty + \frac{d}{2}\log\frac{3\pi}{m\beta} + \beta\left(\frac{L\kappa_0}{3} + B\sqrt{\kappa_0} + A + \frac{b}{2}\log 3\right)\right)}, \tag{4.61}$$

$$\mathcal{C}_3 = \sqrt{2L^2\hat{\sigma}_2^2 + 2B^2}. \tag{4.62}$$

*Además, tenemos que*

$$\hat{\sigma}_1^2,\ \hat{\sigma}_2^2 \leq \frac{b + d/\beta}{m}. \tag{4.63}$$

Observamos que, a diferencia de los resultados anteriores, este teorema ocupa la esperanza total del modelo, dejando a la muestra aleatoria de $\mathcal{Z}^n$ como la variable $\mathbf{Z}$ que no está fija. Esto no es problema desde que las cotas encontradas en los teoremas anteriores no dependen de la muestra $\mathbf{z}$, por lo que es posible encontrar el mismo resultado para la esperanza condicional y concluir por la monotonía de la esperanza.

### 4.6.1. Codependencia entre $\eta$, $T_0$ y $\beta$.

Una pregunta interesante es evaluar la posibilidad de obtener un valor mínimo de la cota del Teorema 4.4, para algún $T_0$, poder evaluar el comportamiento del error cuando $\beta \to \infty$ y, en el mejor de los casos, encontrar condiciones para cumplir que

$$\mathcal{C}_3\mathcal{C}_2 e^{-T_0/\beta c_{LS}} + \frac{d}{2\beta}\log\left(\frac{eL}{m}\left(\frac{b\beta}{d}+1\right)\right) \to 0.$$

Lamentablemente esto no es posible. El análisis anterior parte cuantificando el error que se obtiene al momento de discretizar la dinámica y finaliza trabajando con la distribución estacionaria de la dinámica que se quizo discretizar inicialmente. Usualmente, para llegar a esa distribución estacionaria, el proceso estocástico debe recorrer hasta un tiempo infinito, lo que claramente no es factible de implementar. En este apartado se describirá la codependencia entre las variables $\eta$, $T_0$ y $\beta$.

Partiremos notando que cualquier tipo de sucesión tal que $T_0 \to \infty$ no es compatible con el resultado del Lema 4.6,cuya condición exige que el valor de $T_0$ sea finito. Esto implica que el primer término de error en el Teorema 4.4, es decir;

$$\mathcal{C}_1\,\varphi(\eta),$$

con

$$\varphi(\eta) = \left(\mathcal{Q}_1\eta^2 + \mathcal{Q}_2\eta\right)^{1/2},$$

diverge cuando $T_0 \to \infty$, puesto que los valores de $\mathcal{Q}_1$ (ver (4.59)) y $\mathcal{Q}_2$ (ver (4.60)) lo hacen para todo par de valores $\eta > 0$ y $\beta > 0$. Por lo tanto, podemos concluir que la condición de finitud para $T_0$ es necesaria en el resultado. Por último, recordemos que, como $T_0$ es el tiempo final del proceso que replica el algoritmo SGLD, la relación respecto a $\eta$ es de la forma $T_0 = K\eta$, con $K \in \mathbb{N}$, algún natural finito que representa la cantidad de pasos del algoritmo. Por lo tanto, podríamos afirmar que tomar un $\eta$ tan chico como sea necesario es posible si asumimos que el algoritmo es ejecutable para cualquier cantidad de pasos.

El tercer término del error del algoritmo SGLD del Teorema 4.4 es

$$\frac{d}{2\beta}\log\left(\frac{eL}{m}\left(\frac{b\beta}{d}+1\right)\right),$$

y corresponde al error de aproximación del valor medio de la función, bajo la distribución de Gibbs, al valor óptimo de la función objetivo, expresado en el Teorema 4.3. No es difícil mostrar que este valor converge a 0 cuando $\beta \to \infty$, lo que implica la convergencia de la distribución y la disminución del error del algoritmo. Sin embargo, el segundo término de error será aquel que dificulte el movimiento de $\beta$. En efecto, dicho término está dado por

$$\mathcal{C}_3\mathcal{C}_2 e^{-T_0/\beta c_{LS}}, \tag{4.64}$$

con

$$\mathcal{C}_2 = \sqrt{2c_{LS}\left(\log\|p_0\|_\infty + \frac{d}{2}\log\frac{3\pi}{m\beta} + \beta\left(\frac{L\kappa_0}{3} + B\sqrt{\kappa_0} + A + \frac{b}{2}\log 3\right)\right)},$$

$$\mathcal{C}_3 = \sqrt{2L^2\hat{\sigma}_2^2 + 2B^2}.$$

Podemos simplificar la expresión (4.64) considerando que $\mathcal{C}_3$ no depende de $\beta$, ni de $T_0$, y que todos los demás términos que aparecen son constantes positivas. De esta forma podemos reescribir (4.64) como

$$\left(\sqrt{A\log\frac{B}{\beta}+C\beta+D}\right)\, e^{-T_0/\beta c_{LS}}, \tag{4.65}$$

donde $A, B, C, D > 0$ son constantes. Notamos que, para valores de $T_0$ constantes, el valor de la expresión (4.65) crece hacia infinito cuando $\beta \to \infty$. Por lo tanto, es necesario que $T_0$ crezca lo suficiente como para contrarrestar el crecimiento provocado por $\beta$. Supongamos que $T_0$ es de la forma $\beta^l$, con $l > 1$, entonces tenemos que:

$$\left(\sqrt{A\log\frac{B}{\beta}+C\beta+D}\right)\, e^{-T_0/\beta c_{LS}} = \frac{\left(\sqrt{A\log\frac{B}{\beta}+C\beta+D}\right)}{e^{\beta^{l-1}/c_{LS}}}. \tag{4.66}$$

Para valores altos de $\beta$, el numerador de la expresión anterior crece a un ritmo parecido al de $\sqrt{\beta}$, lo que se ve completamente dominado por el crecimiento de la exponencial en el denominador, puesto que $l - 1 > 0$. Por lo tanto, esta condición basta para para asegurar el control sobre el segundo término del error medio del algoritmo SGLD. Sin embargo, la conclusión anterior condiciona el valor de $T_0$ a una cantidad de al menos $\beta^l$, por lo que la convergencia del error para $\beta \to \infty$ está condicionada a $T_0 \to \infty$, tendencia que ya se argumentó como infactible.

A pesar de lo anterior, es posible ajustar los parámetros para obtener valores pequeños (pero no nulos) para el error acotado por el Teorema 4.4, justificado por la convergencia del término $\mathcal{C}_4(\beta)e^{-t/\beta c_{LS}}$. En efecto, sea $\varepsilon > 0$. Existe $\beta_0 > 0$, tal que $\forall \beta \geq \beta_0$ tenemos que

$$\mathcal{C}_4(\beta)e^{-t/\beta c_{LS}} \leq \varepsilon.$$

Como $\mathcal{C}_3\mathcal{C}_2(\beta)e^{-t/\beta c_{LS}}$ es decreciente en $\beta$, la peor elección que cumpla un error del orden de $\varepsilon$ será $\beta_0$. Esta elección de peor valor puede ser reemplazada por otro electo con el criterio de ser el más "barato" en términos de la ejecución del algoritmo (tiempo, memoria computacional, etc.), acá se asumirá que la elección es la misma.

Para el resto del error, basta con una elección de $T_0$ tal que $\beta_0 < T_0$ lo suficientemente grande para asegurar un valor pequeño, que contrarreste el crecimiento otorgado por el alto valor de $\beta_0$ (puesto que $\mathcal{C}_4(\beta)e^{-t/\beta c_{LS}}$ es creciente con respecto a $\beta$). De esta forma, un error "pequeño" del sesgo del algoritmo SGLD cuando $\eta \to 0$, estará dado por

$$\mathcal{C}_3\mathcal{C}_2(\beta_0)(\beta_0)e^{-T_0/\beta_0 c_{LS}} + \varepsilon \leq \tilde{\varepsilon}.$$

Buscar un $T_0$ que mayore a $\beta_0$ lo suficiente no debería ser tan costoso en comparación con $\beta_0$, sin embargo, esto dependerá fuertemente del aporte de las constantes que participan en $\mathcal{C}_2$ y $\mathcal{C}_3$. Si el orden de las constantes es notoriamente más pequeño que el orden numérico de $\beta_0$, entonces pequeños aumentos de $T_0$ asegurarían errores deseados.

**Observación:** Para cerrar esta discusión, se agregará que el análisis recién incluido no considera la repercusión que puede representar en el costo o la dificultad que puede existir al querer ejecutar el algoritmo simulando una dinámica en un tiempo $T_0$ alto, pero finito, con $\eta$ pequeño.

# Capítulo 5

# Algoritmo SGLD para el problema de EMV

Recordamos el estimador del estimador de máxima verosimilitud, valor que resuelve el problema de optimización (escrito en forma de minimización esta vez) (2.7):

$$\underset{\theta\in\Theta}{\textbf{minimize}} -\frac{1}{n}\sum_{i=1}^{n}\log p_\theta(z_i). \tag{5.1}$$

Equivalente el problema de riesgo empírico esperado (2.2), tomando $f(\theta, z_i) = -\log p_\theta(z_i)$ como la función a minimizar evaluado en una muestra aleatoria $z_i$.

El algoritmo SGLD expresado en (3.3), para este caso particular tiene la forma

$$\theta_{k+1} = \theta_k - \eta g_k + \sqrt{2\eta\beta^{-1}}\xi_k. \tag{5.2}$$

La variable $g_k$ es un estimador insesgado del gradiente de la función objetivo, es decir:

$$\mathbb{E}\, g_k = \nabla_\theta\left(-\frac{1}{n}\sum_{i=1}^{n}\log p_{\theta_k}(z_i)\right) = -\frac{1}{n}\sum_{i=1}^{n}\frac{1}{p_{\theta_k}(z_i)}\nabla_\theta p_{\theta_k}(z_i).$$

Una de las elecciones usuales para $g_k$ es la usada en el método de *mini-batch*, esto es; para $h$, un número natural tal que $h \leq n$, se selecciona un subconjunto de índices aleatoriamente, $I \subseteq \{1, \cdots, n\}$. Más precisamente, eligiendo

$$\tilde{g}_k = -\frac{1}{h}\sum_{i\in I}\frac{1}{p_{\theta_k}(z_i)}\nabla_\theta p_{\theta_k}(z_i),$$

si $I = (i_1, \cdots, i_h)$ es una variable aleatoria que distribuye uniformemente entre todos los subconjuntos de $\{1, \cdots, n\}$ de tamaño $h$, entonces $\tilde{g}_k$ es un estimador insesgado con respecto al gradiente total de la función objetivo. Más aún, si se cumple la Condición 1, entonces, para este estimador, también se cumple la Condición 3 con $\delta = 1/h$. En efecto, podemos definir el mecanismo de oráculo de gradiente estocástico (Definición 3.1) tomando $U_{\mathbf{z}} = I$ junto a la función $g : \mathbb{R}^d \times \{1, \cdots, n\}^n \mapsto \mathbb{R}^d$ definida como

$$g(\theta, U_{\mathbf{z}} = I) = -\frac{1}{h}\sum_{i\in I}\nabla f(\theta, z_i). \tag{5.3}$$

De esta forma tenemos que $g$ es un estimador insesgado para el gradiente de la función objetivo del problema de minimización definido en (2.7) y

$$\mathbb{E}_I\left(\left\|\frac{1}{h}\sum_{i\in I}\nabla f(\theta,z_i)-\frac{1}{n}\sum_{i=1}^{n}\nabla f(\theta,z_i)\right\|^2\right)\leq\frac{1}{h^2}\mathbb{E}_I\left(\sum_{i=1}^{n}\|\nabla f(\theta,z_i)\|^2\left[\mathbb{1}_{i\in I}-\frac{h}{n}\right]^2\right). \quad (5.4)$$

Podemos notar que el término $\left[\mathbb{1}_{i\in I}-\frac{h}{n}\right]^2\leq 1$ para todo $i$ entre 1 y $n$, puesto que $h\leq n$. Por lo tanto, el último valor esperado puede escribirse de forma implícita sumando sobre la familia $\mathcal{I}$ de los $\binom{n}{h}$ subconjuntos de tamaño $h$, presentando la siguiente desigualdad

$$\mathbb{E}_I\left(\sum_{i=1}^{n}\|\nabla f(\theta,z_i)\|^2\left[\mathbb{1}_{i\in I}-\frac{h}{n}\right]^2\right)\leq\mathbb{E}_I\left(\sum_{i=1}^{n}\|\nabla f(\theta,z_i)\|^2\right)=\frac{1}{\binom{n}{h}}\sum_{I\in\mathcal{I}}\sum_{i=1}^{n}\|\nabla f(\theta,z_i)\|^2,$$

un pequeño argumento combinatorial nos permite mostrar que, por cada índice $i$ entre 1 y $n$, existen $\binom{n-1}{h-1}$ subconjuntos de tamaño $h$ que contengan a $i$, por lo que doble suma anterior se reduce a

$$\frac{1}{\binom{n}{h}}\sum_{I\in\mathcal{I}}\sum_{i=1}^{n}\|\nabla f(\theta,z_i)\|^2=\frac{\binom{n-1}{h-1}}{\binom{n}{h}}\sum_{i=1}^{n}\|\nabla f(\theta,z_i)\|^2=\frac{h}{n}\sum_{i=1}^{n}\|\nabla f(\theta,z_i)\|^2$$

Si asumimos que se cumple la Condición 1, entonces, por el Lema 4.1 tenemos que

$$\|\nabla f(\theta,z_i)\|^2\leq 2(L^2\|\theta\|^2+B^2),\quad\forall\, i=1,\cdots,n.$$

Juntando con la desigualdad (5.4) tenemos que

$$\mathbb{E}_I\left(\left\|\frac{1}{h}\sum_{i\in I}\nabla f(\theta,z_i)-\frac{1}{n}\sum_{i=1}^{n}\nabla f(\theta,z_i)\right\|^2\right)\leq\frac{1}{h^2}\frac{h}{n}2n(L^2\|\theta\|^2+B^2)\leq\frac{2}{h}(L^2\|\theta\|^2+B^2),$$

cumpliéndose la Condición 3 con $G=L$, $N=B$ y $\delta=1/h$.

De esta forma, el algoritmo descrito en (5.2), para la función de log-verosimilitud, etá dado por

$$\theta_{k+1}=\theta_k+\frac{\eta}{h}\sum_{i\in I_k}\frac{1}{p_{\theta_k}(z_i)}\nabla_\theta p_{\theta_k}(z_i)+\sqrt{2\eta\beta^{-1}}\xi_k, \quad (5.5)$$

con $\{I_k\}_{k\geq 1}$ una colección de variables i.i.d. que distribuyen uniformemente sobre $\{1,\cdots,n\}^m$, $\{\xi_k\}_{k\geq 1}$ una colección de variables i.i.d. normales estándar $d$-dimensionales, independientes de $\{I_k\}_{k\geq 1}$, $\eta>0$ y $\beta>0$.

En el contexto de este problema nos interesa dos casos típicos de verosimilitud de una muestra: el primer caso ocurre cuando la función de densidad corresponde a una densidad posteriori $p(\theta|\mathcal{D})$. El estimador resultante de este problema es conocido como *Estimador Máximo a Posteriori* (MAP). El segundo caso se presenta cuando la variable $Z$ no es el único fenómeno aleatorio, sino que forma parte de algún vector aleatorio observable. En este caso será necesario marginalizar la densidad conjunta, integrando en las variables que no sean parte del problema de optimización.

## 5.1. Estimador máximo a posteriori

El principal propósito de este contexto es encontrar el parámetro $\theta$ de una ley de probabilidad de densidad $p_\theta$, que mejor explique una muestra $\mathcal{D}$ de datos i.i.d. finita de ley $p_\theta$. Sin embargo, es de esperarse que, para el conjunto de observaciones $\mathcal{D}$ fijo, el parámetro $\theta$ pueda tener, no sólo uno, sino un rango de valores y hasta su propia distribución de probabilidad. Este punto de vista tiene lógica bajo la idea de que la muestra $\mathcal{D}$ fue extraída aleatoriamente según la ley $p_\theta$, por lo que encontrar una ley óptima sólo para el conjunto finito $\mathcal{D}$ y no para cualquier muestra, podría llevarnos al sobreajuste.

Las siguientes definiciones extraídas de Held y Bové, 2013 (sección 6) nos ayudarán a introducir los principales elementos de la estadística bayesiana.

**Definición 5.1** *Sea $\mathcal{D}$ un conjunto de observaciones de una variable aleatoria $Z$ con función de densidad $p(z|\theta)$. Se define la densidad de la distribución a posteriori como*

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)\varsigma(\theta)}{\int p(\mathcal{D}|\theta)\varsigma(\theta)d\theta} \propto p(\mathcal{D}|\theta)\varsigma(\theta), \tag{5.6}$$

*donde la función $\varsigma(\theta)$ es una densidad de probabilidad para $\theta$, denominada distribución a priori.*

Supongamos que el conjunto $\mathcal{D} = \{z_i\}_{i=1}^n$ corresponde a variables *i.i.d.*, entonces podemos escribir la ley conjunta de $\mathcal{D}$ como

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{n} p(z_i|\theta).$$

Si modificamos el problema de verosimilitud, y se considera a la ley posteriori, $p(\theta|\mathcal{D})$, en vez de la verosimilitud, $L(\theta;\mathcal{D})$, entonces la función objetivo del problema (2.7) se transforma en[3]

$$\hat{F}_{\mathbf{z}}(\theta) := -\log\varsigma(\theta) - \frac{1}{n}\sum_{i=1}^{n}\log p(z_i|\theta), \tag{5.7}$$

A un valor de $\theta$ que resuelva esta nueva modificación del problema es conocido como *estimador máximo a posteriori*, pues maximiza la distribución posterior de la muestra. Desde el punto de vista del problema de minimización expuesto, $\hat{F}_{\mathbf{z}}$ sólo difiere del problema original de verosimilitud en el término $\log\varsigma(\theta)$, que representa un término de regularización adecuado al contexto de la naturaleza del parámetro $\theta$.

En Ma y col., 2015, Chen y col., 2014 y Welling y Teh, 2011, la función $\hat{F}_{\mathbf{z}}$ es llamada *función de energía potencial* y es posible analizarla desde el punto de vista de una dinámica Hamiltoniana, aprovechando algunas propiedades peculiares que se logran ver. Esto permite otros tipos de enfoques de estudio, sin embargo, este aspecto no será abarcado en este trabajo.

A continuación, veremos condiciones suficientes sobre los datos del problema que permitan aplicar los resultados del Capítulo 4, justificando el uso del algoritmo SGLD en este contexto.

1. **Condición de suavidad**

Para asegurar la Condición (1) bastará con pedir que, tanto $\log p(z_i|\theta)$ como $\log\varsigma(\theta)$ perte-

[3] El óptimo de $p(\theta|\mathcal{D})$ se mantiene para todas las funciones que sea proporcionales a ella, por lo que puedo intercambiarla por $p(\mathcal{D}|\theta)\varsigma(\theta)$ aunque no sean completamente iguales.

nezcan por separado a alguna familia de distribuciones tal que la derivada de su densidad sea una función Lipschitz. Si consideramos a $f(\theta, z_i) = -\log p(z_i|\theta) - \log \varsigma(\theta)$ en el contexto del problema (2.2), tenemos que; como la condición se mantiene bajo aditividad de funciones, $f(\theta, z_i)$ también las cumplirá.

En efecto, asumiendo la diferenciabilidad de $\log \varsigma(\theta)$ y $\log p(z, \theta)$ con respecto a la variable $\theta$, y suponiendo que ambos gradientes son Lipschizt para $\theta$, de constantes $L_1$ y $L_2$, respectivamente. Sea $z_i \in \mathcal{Z}$, arbitrario, entonces tenemos que

$$\|\nabla f(\theta_1, z_i) - \nabla f(\theta_2, z_i)\| =$$

$$\|\nabla \log \varsigma(\theta_2) - \nabla \log \varsigma(\theta_1) + \nabla \log p(z_i|\theta_2) - \nabla \log p(z_i|\theta_1)\|$$
$$\leq \|\nabla \log \varsigma(\theta_2) - \nabla \log \varsigma(\theta_1)\| + \|\nabla \log p(z_i|\theta_2) - \nabla \log p(z_i|\theta_1)\|$$
$$\leq L_1\|\theta_1 - \theta_2\| + L_2\|\theta_1 - \theta_2\| = (L_1 + L_2)\|\theta_1 - \theta_2\|.$$

Otra forma de corroborar esta condición es asegurar diferenciabilidad de segundo orden para estas funciones. En caso de poseer segunda derivada acotada, aplicando el teorema de valor medio, podemos encontrar opciones de valores para $L_1$ y $L_2$, de la forma

$$L_1 = \sup_{\theta} \nabla^2 \log \varsigma(\theta), \quad \text{y,} \quad L_2 = \sup_{\theta} \max_{i=1,\cdots,n} \nabla^2 \log p(z_i|\theta) \tag{5.8}$$

2. **Condición de disipatividad**

Para las Condición (2) el procedimiento es parecido. Suponiendo que $-\nabla \log p(z_i|\theta)$ cumple la condición de disipatividad con constantes $m_1$ y $b_1$ y $-\nabla \log \varsigma(\theta)$ la cumple con constantes $m_2$ y $b_2$, entonces si

$$\langle \theta\,,\, -\nabla \log p(z_i|\theta) \rangle \geq m_1\|\theta\|^2 - b_1,$$

$$\langle \theta\,,\, -\nabla \log \varsigma(\theta) \rangle \geq m_2\|\theta\|^2 - b_2$$

tenemos que

$$\langle \theta\,,\, -\nabla \log p(z_i|\theta) - \nabla \log \varsigma(\theta) \rangle \geq m_3\|\theta\|^2 - b_3, \tag{5.9}$$

con $m_3 = m_1 + m_2$ y $b_3 = b_1 + b_2$.

**Ejemplo** (Distribución Normal multivariada) Sean $\vec{\mu} \in \mathbb{R}^d$ y $X \sim \mathcal{N}(\vec{\mu}, \Sigma)$, donde $\Sigma$ es una matriz cuadrada de dimensiones $d \times d$, simétrica, invertible y definida positiva. Sea $\mathcal{X} = \{x_i\}_{i=1}^n$ una muestra finita de observaciones de $X$, suponiendo que el parámetro $\vec{\mu}$ es desconocido, pero que si poseemos la información de cómo es la matriz $\Sigma$, la función de verosimilitud dado el parámetro $\vec{\mu}$ está dada por

$$p(x|\vec{\mu}) = \frac{1}{(2\pi)^{n/2}\,|\det \Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \vec{\mu})^T\Sigma^{-1}(x - \vec{\mu})\right), \tag{5.10}$$

por lo tanto

$$-\nabla_{\vec{\mu}} \log p(x|\vec{\mu}) = \Sigma^{-1}(\vec{\mu} - x) \tag{5.11}$$

y

$$\langle \vec{\mu}\,,\, -\nabla_{\vec{\mu}} \log p(x|\vec{\mu}) \rangle = \vec{\mu}^T\Sigma^{-1}\vec{\mu} - \vec{\mu}^T\Sigma^{-1}x. \tag{5.12}$$

Cumplir la Condición (2), es decir, encontrar valores $m, b > 0$ tal que

$$\langle \vec{\mu}\,,\, -\nabla_{\vec{\mu}} \log p(x|\vec{\mu}) \rangle \geq m\|\vec{\mu}\|^2 - b,$$

es equivale a cumplir la siguiente desigualdad

$$\vec{\mu}^T \left(\Sigma^{-1} - m\mathbb{I}\right) \vec{\mu} - \vec{\mu}^T \Sigma^{-1} x + b \geq 0, \tag{5.13}$$

donde $\mathbb{I}$ es la matriz identidad de dimensiones $d \times d$. Sean $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_d > 0$ los valores propios de $\Sigma^{-1}$, entonces si tomamos $m \in (0, \alpha_d)$, la expresión a la izquierda de la desigualdad (5.13) es una forma cuadrática convexa para $\vec{\mu}$, cuyo valor mínimo es

$$-\frac{1}{4} x^T \Sigma^{-1} \left(\Sigma^{-1} - m\mathbb{I}\right)^{-1} \Sigma^{-1} x + b.$$

Imponiendo que esta última expresión sea mayor que cero y tomando el máximo en $x \in \mathcal{X}$ para que la desigualdad se cumpla en todo punto de la muestra, tenemos que si $m \in (0, \alpha_d)$, entonces el $b > 0$ necesario cumple que

$$b > \frac{1}{4} \max_{x \in \mathcal{X}} \; x^T \Sigma^{-1} \left(\Sigma^{-1} - m\mathbb{I}\right)^{-1} \Sigma^{-1} x. \tag{5.14}$$

3. **Otras Condiciones**

Las condiciones faltantes: desigualdad (4.2) de la Condición (1), Condición (3), Condición (4) y Condición (5), dependerán directamente del caso particular que se tome, es decir; de la verosimilitud, densidad a priori ocupada, el estimador de $\nabla F_{\mathbf{z}}(\theta)$ y la elección inicial del algoritmo.

De esta forma, aplicando el algoritmo especificado en (5.5) para este caso, tenemos la siguiente regla de actualización para $\theta_k$:

$$\theta_{k+1} = \theta_k + \frac{\eta}{\varsigma(\theta_k)} \nabla_\theta \varsigma(\theta_k) + \frac{\eta}{h} \sum_{i \in I_k} \frac{1}{p(z_i|\theta_k)} \nabla_\theta p(z_i|\theta_k) + \sqrt{2\eta\beta^{-1}}\xi_k. \tag{5.15}$$

El algoritmo anterior puede verse bastante beneficiado en casos en que las leyes son tomadas de algún caso particular (pero común) como lo son las familias exponenciales, donde perdemos la dependencia del término divisor en el paso y el cálculo de los gradientes llega a simplificarse de una manera importante, como lo es (por ejemplo) en los modelos normales.

## 5.2. Distribución marginal desde distribuciones conjuntas

Actualmente es difícil pensar en modelos matemáticos que describan de buena forma un fenómeno natural o social con tan sólo una variable. Lo usual y conveniente es encontrar modelos multidimensionales. De esta forma, si nuestro interés es estimar una distribución de alguna variable $Z$ no observada, será natural buscar una regla, dependencia o distribución conjunta con respecto a otra variable $Y$, cuyas realizaciones sí se observan.

Sea $p_\theta(z, y)$ la distribución conjunta del vector aleatorio $(Z, Y)$, entonces la distribución

marginal, objeto de interés del problema de verosimilitud, está dada por:

$$p_\theta(z) = \int p_\theta(z, y)dy. \tag{5.16}$$

Al aplicarle logaritmo a esta función y asumir la diferenciabilidad suficiente por parte de la función $\theta \mapsto p_\theta(z, y)$, tenemos que el gradiente estará dado por la expresión

$$\nabla_\theta \log p_\theta(z) = \frac{\int \nabla_\theta p_\theta(z, y)dy}{\int p_\theta(z, y)dy}, \tag{5.17}$$

Sobre la cual se puede verificar si se cumplen las condiciones revisadas en los capítulos anteriores.

A diferencia del caso anterior, con la densidad posteriori, en el caso de la densidad marginal no hay variadas generalidades que se puedan tomar más que ver en cada caso cuál es la forma de la función de log-verosimilitud (5.17). Puede ocurrir que el espacio donde se encuentra definido la variable $Y$ sea acotado o de medida finita, en tal caso, la condición de suavidad puede caer sólo sobre el integrando de la expresión anterior. En efecto, sean $\theta_1$ y $\theta_2$, supongamos que la función $\frac{\nabla_\theta p_\theta(z,y)}{p_\theta(z)}$ es localmente Lipschitz, de constante $L > 0$, para la variable $\theta$, y que la variable aleatoria $Y$ sólo toma valores en el conjunto acotado $\mathcal{Y}$ con medida de Lebesgue finita $\#(\mathcal{Y})$, entonces

$$\begin{aligned}\left\| \int_{\mathcal{Y}} \frac{\nabla_\theta p_{\theta_1}(z, y)}{p_{\theta_1}(z)} dy - \int_{\mathcal{Y}} \frac{\nabla_\theta p_{\theta_2}(z, y)}{p_{\theta_2}(z)} dy \right\| &\leq \int_{\mathcal{Y}} \left\| \frac{\nabla_\theta p_{\theta_1}(z, y)}{p_{\theta_1}(z)} - \frac{\nabla_\theta p_{\theta_2}(z, y)}{p_{\theta_2}(z)} \right\| dy \\ &\leq \int_{\mathcal{Y}} L\|\theta_1 - \theta_2\| dy = \#(\mathcal{Y})L\|\theta_1 - \theta_2\|. \end{aligned} \tag{5.18}$$

# Capítulo 6

# Aplicación: Tiempo mínimo de viaje y tomografía sísmica

A continuación, se establecerá el contexto del problema inverso del tiempo mínimo de viaje para ondas sísmicas en un dominio acotado, que consiste en reconstruir la distribución de velocidades de onda del medio por donde se propagan a partir de una base de registros de llegadas de sismos a un conjunto de sensores. Este problema, conocido como el problema de tomografía sísmica, típicamente es un problema inverso mal puesto, su modelamiento y soluciones han representado un reto en el área de las matemáticas y datos, por lo que un método de solución numérico será el desafío propuesto para el algoritmo SGLD.

En los siguientes capítulos se modelará la función del tiempo mínimo de viaje de una onda sísmica y su relación con el campo de lentitud fijo en un medio, y cómo estos elementos pueden incluirse en un modelo paramétrico de estimación. Conseguir el nivel de diferenciación necesaria para aplicar el algoritmo SGLD y analizar de qué forma se podría (o no) verificar las condiciones necesarias para aplicar el resultado mostrado en capítulos anteriores.

## 6.1. Tiempo mínimo de viaje y campo de lentitud

El siguiente capítulo recopila y resume el desarrollo expuesto en Delplancke y col., 2023 para la comprensión de la elección del modelo de prueba del algoritmo SGLD y las hipótesis necesarias para su uso. En lo que sigue se trabajará en un espacio tridimensional, sin embargo, esta teoría es aplicable en dos dimensiones, caso que se observará en los resultados de la aplicación.

Sean $a, b \in \mathbb{R}^3$ dos puntos, se define el conjunto de los caminos entre $a$ y $b$ como $\mathcal{P}_{a,b}$, es decir; si $p \in \mathcal{P}_{a,b}$ entonces $p = \{\gamma(t) \,:\, t \in [0, t_\gamma]\}$, donde $\gamma$ es una parametrización de este camino, Lipschizt continua, y $t_\gamma$ es el tiempo en que se demora en llegar por primera vez desde $a$ hasta $b$, es decir; $\gamma(0) = a$ y $\gamma(t_\gamma) = b$, con $\gamma(t) \neq b$ para todo $t \in (0, t_\gamma)$.

Además se denota por $\mathcal{B}(\mathbb{R}^3, (0, \infty))$ en conjunto de las funciones medibles y acotadas desde $\mathbb{R}^3$ al conjunto $(0, \infty)$. Se dota a $\mathcal{B}(\mathbb{R}^3, (0, \infty))$ por la norma uniforme $\| \cdot \|_\infty$ y el producto natural entre él y su dual, $\langle \cdot, \cdot \rangle$.

Para cada $p \in \mathcal{P}_{a,b}$ se define la aplicación lineal $\Gamma_p : \mathcal{B}(\mathbb{R}^3, (0, \infty)) \to \mathbb{R}$,a la que a cada campo $S \in \mathcal{B}(\mathbb{R}^3, (0, \infty))$, le asigna su integral a través del camino $p$, es decir

$$\langle \Gamma_p, S \rangle := \int_p S = \int_0^{t_\gamma} S(\gamma(t)) \|\gamma'\| dt. \tag{6.1}$$

Es posible mostrar (ver Delplancke y col., 2023) que la aplicación $\Gamma_p$ no depende de la parametrización $\gamma$, es continua y puede identificarse como una medida de Radon sobre $\mathbb{R}^3$. De esta forma se define, en función de $S \in \mathcal{B}(\mathbb{R}^3, (0,\infty))$, la ecuación del *tiempo mínimo de viaje* entre los puntos $a$ y $b$ a través de $S$ como

$$d_S(a,b) = \inf_{p \in \mathcal{P}_{a,b}} \langle \Gamma_p, S \rangle, \tag{6.2}$$

donde $S$ es interpretado como un campo de lentitud finito, es decir, el inverso de una velocidad no nula, isotrópica y finita en $\mathbb{R}^3$. Si $\gamma : [0, t_\gamma] \to \mathbb{R}^3$ es una parametrización Lipschitz con una lentitud establecida por el campo $S \in \mathcal{B}(\mathbb{R}^3, (0,\infty))$, se cumple que $\|\gamma'(t)\| S(\gamma(t)) = 1$. Por lo tanto, si $\gamma(0) = a$, $\gamma(t_\gamma) = b$ y $\gamma(t) \neq b$ para todo $t \in (0, t_\gamma)$, el tiempo de viaje desde $a$ hasta $b$ estará dado por

$$\langle \Gamma_p, S \rangle = t_\gamma,$$

donde $p \in \mathcal{P}_{a,b}$ es el camino parametrizado por $\gamma$.

Así, para $S \in \mathcal{B}(\mathbb{R}^3, (0,\infty))$, el valor $d_S(a,b)$ representa el tiempo mínimo de viaje de una partícula mediante un camino Lipschitz entre $a$ y $b$ cuya velocidad en el punto $x \in \mathbb{R}^3$ está dado por $1/S(x)$.

También es posible trabajar este mismo valor, $d_S(a,b)$, como la solución viscosa de la ecuación eiokonal, como se hace en Lee y col., 1981, Noble y col., 2014 y Tarantola y Valette, 1982. En efecto, para $a \in \mathbb{R}^3$ fijo, si $S$ toma valores en la cerradura de $G \subset \mathbb{R}^3$, al igual que los caminos que unen $a$ con $b$. Si $S$ es suave tendremos que $d_S(a,\cdot)$ es la solución de la ecuación

$$\|\nabla u(b)\| = S(b), \quad \forall\, b \in G, \;\; u(a) = 0. \tag{6.3}$$

A continuación, se buscará realizar estimaciones sobre el campo $S(x)$, para todo $x$ es un conjunto acotado $G \subset \mathbb{R}^3$, mediante estimadores de máxima verosimilitud, por lo que es necesario imponer un modelo probabilístico sobre alguna base de datos relacionadas con este fenómeno físico. Además, como el campo $S$ tiene infinitos valores en un subconjunto de $\mathbb{R}^3$, será necesario discretizar el dominio en el que se encuentren los datos disponibles e interpretar el resultado anterior para poder derivar la función de verosimilitud y aplicar el algoritmo SGLD con el parámetro adecuado.

## 6.2. Modelo de estimación de $S$

Tanto la construcción de la base de datos, como el modelo paramétrico de estimación del campo es tomado de Delplancke y col., 2020, donde se desarrolló un método basado en el algoritmo SGD clásico. Los resultados de ese experimento serán citados más adelante como comparación con respecto al rendimiento obtenido con el algoritmo SGLD.

Considerando un campo de lentitud fijo $S \in \mathcal{B}(\mathbb{R}^3, (0,\infty))$, se centra el análisis en un conjunto acotado $G \subset \mathbb{R}^3$ desde el cual se han realizado registros de la posición y tiempo en que $M \in \mathbb{N}$ sismos independientes, de origen espacio-temporal desconocido, llegan a un conjunto de sensores. Cada sismo es etiquetado con el super-índice $i \in \{1, \cdots, M\}$ y cada registro consiste en un par ordenado, tiempo y posición, $(t_j^i, r_j^i)$, correspondientes al $j$-ésimo sensor ($j \in \{1, \cdots, N^i\}$) para cada uno de los $N^i \in \mathbb{N}$ sensores, en el $i$-ésimo sismo.

### 6.2.1. Parametrización del campo

Dado que el campo $S \in \mathcal{B}(\mathbb{R}^3, (0,\infty))$ es un parámetro continuo en función de la posición espacial y los modelos de verosimilitud convencionales sólo ocupan parámetros de dimensión finita, es necesario un método de discretización para $S$. Sea $G \subseteq \mathbb{R}^3$, el conjunto acotado contiene a todas las posiciones de los sensores que recolectaron la información de los sismos, $r_j^i$. Particionando el conjunto $G$ en $d$ subdominios y definiendo un valor para el campo de manera constante en cada subdominio, podemos expresar el campo $S$ como un elemento de $\mathbb{R}^d$. En Delplancke y col., 2023 se muestra que existe una relación lineal e inyectiva entre $S \in \mathcal{B}(G, (0,\infty))$ y su parametrización construida como se acaba de mencionar, $s \in \mathbb{R}^d$. Esto permite definir la función cóncava $\mathcal{T} : \mathbb{R}_+^d \mapsto \mathbb{R}_+$, que expresa el tiempo mínimo de viaje entre los puntos $a$ y $b$ en función de $s$. El siguiente Teorema[4] de Delplancke y col., 2023, permite asegurar la diferenciabilidad de la función de $s$, una parametrización del campo $S$.

**Teorema 6.1** (Teorema 4.3 de Delplancke y col., 2023) *Sea $\hat{s} \in \mathbb{R}^d$ la parametrización por bloques respectiva de $S_{\hat{s}} \in \mathcal{B}(\mathbb{R}^3, [0,\infty))$, se define el cono abierto de parametrizaciones de campos no negativos como*

$$\mathcal{S}_J := \left\{ \hat{s} \in \mathbb{R}^d : S_{\hat{s}} \in \bigcup_{m>0} \mathcal{B}(\mathbb{R}^3, [m,\infty)) \right\}. \tag{6.4}$$

*Entonces, la función $s \mapsto \mathcal{T}(s)$ es diferenciable c.t.p. en $\mathcal{S}_J$. Más aún, para $s \in \mathcal{S}_J - c.t.p.$, existe un único vector $w(s) \in \mathbb{R}_+^d$ tal que*

$$\mathcal{T}(s) = \langle w(s), s \rangle_{\mathbb{R}^d} = d_S(a,b), \tag{6.5}$$

*y*

$$D_s\mathcal{T}(h) = \langle \nabla\mathcal{T}(s), h \rangle_{\mathbb{R}^d} = \langle w(s), h \rangle_{\mathbb{R}^d}, \quad \forall\, h \in \mathbb{R}^d. \tag{6.6}$$

### 6.2.2. Función de verosimilitud

Sea $\{t_j^i : 1 \le j \le N^i, 1 \le i \le M\}$ el conjunto de los tiempos correspondientes a $M$ sismos registrados. $t_j^i$ corresponde al tiempo de la primera vez en que se registra la $i$-ésima onda sísmica por el $j$-ésimo sensor (del total de $N^i$ sensores para tal sismo) ubicado en el punto espacial $r_j^i$.

Se asume que cada $t_j^i \in [0,\infty)$ corresponden a valores positivos y $r_j^i \in \mathbb{R}^3$, y que existen orígenes de cada sismo $(t^i, r^i)$ que serán parámetros desconocidos. Además, se asume que cada sismo se comporta de manera independiente con los $M-1$ restantes y la relación entre estos tiempos y el capo de lentitud $s$ es de la forma[5]

$$t_j^i = t^i + d_s(r^i, r_j^i) + \tilde{\epsilon}_j^i, \;\; 1 \le j \le N^i, \tag{6.7}$$

donde $\tilde{\epsilon}_j^i$, $i \in \{1, \cdots, M\}$, $j \in \{1, \cdots, \mathbb{N}^i\}$, representan medidas de error de medición correspondientes a variables aleatorias gaussianas centradas de varianza común $\sigma^2 > 0$. En palabras simples, el tiempo registrado debe ser igual al tiempo de referencia origen del sismo más el tiempo mínimo de viaje ente el punto de origen del sismo $r^i$ y el punto de registro del sensor

[4] La demostración del Teorema 6.1 se puede encontrar en el Anexo A.

[5] Desde ahora y en adelante se tomarán sólo campos discretizados, justificado por lo mencionado anteriormente.

$r_j^i$, perturbado por un error de medición independiente del fenómeno, donde $\sigma$ está definido completamente por la calibración de los sensores.

Tal como se menciona en Delplancke y col., 2020 y Tarantola y Valette, 1982, es posible modificar los datos disponibles para perder la dependencia del origen temporal de los sismos. Para ello se trabajará con vectores de datos temporales centrados

$$\mathbf{t}^i := \left( t_j^i - \frac{1}{N^i} \sum_{j=1}^{N^i} t_j^i \right)_{1 \leq j \leq N^i},$$

$$\mathbf{F}(r^i, s) := \left( d_s(r^i, r_j^i) - \frac{1}{N^i} \sum_{j=1}^{N^i} d_s(r^i, r_j^i) \right)_{1 \leq j \leq N^i}.$$

Con esta modificación se satisface la siguiente ecuación

$$\mathbf{t}^i = \mathbf{F}(r^i, s) + \epsilon^i, \quad \forall\, i = 1, \cdots, M \tag{6.8}$$

donde $\epsilon^i$ son vectores gaussianos $N^i$-dimensionales. Si denotando por $I_{N^i}$, la matriz identidad de $\mathbb{R}^{N^i} \times \mathbb{R}^{N^i}$ y por $\mathbf{1}_{N^i}$, a la matriz cuadrada en $\mathbb{R}^{N^i} \times \mathbb{R}^{N^i}$ que sólo contiene 1 en sus entradas, entonces la matriz de covarianzas de $\epsilon^i$ estará dada por

$$\Sigma^i := \sigma^2 \left( I_{N^i} - \frac{1}{N^i} \mathbf{1}_{N^i} \right). \tag{6.9}$$

Dado que la matriz $\Sigma^i$ es singular y no es invertible, el vector $\epsilon^i$ no posee densidad en $\mathbb{R}^{N^i}$. Sin embargo, como se menciona en el Capítulo 8 de Rao, 1973, es posible encontrar una densidad para este vector Gaussiano en un subespacio vectorial contenido en $\mathbb{R}^{N^i}$. En el siguiente lema de Delplancke y col., 2023, cuya demostración se puede encontrar en el Anexo B, se muestra que este subespacio es aquel espacio ortogonal al vector $(1, \cdots, 1)$ en $\mathbb{R}^{N^i}$. Por lo tanto, es posible encontrar función de densidad concentrada en $\mathbb{R}^{N^i - 1}$.

**Lema 6.1** (Lema 3.1 en Delplancke y col., 2023) *Para cada sismo $i = \{1, \cdots, M\}$, sea $\mathcal{H}$ el hiperplano de $\mathbb{R}^{N^i}$ ortogonal al vector $(1, \cdots, 1)$, $\epsilon_i$ un vector gaussiano en $\mathbb{R}^{N^i}$ con matriz de covarianzas $\Sigma^i$, y $\bar{\epsilon}_i \in \mathbb{R}^{N^i - 1}$ las coordenadas de $\epsilon_i$ con respecto a una base ortonormal de $\mathcal{H}$ dada. Entonces $\bar{\epsilon}_i \in \mathbb{R}^{N^i - 1}$ admite la siguiente función de densidad*

$$\left( \frac{1}{\sqrt{2\pi}\sigma} \right)^{N^i - 1} \exp\left( -\frac{\|\bar{\epsilon}^i\|^2}{2\sigma^2} \right) = \left( \frac{1}{\sqrt{2\pi}\sigma} \right)^{N^i - 1} \exp\left( -\frac{\|\epsilon^i\|^2}{2\sigma^2} \right),$$

*donde $\|\cdot\|$ es la norma Euclideana en $\mathbb{R}^{N^i - 1}$ o $\mathbb{R}^{N^i}$.*

Por los argumentos mencionados anteriormente, la densidad del vector Gaussiano $\mathbf{t}^i$, condicionado al valor de $s \in \mathbb{R}_+^d$ y al punto de origen del sismo $r^i$ es

$$p(\mathbf{t}^i | r^i, s) = \left( \frac{1}{\sqrt{2\pi}\sigma} \right)^{N^i - 1} \exp\left( -\frac{\|\mathbf{t}^i - \mathbf{F}(r^i, s)\|^2}{2\sigma^2} \right). \tag{6.10}$$

Como no se tiene información alguna sobre los parámetros de hipocentro de origen del sismo será necesario imponer una distribución a priori para encontrar, marginalmente, la distribu-

ción de $\mathbf{t}^i$ condicionada ahora sólo a $s$. Esta es

$$p(\mathbf{t}^i|s) = \int p(\mathbf{t}^i, r^i|s)dr^i = \int p(\mathbf{t}^i|r^i, s)p_{prior}(r^i)dr^i, \tag{6.11}$$

donde $p_{prior}(r^i)$ es la densidad a priori para los hipocentros $r^i$.

Finalmente, ocupando que los sismos aparecen de manera independiente, e imponiendo una distribución a priori para el campo $s$ esta vez, podemos encontrar una distribución conjunta para la base de datos $\mathcal{D} = \{\mathbf{t}^i\}_{i=1}^M$. Gracias a la independencia de los sismos, definimos como

$$p(\mathcal{D}|s) = \prod_{i=1}^{M} p(\mathbf{t}^i|s), \tag{6.12}$$

a la verosimilitud de $\mathcal{D}$, condicionada a $s$. Llamando a $p_{prior}(s)$ a la distribución a priori para $s$, por lo visto en el capítulo 5.1, obtenemos nuestra función de riesgo esperado empírico, a minimizar, dada por

$$\begin{aligned} L(s, \mathcal{D}) &= -\log \varsigma(s) - \frac{1}{M}\sum_{i=1}^{M} \log p(\mathbf{t}^i|s) \\ &= -\log \varsigma(s) - \frac{1}{M}\sum_{i=1}^{M} \log \int p(\mathbf{t}^i|r^i, s)p_{prior}(r^i)dr^i. \end{aligned} \tag{6.13}$$

Vale la pena notar que, dada la distribución a priori impuesta para $r^i$, la integral que aparece en la función de pérdida no necesariamente tiene que estar integrando todo $\mathbb{R}^3$. En la práctica, esta densidad tendrá soporte en un conjunto acotado $G$, lo suficientemente grande.

La forma en que se buscará optimizar esta función objetivo es el algoritmo SGLD. Como el campo $S$ es una función de $\mathbb{R}^3$ a $\mathbb{R}_+$, es necesario generar una discretización del campo en algún subconjunto de su dominio para generar una estimación, por lo tanto, también será necesario conocer de qué forma representar el diferencial $D_S$ en esta discretización, para ser usado en el algoritmo de gradiente.

## 6.3. Discretización del campo de lentitud y existencia del gradiente

Ya obtenida la función de log-verosimilitud en (6.13), el algoritmo amerita el cálculo del gradiente de la función objetivo para el algoritmo SGLD. Afortunadamente, el Teorema 5.1 de Delplancke y col., 2023 nos asegura la existencia de $\nabla_s p(\mathbf{t}^i|s)$.

**Teorema 6.2** (Teorema 5.1 de Delplancke y col., 2023) *El mapeo $s \mapsto p(\mathbf{t}^i|s)$, definido en (6.11), es diferenciable en casi todo punto del cono $\mathcal{S}_J \subseteq \mathbb{R}^d$, definido en (6.4). Más aún, tenemos que*

$$\nabla_s \log p(\mathbf{t}^i|s) = \int_G (\nabla_s \log p(\mathbf{t}^i|r^i, s))p_{post}(r^i|\mathbf{t}^i, s)dr^i, \tag{6.14}$$

*donde*

$$p_{post}(r^i|\mathbf{t}^i, s) \propto \exp\left(-\frac{\|\mathbf{t}^i - \mathbf{F}(r^i, s)\|^2}{2\sigma^2}\right) p_{prior}(r^i) \tag{6.15}$$

*y*

$$\nabla_s \log p(\mathbf{t}^i|r^i, s) = \sigma^{-2}[\nabla_s \mathbf{F}(r^i, s)]^T(\mathbf{t}^i - \mathbf{F}(r^i, s)), \tag{6.16}$$

*con $\nabla_s \mathbf{F}(r^i, s)$ la matriz de dimensión $N^i \times d$, en que la $j$-ésima fila está dada por el vector*

$$w_j^i(s) - \frac{1}{N^i} \sum_{i=1}^{N^i} w_j^i(s) \in \mathbb{R}^d, \tag{6.17}$$

*y donde $w_j^i(s)$ es aquel el único elemento de $\mathbb{R}^d$ tal que $d_S(r^i, r_j^i) = \langle w_j^i(s), s \rangle_{\mathbb{R}^d}$, mencionado en el Teorema 6.1, cuando $s \in \mathbb{R}^d$ es la parametrización constante por bloques del campo de lentitud $S \in \mathcal{B}(G, (0, \infty))$.*

La demostración del Teorema 6.2 se encuentra disponible en el Anexo C. En base a lo descrito, ya es posible formular el algoritmo SGLD para tomografía sísmica, resolviendo el problema de máxima verosimilitud equivalente a resolver el siguiente problema de minimización

$$\min_{s \in \mathbb{R}_+^d} L(s, \mathcal{D}), \tag{6.18}$$

donde $L(s, \mathcal{D})$ es la función expresada en (6.13).

## 6.4. Aplicación del algoritmo

La sección anterior explica las razones por la cual el gradiente de la función objetivo depende fuertemente de los factores $w_j^i$ que variarán según la parametrización utilizada en la discretización del campo. Por lo tanto, lo primero en el algoritmo será establecer la parametrización utilizada, en pos de identificar los parámetros $w_j^i$.

Idealmente, las funciones objetivos de los problemas de optimización en los que se plantea el uso del algoritmo SGLD son tales que se puedan cumplir la condiciones necesarias para presentar los resultados anteriores (como el del Teorema 4.4). Sin embargo, en el presente contexto no es posible asegurar cada una de estas condiciones. Por este motivo se discutirán formas en que el algoritmo sigue funcionando si se relajan o se cambian condiciones por otras de la misma naturaleza, condiciones más débiles pero fáciles de hallar en casos de la aplicación o simulación computacional.

### 6.4.1. Parametrización por bloques

El Ejemplo 4.1 de Delplancke y col., 2023 se esquematiza el siguiente método de parametrización por bloques. Sean $(H_k)_{1 \leq k \leq d}$ una partición de $G \subseteq \mathbb{R}^3$ en $d \in \mathbb{N}$ bloques, para cada $s \in \mathbb{R}^d$ se escribirá el campo de lentitud de la forma $S(x) = \sum_{k=1}^d s_k \mathbb{1}_{H_k}(x)$, para todo $x \in G$.

Para $a, b \in G$ fijos, sea $\bar{p} \in \mathcal{P}_{a,b}$ el camino que minimiza el tiempo de viaje entre $a$ y $b$, entonces tenemos que

$$\begin{aligned} d_S(a,b) &= \textstyle\int_{\bar{p}} S \\ &= \textstyle\sum_{k=1}^d s_k \int_{\bar{p}} \mathbb{1}_{H_k} \\ &= \textstyle\sum_{k=1}^d s_k Len(\bar{p} \cap H_k) \\ &= \langle (Len(\bar{p} \cap H_k))_{1 \leq k \leq d} \ , \ s \rangle_{\mathbb{R}^d}, \end{aligned} \tag{6.19}$$

donde $Len(\bar{p} \cap H_k)$ es el largo total del segmento de $\bar{p}$ que cruza por el sector de la partición $H_k$. De esta forma, $w_j^i(s) = (Len(\bar{p} \cap H_k))_{1 \leq k \leq d}$ cuando $\bar{p}$ sea el camino que minimice el tiempo de viaje entre $r^i$ y $r_j^i$ del sismo que pase a través del campo parametrizado $s$.

El algoritmo SGLD para la estimación de $s \in \mathbb{R}^d$ quedará definido por la siguiente iteración

$$s_{k+1} = s_k - \eta \nabla_s \tilde{L}(s, \mathcal{D}) + \sqrt{2\eta\beta^{-1}}\xi_k, \tag{6.20}$$

donde $\tilde{L}(s, \mathcal{D})$ es una estimador insesgado de la función de riesgo $L(s, \mathcal{D})$. Recordando la regla de actualización definida en (5.15), el método de batches como estimador insesgado del gradiente, y el gradiente de la log-verosimilitud dado por (6.14), la forma expandida del algoritmo en (6.20) se expresa como

$$s_{k+1} = s_k + \frac{\eta}{\varsigma(s)} \nabla_s \varsigma(s_k) + \frac{\eta}{h} \sum_{i \in I_k} \int_G \left( \nabla_s \log p(\mathbf{t}^i | r^i, s) \right) p_{post}(r^i | \mathbf{t}^i, s) dr^i + \sqrt{2\beta^{-1}\eta}\xi_k, \tag{6.21}$$

donde $\varsigma(s)$ es la distribución a priori (diferenciable) de $s$, $I_k$ es el batch de tamaño $m$ seleccionado en la iteración $k$-ésima, $\xi_k$ son variables normales estándar multivariada en $\mathbb{R}^d$ y $G$ es el conjunto acotado en que se encuentran los hipocentros $r^i$.

### 6.4.2. Condición de suavidad

Para asegurar que el comportamiento asintótico del algoritmo converja a un punto que minimice la función de riesgo, es necesario verificar las condiciones solicitadas a la función objetivo para recurrir al resultado del Teorema 4.4 y estimar un sesgo. En ocasiones, estas condiciones no se cumplen con la misma rigurosidad en que fueron enunciadas en Sección 4.1, sino que se recurre a argumentos de finitud de iteraciones, conjuntos de observaciones finitos y conjuntos de parámetros acotados. Ciertamente, tomar conjuntos acotados de parámetros añade restricciones que inicialmente no son consideradas en el problema original pero que nunca serán restricciones activas en los puntos óptimos encontrados en el algoritmo, tan sólo conocemos su existencia por el contexto del problema y del hecho de que; como la cantidad de iteraciones es finita, también lo será la cantidad de estimaciones encontradas. Actualmente en la literatura no existe el análisis de la convergencia del algoritmo SGLD cuando hablamos de un conjunto de parámetros que deba responder a cierta restricción, sino que se considera completamente el conjunto $\mathbb{R}^d$, esto no suele ocurrir en la práctica.

Si observamos la forma de la función $\nabla_s \log p(\mathbf{t}^i | s)$,

$$\nabla_s \log p(\mathbf{t}^i | s) = \int_G (\nabla_s \log p(\mathbf{t}^i | r^i, s)) p_{post}(r^i | \mathbf{t}^i, s) dr^i, \tag{6.22}$$

notamos que corresponde a una función que integra sobre todos los puntos posibles de hipocentros $r^i$. Esto impulsa a utilizar la observación (5.18), de la sección 5.2, que nos indica que podemos obtener una constante de Lipschitz para $\nabla_s \log p(\mathbf{t}^i | s)$ si el integrando de la función también resulta ser una función Lipschitz, dado que el conjunto en que se está integrando, $G$, es acotado. Sin embargo, probar que la función

$$s \longmapsto (\nabla_s \log p(\mathbf{t}^i | r^i, s)) p_{post}(r^i | \mathbf{t}^i, s)$$

es una función uniformemente Lipschitz en $r^i$ y $\mathbf{t}^i$, no es una tarea sencilla. Para abordar este problema, se recurre a la observación (5.8), del capítulo 5.1, donde se concluye que si

la función posee segunda derivada acotada, entonces tendremos un candidato a constante de Lipschitz.

El procedimiento mencionado equivale a verificar la diferenciabilidad de orden superior del mapeo $s \mapsto \mathcal{T}(s)$, definido en (6.5). Afortunadamente, dado lo mencionado en la sección 6.2.1, la función $\mathcal{T}(s)$ es una función cóncava, con esta información podemos recurrir un resultado de Alberti y Ambrosio, 1999 referentes a funciones convexas, también aplicable a funciones cóncavas.

**Teorema 6.3** (Proposición 7.11 en Alberti y Ambrosio, 1999) *Sea $\mathcal{R}$ un subconjunto convexo y abierto en $\mathbb{R}^d$. Si $f : \mathcal{R} \to \mathbb{R}$ es una función convexa, entonces $Df$ es un operador monótono, y $D^2 f$ es una medida positiva y simétrica (a valores en las matrices cuadradas de dimensión $d \times d$ y localmente acotada). Por otro lado, si $f \in L^1_{loc}(\mathcal{R})$ y $D^2 f$ es una distribución positiva (a valores en matrices) en $\mathcal{R}$, entonces $f$ coincide en casi todo punto de $\mathcal{R}$ con una función convexa $g$ con $\mathcal{R} \subset Dom\, g$.*

El teorema anterior nos asegura la existencia de $D_s^2\mathcal{T}$ para casi todo punto $s \in Dom\mathcal{T}$. Esto motiva a la búsqueda de

$$\nabla_s \left( (\nabla_s \log p(\mathbf{t}^i | r^i, s)) p_{post}(r^i | \mathbf{t}^i, s) \right) =$$

$$\nabla_s \left( \sigma^{-2} [\nabla_s \mathbf{F}(r^i, s)]^T (\mathbf{t}^i - \mathbf{F}(r^i, s)) p_{post}(r^i | \mathbf{t}^i, s) \right).$$

Para $a, b \in \mathbb{R}^3$, fijos, sea $w(s) \in \mathbb{R}^d$ el vector tal que, para la parametrización del campo $S$, $s \in Dom\mathcal{T} \subset \mathbb{R}^d$, se cumple que

$$\mathcal{T}(s) = \langle w(s), s \rangle_{\mathbb{R}^d}, \quad \text{y} \quad \langle \nabla \mathcal{T}(s), h \rangle_{\mathbb{R}^d} = \langle w(s), h \rangle_{\mathbb{R}^d}, \ \forall h \in \mathbb{R}^d. \tag{6.23}$$

Dado que la función $\mathcal{T}$ es cóncava, por el Teorema 6.3, $D_s^2\mathcal{T}$ existe, es una aplicación bilineal y cumple que

$$D_s^2 \mathcal{T}(s)[h, v] = \langle \nabla^2 \mathcal{T}(s) v, h \rangle_{\mathbb{R}^d}, \quad \forall v, h \in \mathbb{R}^d, \tag{6.24}$$

donde $\nabla^2 \mathcal{T}(s) = \nabla w(s)$ es una matriz cuadrada de dimensiones $d \times d$, simétrica y definida negativa (en caso de la función cóncava). De esta forma, caracterizamos a $\nabla_s^2 \mathbf{F}(r^i, s)$ como el tensor de rango 3, es decir, un arreglo vectorial de dimensiones $N^i \times d \times d$, tal que en la $j$-ésima posición en la primer dimensión se sitúa a la matriz

$$\nabla w_j^i(s) - \frac{1}{N^i} \sum_{i=1}^{N^i} \nabla w_j^i(s) \ \in \ \mathbb{R}^{d \times d}, \tag{6.25}$$

donde $w_j^i$ corresponde a los vectores definidos por 6.5.

Con lo anterior mencionado, definimos la derivada de segundo orden de la función objetivo como

$$\nabla_s^2 \log p(\mathbf{t}^i | s) = \int_G \nabla_s \left( (\nabla_s \log p(\mathbf{t}^i | r^i, s)) p_{post}(r^i | \mathbf{t}^i, s) \right) dr^i \tag{6.26}$$

$$= \sigma^{-2} \int_G p_{post}(r^i | \mathbf{t}^i, s) \left( [\nabla_s^2 \mathbf{F}(r^i, s)]^T (\mathbf{t}^i - \mathbf{F}(r^i, s)) - [\nabla_s \mathbf{F}(r^i, s)]^T \nabla_s \mathbf{F}(r^i, s) \right) \tag{6.27}$$

$$+ \nabla_s p_{post}(r^i | \mathbf{t}^i, s) \left( [\nabla_s \mathbf{F}(r^i, s)]^T (\mathbf{t}^i - \mathbf{F}(r^i, s)) \right)^T dr^i, \tag{6.28}$$

donde la operación $[\nabla_s^2 \mathbf{F}(r^i, s)]^T(\mathbf{t}^i - \mathbf{F}(r^i, s))$ se define como la matriz cuadrada de dimensiones $d \times d$ tal que su $(m, k)$-componente es

$$[\nabla_s^2 \mathbf{F}(r^i, s)]^T(\mathbf{t}^i - \mathbf{F}(r^i, s))_{m,k} = \sum_{j=1}^{N^i} \frac{\partial^2 F_j(s)}{\partial s_m \partial s_k}(t_j^i - F_j(r^i, s)), \tag{6.29}$$

donde

$$F_j(r^i, s) = d_S(r^i, r_j^i) - \frac{1}{N^i}\sum_{j=1}^{N^i} d_S(r^i, r_j^i), \quad \text{y} \quad \left(\nabla_s^2 \mathbf{F}(r^i, s)\right)_{j,m,k} = \frac{\partial^2 F_j(s)}{\partial s_m \partial s_k}.$$

Y el producto $\nabla_s p_{post}(r^i|\mathbf{t}^i, s)\left([\nabla_s \mathbf{F}(r^i, s)]^T(\mathbf{t}^i - \mathbf{F}(r^i, s))\right)^T$ es la multiplicación del vector $\nabla_s p_{post}(r^i|\mathbf{t}^i, s)$, de dimensión $d\times 1$, con el vector $\left([\nabla_s \mathbf{F}(r^i, s)]^T(\mathbf{t}^i - \mathbf{F}(r^i, s))\right)^T$, de dimensión $1 \times d$, formando una matriz cuadrada de dimensiones $d \times d$.

Basta la verificación de que $\nabla_s^2 \log p(\mathbf{t}^i|s)$ es localmente acotado, sumado a la precompacidad de $G$, es posible asegurar una constante para el criterio de suavidad del algoritmo SGLD.

**Observación:** Aún cuando sólo hemos verificado la existencia de $D^2\mathcal{T}(s)$ (mas no una expresión cerrada), podemos notar que: para $a, b \in \mathbb{R}^3$ fijos y sea $s \in \mathbb{R}_+^d$ un punto donde $D^2\mathcal{T}(s)$ existe, como $w(s)$ es aquel vector que minimiza el tiempo mínimo de viaje entre $a$ y $b$, bajo el campo de lentitud $s$, entonces el valor $\langle w(s), s\rangle$ es mínimo por sobre cualquier elección de $s$. Más aún, por la unicidad de $w(s)$, este valor también es mínimo bajo cualquier elección de $w \in \mathbb{R}^d$. Entonces, tomando $h \in \mathbb{R}_+^d$ y $t > 0$ suficientemente pequeño para que $s + th \in \mathbb{R}_+^d$, tenemos que

$$\langle w(s), s\rangle_{\mathbb{R}^d} \leq \langle w(s + th), s\rangle_{\mathbb{R}^d},$$

es decir,

$$\begin{aligned} 0 \quad &\leq \langle w(s + th), s\rangle_{\mathbb{R}^d} - \langle w(s), s\rangle_{\mathbb{R}^d}, \\ &\leq \langle w(s + th) - w(s), s\rangle_{\mathbb{R}^d}, \quad , \quad \forall\, h \in \mathbb{R}^d, \quad \text{y t pequeño.} \end{aligned}$$

Dividiendo por $t > 0$ y haciendo $t \downarrow 0$, tenemos que

$$0 \leq \langle \lim_{t\downarrow 0} \frac{w(s + th) - w(s)}{t}\,,\, s\rangle_{\mathbb{R}^d}.$$

Por otro lado, tomando $t < 0$ y haciendo $t \uparrow 0$ tal que $s + th \in \mathbb{R}_+^d$ para todo $t$, al dividir por $t$ invertimos la desigualdad, por lo tanto

$$0 \geq \langle \lim_{t\uparrow 0} \frac{w(s + th) - w(s)}{t}\,,\, s\rangle_{\mathbb{R}^d}.$$

Recordando que $w(s) = \nabla\mathcal{T}(s)$, y recordando el Teorema 6.3, $w(s)$ es diferenciable (en particular) en el sentido de Gateaux, entonces

$$\lim_{t\downarrow 0} \frac{w(s + th) - w(s)}{t} = \lim_{t\uparrow 0} \frac{w(s + th) - w(s)}{t} = \nabla w(s)^T h = \nabla^2\mathcal{T}(s)^T h, \quad \forall h \in \mathbb{R}^d.$$

Al juntar ambas desigualdades, encontramos una condición necesaria para el cálculo de

$\nabla^2\mathcal{T}(s)$,

$$\langle \nabla^2\mathcal{T}(s)^T h, s\rangle_{\mathbb{R}^d} = 0, \quad \forall\, h \in \mathbb{R}^d,$$

es decir,

$$h^T\nabla^2\mathcal{T}(s)s = 0, \quad \forall\, h \in \mathbb{R}^d \quad \Longrightarrow \quad \nabla^2\mathcal{T}(s)s = \vec{0} \in \mathbb{R}^d.$$

## 6.5. Implementación

### 6.5.1. Generación de datos

La metodología de implementación replica el experimento realizado en Delplancke y col., 2020. Utilizando una grilla fina del campo de velocidades del modelo Marmousi en 2 dimensiones de Versteeg, 1994, que representa un espacio bidimensional de $10000 \times 3000$ metros, se sitúan 48 puntos equiespaciados que representarán los sensores que puedan registrar la base de datos de tiempos de sismos al interior del campo. Se entenderá al *dominio* $G$, como el interior del campo de lentitud original del modelo.

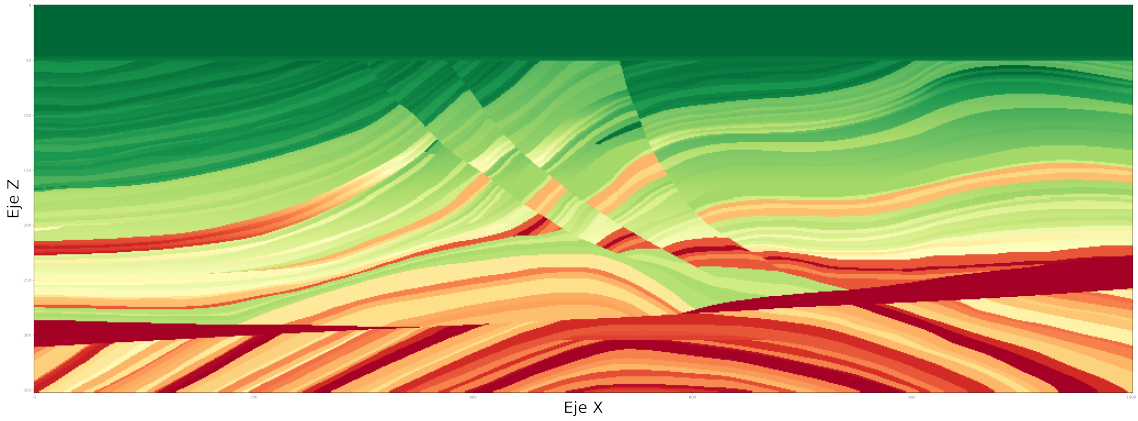


Figura 6.1: Campo de velocidades original del modelo Marmousi 2D, discretizado en una grilla de $1000 \times 30$ bloques.

Usando los mismos métodos numéricos que en Delplancke y col., 2020, se simulan $M = 10000$ sismos sintéticos cuyos hipocentros distribuyen uniformemente en el dominio y se asume que son registrados sólo en los primeros $N^i = 30$ sensores que captan la onda sísmica. De estos registros se formará la base de datos de tiempos $\{t^i_j : 1 \leq j \leq N^i, 1 \leq i \leq M\}$, ignorándose la posición del hipocentro y el tiempo inicial del inicio del sismo simulado. Por último, de esta base de datos se separará un subconjunto de 500 datos, denominado conjunto Test, que se ocupará para evaluar la función objetivo y observar el desempeño del algoritmo. El resto de la base de datos, es decir, todo el conjunto menos el conjunto Test, será ocupado para la extracción de los batchs en el algoritmo.

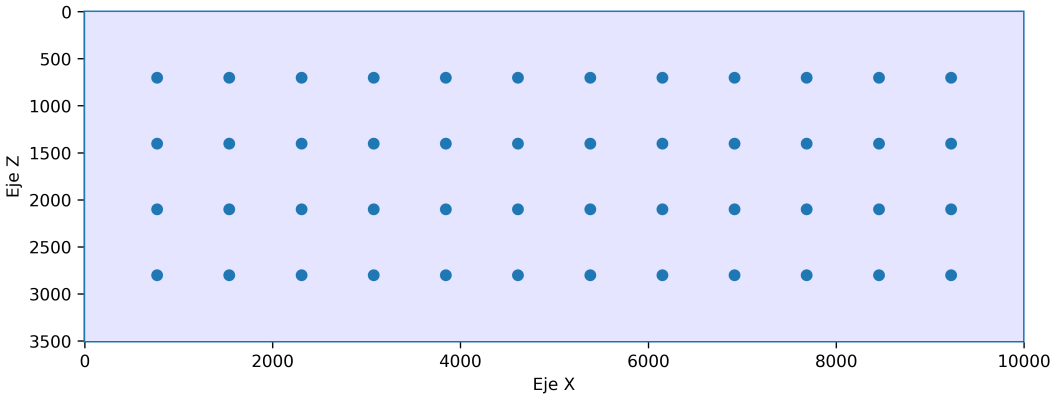


Figura 6.2: Distribución de sensores en el dominio, posición fija, estilo grilla.

### 6.5.2. Método de cascada

La resolución del problema de tomografía sísmica mediante los algoritmos de descenso de gradiente estocástico resulta ser muy sensible a la elección del punto inicial, para valores muy alejados y empeora cuando aumenta la dimensionalidad del parámetro a estimar (el campo de lentitud). Por esta razón se propone, en Delplancke y col., 2020, un método de refinamiento del campo de lentitud (método de cascada), que consiste en una serie de rutinas de optimización usando el descenso de gradiente tal que en cada nueva rutina se aumenta la dimensión de la parametrización del campo, otorgando control al algoritmo de minimización.

En cada rutina de gradiente estocástico (desde ahora *loop*), se utiliza el algoritmo de descenso con la cantidad de iteraciones necesarias para que el valor de la función objetivo alcance cierta estabilidad. Posteriormente, se propone una parametrización más fina, subdividiendo cada bloque en 4 sub-bloques de igual tamaño, se utiliza la última estimación del loop anterior como punto de inicio de una nueva rutina de descenso de gradiente estocástico con parámetros reajustados ad hoc a la nueva dimensión del campo.

Partiendo de un campo subdividido por una partición de $3 \times 8$ bloques se tiene una parametrización de 24 bloques de valores constantes. Por lo tanto, realizando la división señalada anteriormente, a través de 5 loops se llega a un total de 6144 bloques, es decir, una partición del campo de $48 \times 128$ bloques.

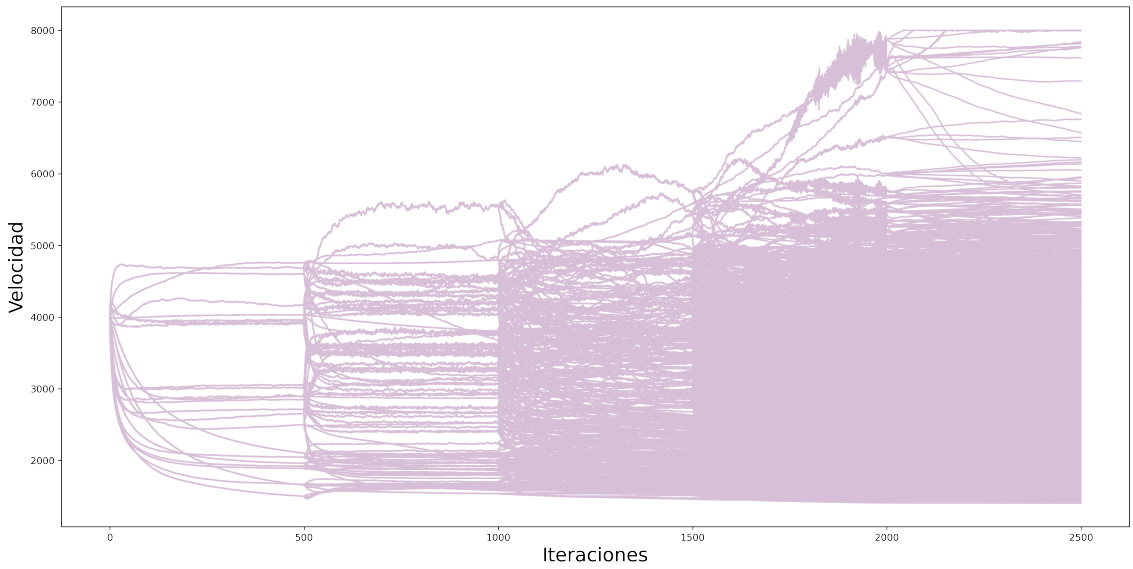


Figura 6.3: Esquema de evolución de las velocidades en cada iteración. Cada 500 iteraciones (cambio de loop, para 5 loops) cada valocidad se subdivide en 4 diferentes dada la refinación de los bloques.

### 6.5.3. Parámetros de la simulación

Cada loop del método de implementación amerita un reajuste de los parámetros del modelo. En el caso de la desviación estándar del modelo paramétrico de verosimilitud, $\sigma$, a través de los 5 loops va presentando el siguiente decrecimiento: 0.05, 0.03, 0.01, 0.0075, 0.006. Estos valores han sido ajustados manualmente bajo repeticiones de pruebas, donde se ha observado que, con varianzas mayores que estas, el algoritmo de descenso de gradiente no presenta disminución del valor objetivo del problema.

En cuanto al parámetro de temperatura, se establece una regla de aumento de la temperatura en cada iteración que pudiese seguir el incremento comentado en la Sección 4.6.1. De esta forma, el criterio de temperatura se rige por la siguiente fórmula,

$$\beta = 10^5 \times 4^{loop}\sqrt{(iter+1)\eta}, \tag{6.30}$$

donde *loop* es el número del loop actual (de 0 a 4), *iter* es el número de la iteración del algoritmo de gradiente que va desde 0 a 499 (un total de 500 iteraciones) y el parámetro $\eta$ corresponde al paso del algoritmo de descenso de gradiente.

Con respecto a este último parámetro, el paso del algoritmo SGD o SGLD, $\eta$, su naturaleza es constante en cada iteración, mas puede ser reajustado al cambiar de loop. El criterio de ajuste es tal que, en la primera iteración de cada loop, el campo de lentitud pueda presentar una variación máxima de 1/200 s/m en alguno de sus bloques. Posteriormente se discutirá el uso de un paso adaptativo variable en cada iteración.

El único parámetro que se decido mantener constante durante todo el experimento es el tamaño de batch para la estimación del gradiente de la función objetivo, cuyo valor corresponde a $m = 500$.

### 6.5.4. Iniciación del algoritmo

Para iniciar el algoritmo se propuso un punto inicial de 24 valores iguales en el campo de lentitud, es decir, para los $3 \times 8$ bloques iniciales del primer loop, el algoritmo inicia con el campo de valor 1/4000 s/m en cada uno de los bloques.

## 6.6. Resultados

Los dos aspectos importantes a considerar al momento de presentar los resultados son: primero el valor de la log verosimilitud del modelo, puesto que el principal fin del algoritmo es maximizar la verosimilitud. Dado que en este caso disponemos del campo de velocidades real del modelo de donde se extrajo la base de datos, el segundo aspecto a considerar es la semejanza del campo estimado con respecto al real, sin embargo, en este caso no se cuantificará esta "semejanza" y más delante se explicará por qué esta decisión y qué tipo de cuantificación es posible realizar.

Lo primero que podemos observar, en la Figura 6.4, es lo que a simple vista se nota sobre el campo estimado por el algoritmo SGLD y el estimado por el algoritmo SGD en Delplancke y col., 2020, ambos comparados con respecto al campo real. En este caso, los detalles más finos se destacan de mejor manera en la estimación realizada con el algoritmo SGLD, reflejando la textura y tendencias del campo real, mientras que en la versión SGD, estos detalles se difuminan perdiendo presencia los valores más altos (rojos). Sin embargo, cabe destacar que tanto los batch como cantidad de loops en la estimación podrían no ser los mismos, por lo que no es posible generar una comparación "más justa" con el campo estimado en Delplancke y col., 2020.

En la Figura 6.5, se puede comparar una estimación del campo realizada por el algoritmo SGD y una estimación del campo realizada por el algoritmo SGLD. Esta vez la comparación es tal que, tanto los puntos iniciales de los algoritmos, los batches utilizados y la cantidad de loops e iteraciones son las mismas, reflejando tan sólo el efecto del ruido gaussiano en cada iteración. En la versión SGLD se vuelve a encontrar mejor detalle en el centro de la imagen, que en la versión SGD, a pesar de que ambos algoritmos tuvieron los mismos niveles de refinamiento. Por lo que aquellos bloques de mayor tamaño que se pueden encontrar en la versión SGD, impidiendo definir detalles más pequeños, sólo se pueden explicar debido a cercanía con algún óptimo local de la función objetivo que evitó que esos bloques siguieran variando.

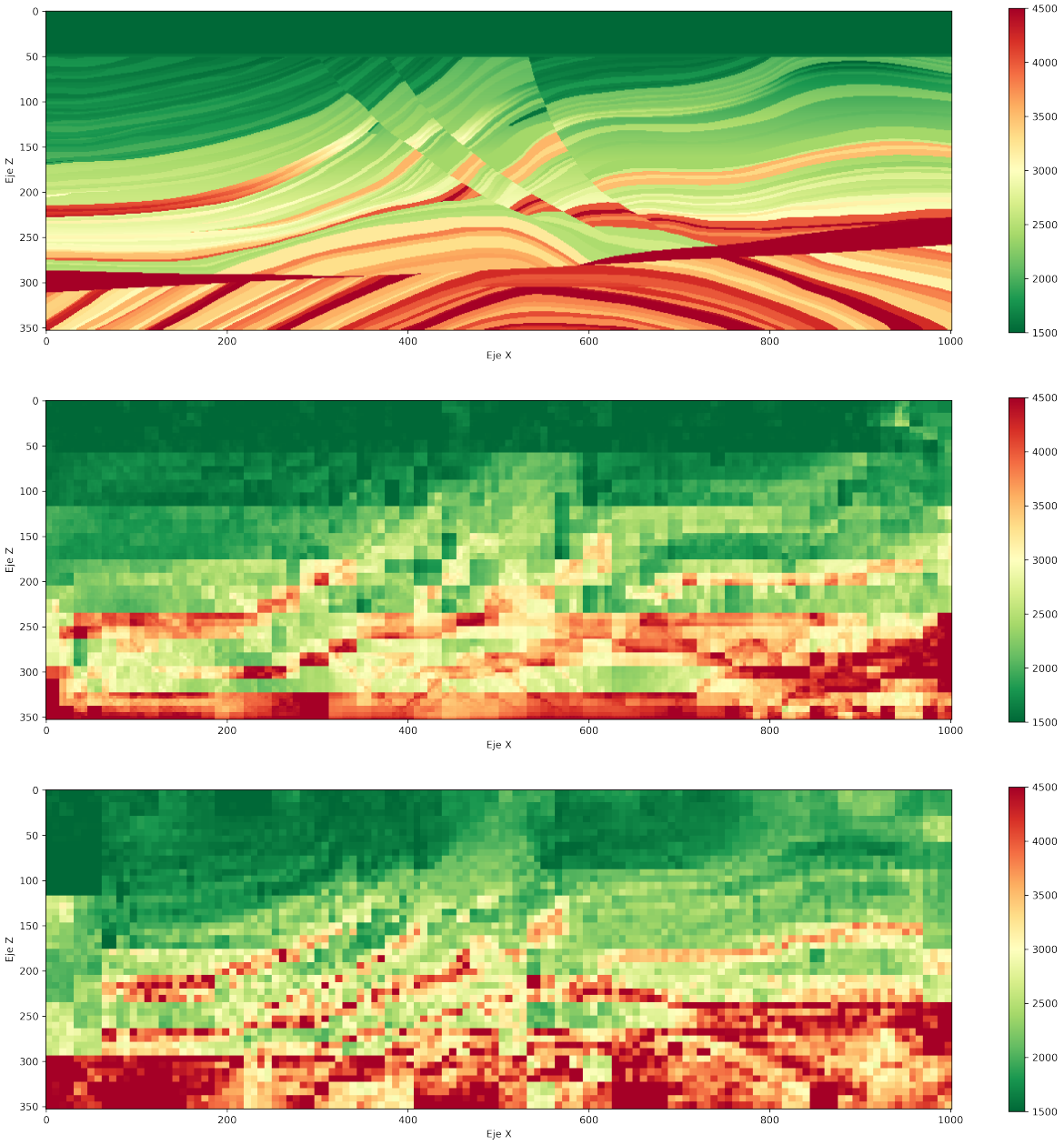


Figura 6.4: Desde arriba para abajo: En la primera imagen, la representación del campo original del modelo Marmousi. La segunda imagen corresponde a la estimación realizada por Delplancke y col., 2020 mediante el algoritmo SGD de 7 loops, con 100 iteraciones en cada loop. La tercera imagen es una estimación del campo realizada con el algoritmo SGLD, con 5 loops de 500 iteraciones en cada loop. Los tres campos con discretización de $1000 \times 300$ bloques.

En cuanto a los valores de la función objetivo del problema, se compara la evolución a medida que avanza cada iteración en cada loop, para las rutinas de optimización que resultan en las estimaciones de la Figura 6.5, para ambas versiones (SGD y SGLD). Los valores de la Log-verosimilitud negativa mantienen una tendencia de decrecimiento, sin embargo, como el algoritmo SGLD presenta una naturaleza aleatoria extra, este decrecimiento se ve perturbado por el ruido, como se puede observar en la Figura 6.6.

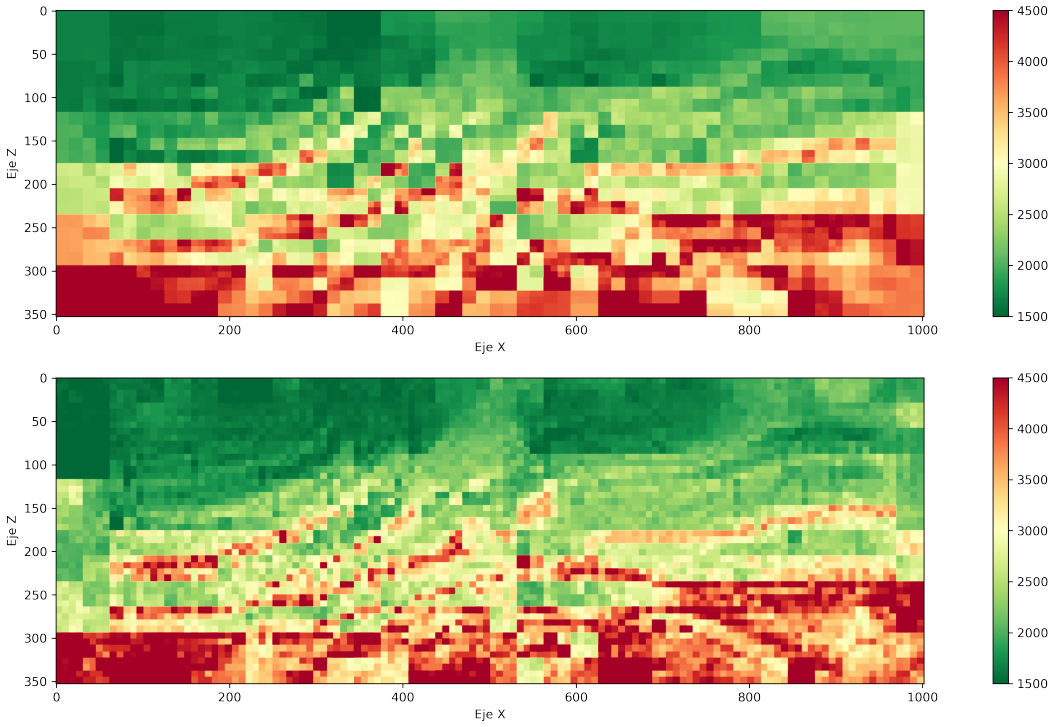


Figura 6.5: En la imagen superior, la estimación generada por el algoritmo SGD. En la imagen inferior, la estimación generada por el algoritmo SGLD. Ambas estimaciones del campo se realizaron con una rutina de 5 loops, con 500 iteraciones por loop. Ambas discretizaciones de $1000 \times 300$ bloques.

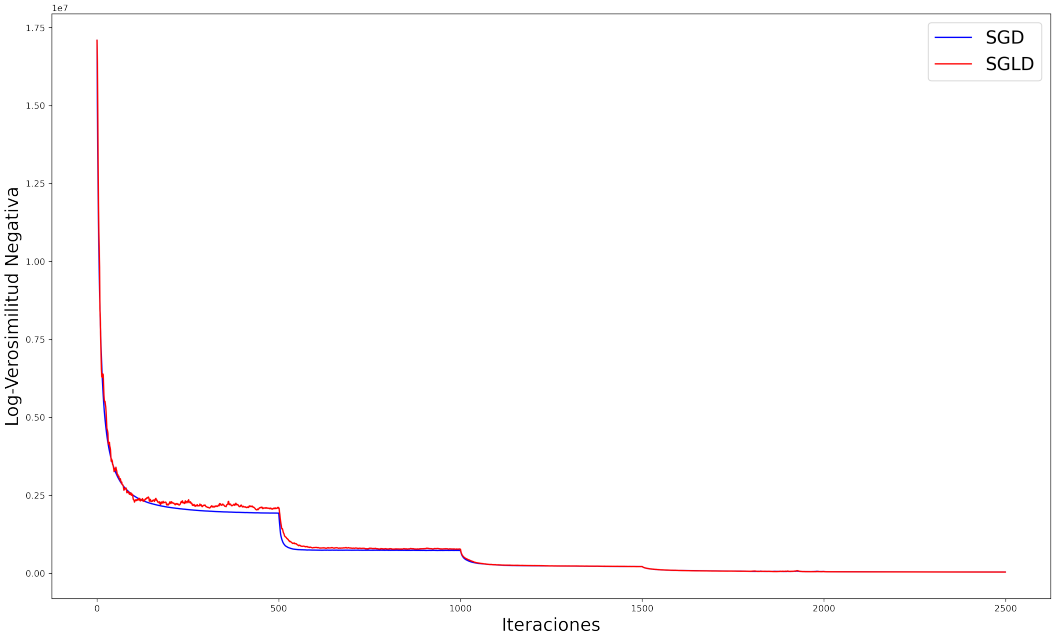


Figura 6.6: Decrecimiento de la función objetivo para los algoritmos SGD (en azul) y SGLD (en rojo), para 5 loops con 500 iteraciones en los primeros 4 loops, 700 en el último. Es posible notar al ruido generado por el algoritmo SGLD, particularmente al final del primer loop.

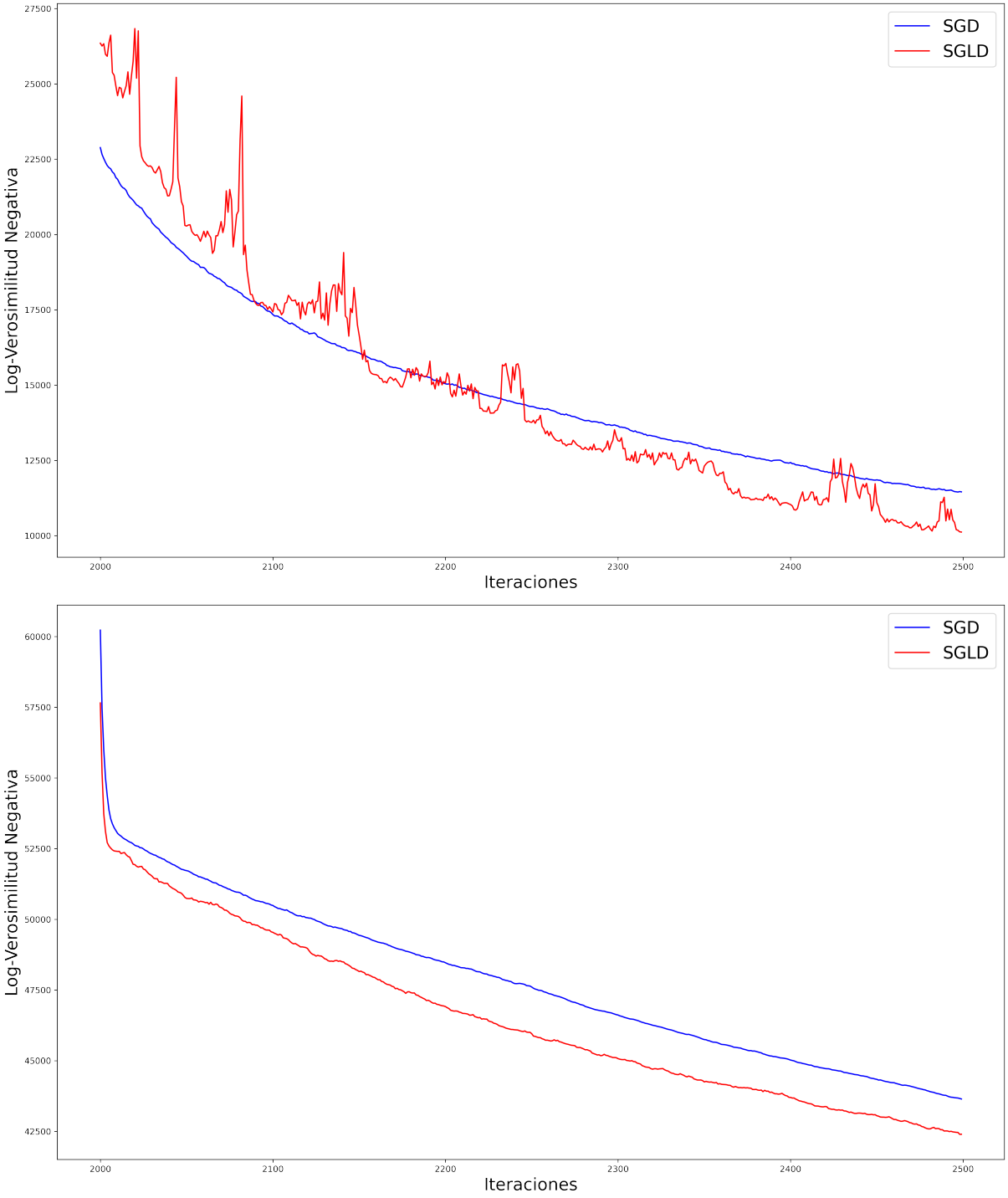


Figura 6.7: Dos casos del último loop (500 iteraciones) de la comparación de los algoritmos. Caso 1 (arriba): en algoritmo SGLD se mantiene oscilando al rededor del algoritmo SGD, manteniendo una tendencia de decrecimiento, pero en ocasiones, encontrando valores mayores al del SGD y, en ocasiones, menores. Caso 2 (abajo): El algoritmo SGLD logra separarse del algoritmo SGD al cambiar de loop encontrando valores siempre menores que el algoritmos SGD. Ambos algoritmos mantienen la tendencia de decrecimiento.

La diferencia entre los valores de la función objetivo al inicio del algoritmo con los valores obtenidos al final del algoritmo es tan grande que no se logra visualizar diferencia alguna en las últimas iteraciones. Sin embargo, al hacerle un acercamiento al gráfico en los últimos dos loops, podemos encontrar comportamientos como los mostrados en la Figura 6.7. En el primer caso, el ruido exógeno del algoritmo SGLD provoca que los valores oscilen en torno al valor de su contraparte SGD. Para este caso el algoritmo SGLD podría (o no) detenerse en iteraciones en que el valor objetivo sea menor que el del algoritmo SGD, dando argumentos para decir que la modificación ruidosa favorece el descenso y la búsqueda del mínimo (recalcar que la conclusión podría ser la contraria en caso de detener el algoritmo en un valor mayor al del otro método). En el otro caso (segunda imagen de la Figura 6.7), en alguna de las iteraciones, el ruido puede provocar que el algoritmo salga de (o entre en) alguna vecindad en torno a un óptimo local creando una bifurcación entre los valores objetivos que genere una notoria diferencia entre los algoritmos. En cual quiera de los dos casos, uno bien podría generar un criterio de detención para garantizar la posible mejora de la función objetivo a través del algoritmo SGLD.

## 6.7. Discusión de resultados

### 6.7.1. Modelo de sensores, bordes y parámetros adecuados

Comparando visualmente el campo real y el generado por el algoritmo SGLD en la Figura 6.4, se puede notar que los sectores de más diferencias son los más cercanos a la frontera (los bordes del campo). Esto es explicado por la configuración del modelo de sensores en forma de grilla. La configuración de sensores en forma de grilla provoca que no se puedan registrar datos en los bordes (puesto que no hay sensores), por lo tanto, el algoritmo de descenso del gradiente no justifica la evolución de los bloques por los que no pasan sismos, menos aún la modificación ruidosa del algoritmo. En respuesta a esta irregularidad se propuso una configuración de sensores de manera aleatoria, donde cada posición del sensor es una variable independiente uniformemente distribuida en el dominio, tal como lo muestra en la Figura 6.8. Esta nueva configuración replica de manera más realista el contexto del problema real de tomografía sísmica e intenta reducir los espacios vacíos que puedan dejar bloques sin datos (aunque el problema de los bloques sin datos realmente está sujeto al tamaño de los bloques y la cantidad de datos).

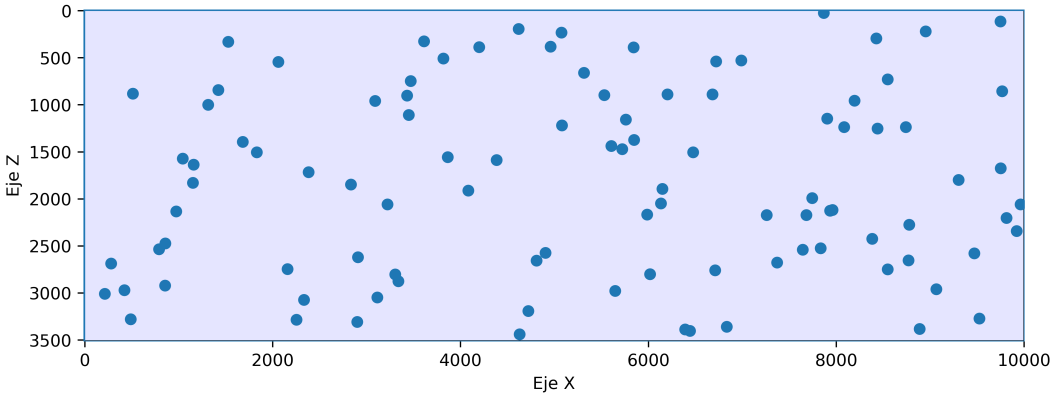


Figura 6.8: Distribución de sensores en el dominio, posición aleatoria.

Si bien, la configuración aleatoria de sensores mejora los errores en el borde de los campos estimados, no logra mejorar significativamente el rendimiento de los algoritmos SGD o SGLD bajo el criterio de la optimización de la verosimilitud, por lo que no se consideró como parte del análisis. Más aún, genera otro problema que dificulta la comparación de las simulaciones: el reajuste de parámetros.

En la sección anterior se dio a conocer los valores de las varianzas del modelo utilizadas en cada loop, $\sigma$, mencionando que se ha observado un comportamiento errático cuando estos valores son demasiado altos. En efecto, este parámetro se encuentra explícito en el cálculo del gradiente de la función objetivo, en la Ecuación (6.16), por lo que su definición debe tomarse con cuidado y dependerá tanto de la varianza original con la que se simularon los datos, como del tamaño de los bloques de la discretización del campo. Sin embargo, se ha podido reconocer (mas no probar) que este parámetro es sensible a la aleatoriedad del modelo y del algoritmo. Por ejemplo, una modificación del algoritmo SGD que se propuso, fue generar un paso $\gamma$ adaptado y dependiente de la condición inicial del campo. Esta condición inicial es de carácter aleatoria desde el segundo loop, por lo que el parámetro $\sigma$ ha de necesitar un valor menor en el segundo loop al valor necesario en el segundo loop si el paso se hubiese considerado constante. El no reajuste de las varianzas del modelo puede provocar valores absurdos en la función objetivo, incluso la divergencia del algoritmo. Caso similar ocurre con la configuración aleatoria de sensores.

Para finalizar esta discusión, al inicio de la sección de Resultados se mencionó formas de cuantificar la semejanza entre los campos. Estas formas corresponden a las normas $L^1$ y $L^2$ de las matrices, es decir; las sumas de los módulos o la suma cuadrática, de la diferencia entre los campos. El modelo de sensores aleatorio influía directamente en estos tipos de errores, pero no en verosimilitud, que corresponde al indicador principal de estos experimentos, por lo que tal análisis no se agregó a este trabajo.

### 6.7.2. Ajuste del parámetro de temperatura $\beta$

En la ecuación (6.30) se presentó el criterio de temperatura a utilizar en cada paso del algoritmo. La presencia de la raíz cuadrada está justificada por el intento de replicar la relación entre la temperatura y el tiempo de simulación del proceso de difusión expresado en la Sección 4.6.1, sin embargo, esto no es suficiente para lograr buenas estimaciones. El buen ajuste del parámetro de temperatura implica un buen control del ruido gaussiano. Es importante que este ruido no sea demasiado grande, para no desviar al algoritmo de la dirección de decrecimiento de la función objetivo, ni demasiado pequeño, para que exista influencia aleatoria que lo diferencie del algoritmo de descenso de gradiente estocástico usual SGD.

En cada refinamiento de bloques se busca aumentar detalles de la estimación del campo, esto provoca que la estimación se vuelva más sensible al ruido a través de los loops. El término $4^{loop}$ en (6.30), cumple la función de disminuir el ruido a medida que la discretización del campo se hace más fina, precisamente mantiene la varianza del ruido gaussiano por unidad de área en toda la rutina de estimación. Por último, el valor $10^5$ en (6.30) se buscó de manera heurística y representa el valor base para la temperatura inicial que permita reflejar variaciones sin desviar completamente la dirección de descenso. Sin embargo, ¿qué tanto ruido puedo considerar sin arriesgar la optimización de la función objetivo? o ¿cuál es la varianza óptima para el ruido exógeno en el algoritmo SGLD?

Para intentar responder estas preguntas se tomó el caso de las simulaciones correspondientes a las Figuras 6.5 y 6.6. Ambas simulaciones sólo se diferencian en el ruido exógeno,

poseen el mismo punto inicial y, en cada iteración, se mueven en la dirección del gradiente calculado con los mismos batch de datos. Estos casos se compararán con tres versiones levemente modificadas de la versión SGLD. Recordemos que el algoritmo SGLD tiene un ruido correspondiente a $\sqrt{2\eta\beta^{-1}}\xi$, donde $\eta$ es el paso, $\beta$ es la temperatura definida en (6.30) y $\xi$ es un vector normal estándar. Las nuevas tres simulaciones tendrán los mismos batch en cada iteración, pero tendrán un ruido de la forma $\alpha\sqrt{2\eta\beta^{-1}}\xi$, con $\alpha$ tomando los valores 2, 5 y 10 (es decir; 2, 5 y 10 veces la desviación estándar del algoritmo SGLD original).

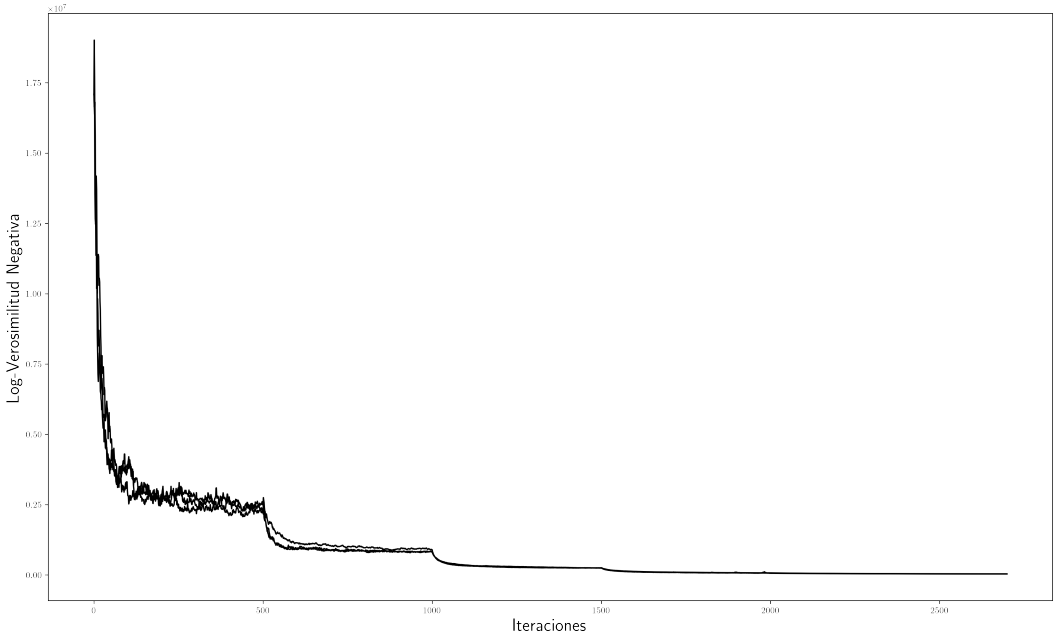


Figura 6.9: Valor de la función de Log-Verosimilitud para tres simulaciones del caso $\alpha = 2$.

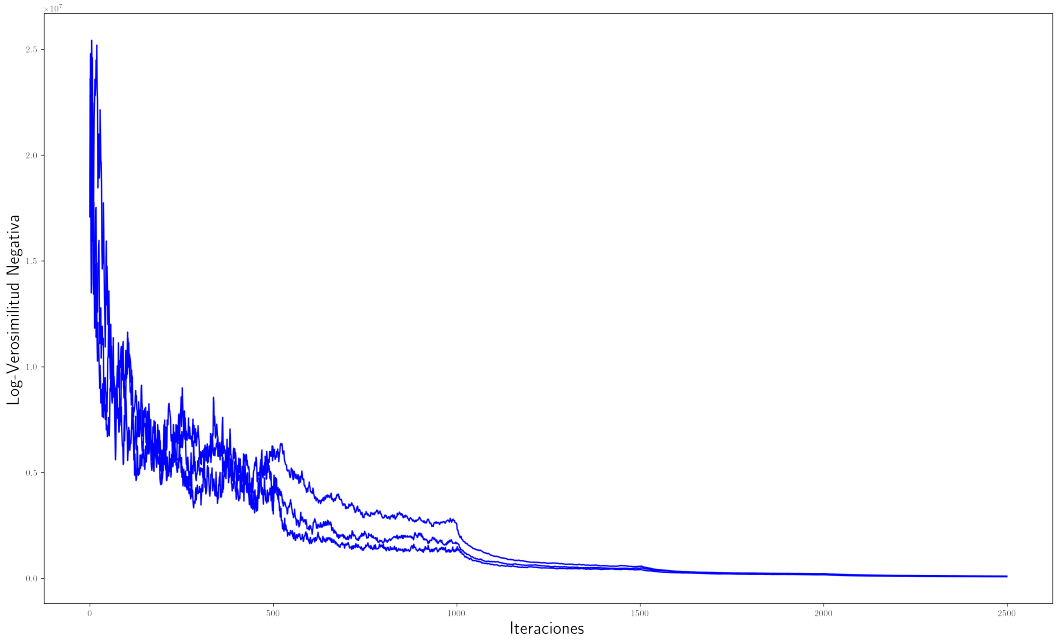


Figura 6.10: Valor de la función de Log-Verosimilitud para tres simulaciones del caso $\alpha = 5$.

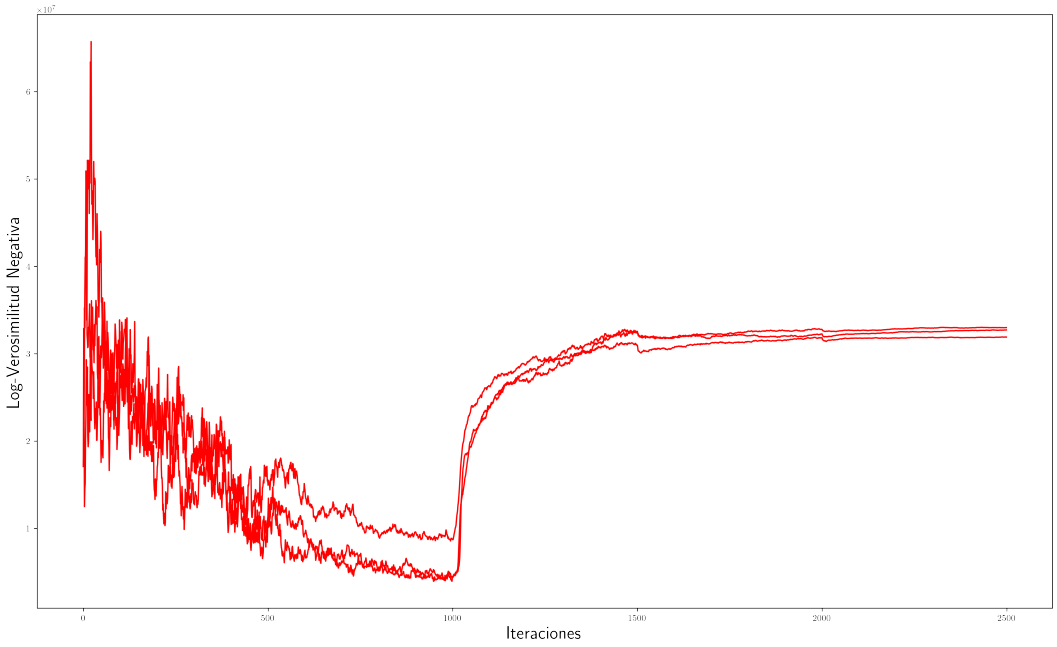


Figura 6.11: Valor de la función de Log-Verosimilitud para tres simulaciones del caso $\alpha = 10$.

De cada simulación (es decir, para cada $\alpha$), se realizaron 3 muestras de cada rutina completa (5 loops), cuya diferencia radica sólo en el conjunto de normales estándar que se utilizaron en cada iteración, independientes una de otra. El punto más distinguible en las Figuras 6.9, 6.10 y 6.11 es el ruido del primer loop, mucho menor al visto en las simulaciones anteriores (Figura 6.6). Para el caso $\alpha = 10$, a partir del tercer loop, notamos que el descenso de la fun-

ción objetivo se pierde. Más aún, la estimación tiende a empeorar en optimalidad a medida que avanza el algoritmo desde ese punto (ver Figura 6.11).

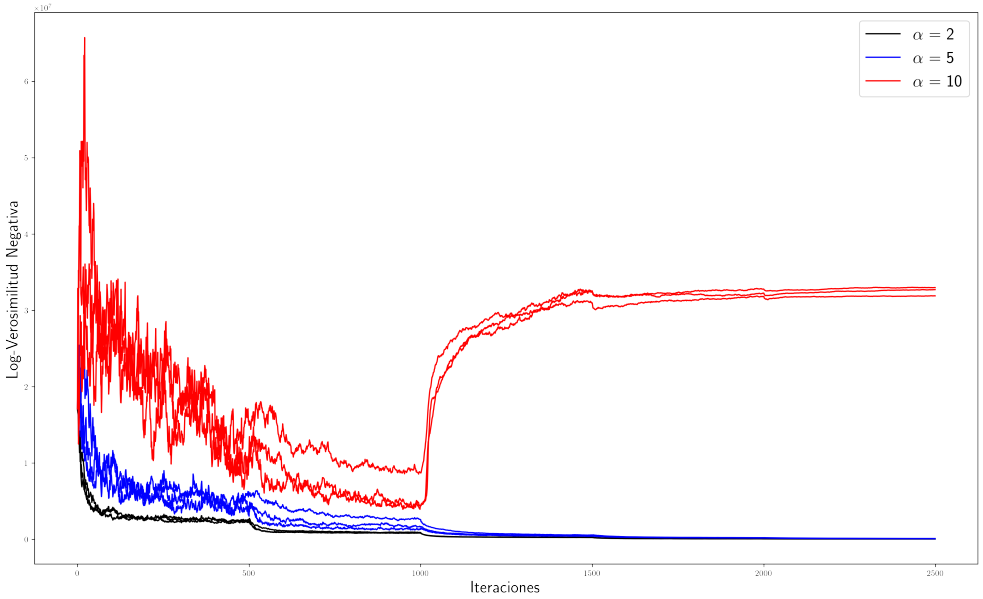


Figura 6.12: Comparación de las tres simulaciones por cada uno de los tres casos de aumento de ruido en el algoritmo SGLD.

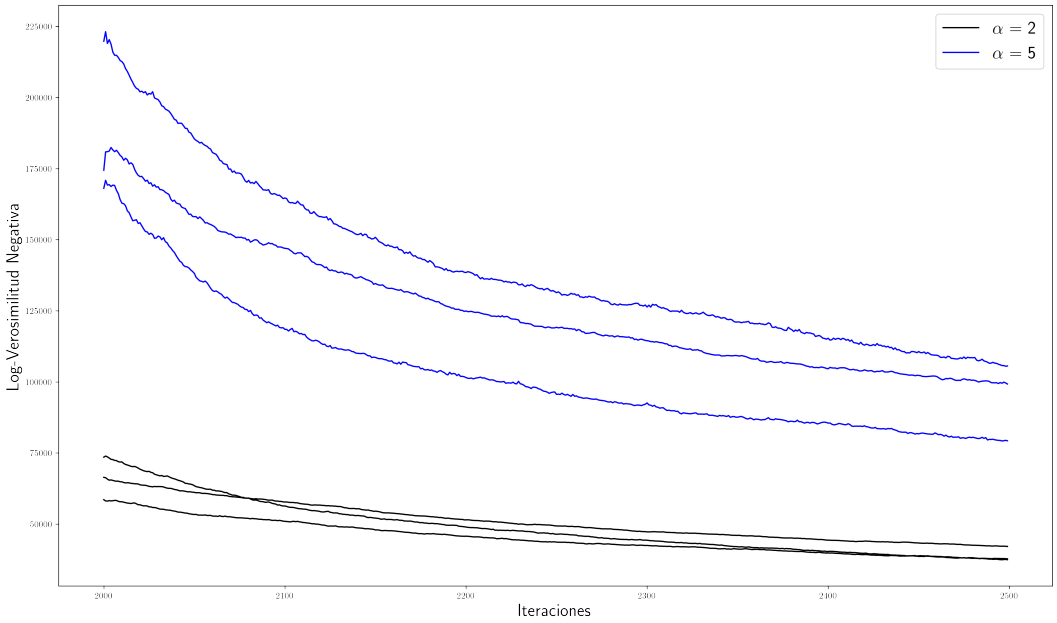


Figura 6.13: Comparación del último loop para los casos $\alpha = 2$ y $\alpha = 5$. El caso $\alpha = 2$ mantiene valores menores de la función objetivo en todo momento.

En la Figura 6.12 se pueden observar las 9 simulaciones, para distinto valores de $\alpha$. Cada una de ellas son fácilmente distinguibles en los dos primeros loops, pero a partir del tercero el caso $\alpha = 10$ marca una diferencia notable ante los otros dos casos que continúan su decrecimiento. Este cambio en el comportamiento de la versión $\alpha = 10$ puede deberse a

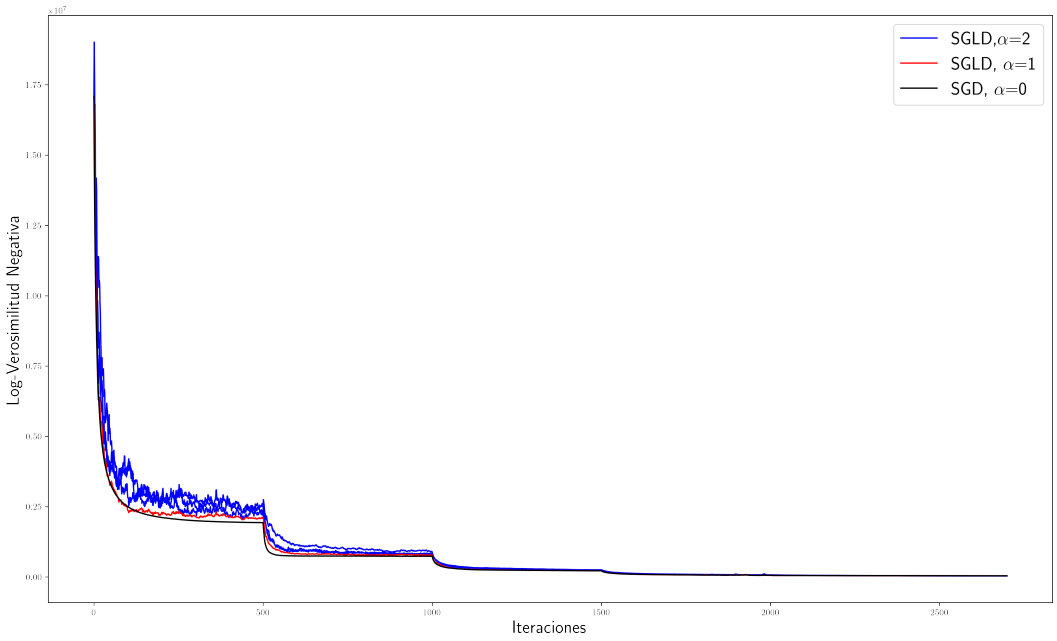


Figura 6.14: Comparación de las simulaciones para el caso $\alpha = 2$ con los algoritmos SGD y SGLD usuales, correspondientes a la Figura 6.6.

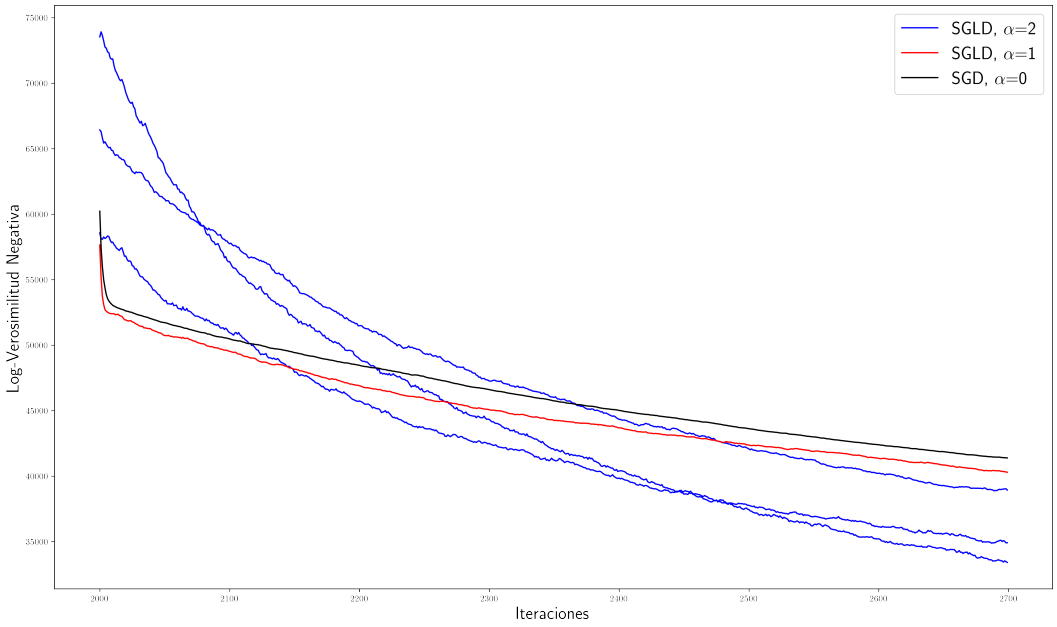


Figura 6.15: Comparación del último loop de las simulaciones para el caso $\alpha = 2$ con los algoritmos SGD y SGLD usuales, correspondientes a la Figura 6.6. Se puede observar cómo las tres simulaciones del caso $\alpha = 2$ logran llegar a valores menores que los dos algoritmos usuales, incluso manteniendo decrecimientos más pronunciados durante todo el loop.

que el ruido a partir del tercer loop es lo suficientemente grande como para inestabilizar el algoritmo cuando se encuentra en ese nivel de discretización del campo, recordando que el campo es más sensible al ruido a medida que aumenta su dimensionalidad. El detalle de la diferencia entre el caso $\alpha = 2$ y $\alpha = 5$ en los últimos loops se puede observar en la Figura 6.13,

en que el algoritmo con menor ruido logra siempre un menor valor de la función objetivo.

Hasta este punto, la función objetivo del problema ha tenido valores peores a medida que se aumenta el ruido en el algoritmo SGLD. Sin embargo, al momento de comparar los métodos usuales con la modificación $\alpha = 2$ se logra observar que un aumento leve de ruido logra disminuir la función objetivo aún más de lo que ocurría con el algoritmo SGD y el SGLD usual (Ver Figuras 6.14 y 6.15). El hecho de que fuesen, no solo una, sino que las 3 muestras independientes del algoritmo con la modificación $\alpha = 2$, da la impresión de que la mejora encontrada realmente sea provocada por el aumento del ruido (puntualmente por el factor 2, en este caso). Dado que se encontró un reajuste del ruido que favorece al algoritmo, quedará pendiente la pregunta de cuál es el ponderador óptimo en esta aplicación del algoritmo SGLD y qué relación pueda tener con el contexto del problema.

### 6.7.3. Efecto y uso adecuado de la aleatoriedad

De acuerdo a lo observado en la Sección 6.6 y en particular, en la Figura 6.7, el comportamiento del valor objetivo de la función, bajo las condiciones específicas de paso $\gamma$ y temperatura $\beta$, puede ser de las siguientes formas: oscilación en torno al descenso sin alteraciones y oscilación cuya media difiere de los valores tomados por el algoritmo SGD (en este último caso la media puede ser mayor o menor que el entregado por el algoritmo SGD). Sin embargo, una observación muy importante que no se debe dejar pasar es que en ambos casos la tendencia del algoritmo siempre es en la dirección de descenso (a pesar de existir momentos en que el valor objetivo aumente, más recurrentemente visto en el algoritmo SGLD). Esto permite idear un plan de acción al momento de usar el algoritmo SGLD para el problema de tomografía, con los siguientes pasos: en caso de que el algoritmo SGLD presente movimientos ruidosos al rededor del valor de decrecimiento, mantener la tendencia y aumentar la cantidad de iteraciones, puesto que eso hará disminuir el ruido de acuerdo a la definición de temperatura en función de la cantidad de iteraciones propuesta en (6.30). De esta forma el valor del algoritmo SGLD será cercano al del algoritmo SGLD. También existe la opción de disiminuir de forma más rápidamente el ruido cuando la oscilación se encuentre en un valor favorable para el algoritmo SGLD, con tal de forzar un mejor rendimiento.

En el caso de que el ruido provoque un salto de la función objetivo que se diferencie considerablemente del valor tomado por el algoritmo SGD, tenemos el sub caso en que esta diferencia es favorable para el algoritmo SGD (es decir, el algoritmo SGD tome valores más bajos). Esto no implica que el algoritmo SGLD no pueda optimizar la función hasta llevar a los valores que pueda tomar el algoritmo SGD, puesto que el valor sigue en decrecimiento, simplemente que demorará más iteraciones en disminuir al rango de valores del algoritmo SGD.

El último caso es el descrito por la segunda imagen de la Figura 6.7, en que el salto favorece al algoritmo SGLD, ahorrando iteraciones y llevando el valor objetivo a valores siempre menores que los encontrados por el algoritmo SGD. Un indicador de este caso podría ser una cantidad importante de iteraciones en que el algoritmo SGLD mantenga la ventaja, con poca variabilidad en la distancia entre los valores objetivos de ambos algoritmos.

Para finalizar, es necesario comentar que el comportamiento de decrecimiento de ambos algoritmos fue propio del contexto del problema (cantidad de bloques, función del problema, etc) y se recomienda su comprensión como método de ajuste para este caso particular más que una heurística de uso. Si bien, el mejor caso encontrado fue en aquel que el salto aleatorio separó la función objetivo del algoritmo SGLD con el del algoritmo SGD, favoreciendo al primero de ellos, nunca se logró verificar que tal salto efectivamente representaba posicionarse

en la vecindad de otro punto crítico (lo que realmente se buscaba), puesto que los valores no alcanzaban a estabilizarse. Por lo tanto, para otros problemas, no es seguro que el algoritmo SGLD vaya a tener un rendimiento como el encontrado en estos experimentos, puesto que el riesgo de caer en un punto crítico peor existe. No obstante, la naturaleza del algoritmo SGLD siempre le da la probabilidad de esquivar o salir de tales puntos, a diferencia del algoritmo SGD.

# Capítulo 7

# Conclusiones

En el marco del análisis del algoritmo de descenso de gradiente estocástico de Langevine, se encuentra una cota para el sesgo respecto al valor óptimo global de la función objetivo del problema de riesgo empírico. A diferencia de Raginsky y col., 2017, que también intenta ejemplificar este error, la cota encontrada es explícita y clara en su dependencia hacia todos los parámetros y variables del problema. Del desglose detallado de las dependencias del sesgo es posible afirmar que la metodología de análisis del algoritmo utilizada en este trabajo no logra concluir la existencia de convergencia entre el valor esperado de la función objetivo evaluada en el algoritmo y el valor óptimo de la función, sino que genera una interacción de *trade off*, como bien se explica en la Sección 4.6.1.

Las condiciones para el resultado general del algoritmo SGLD resultan un punto importante para la aplicación del Teorema 4.4, sin embargo, esto no restringe el uso del algoritmo. Los resultados de proximidad de la Sección 4 son descritos a través del valor esperado de los errores con respecto al valor óptimo de la función, esto quiere decir que la existencia conjuntos de medida nula (en las respectivas leyes de los procesos involucrados) en la que las condiciones no se cumplan, no afectará los resultados encontrados. Por lo tanto, todas las condiciones se pueden reemplazar por sus versiones c.t.p., como lo visto para la condición de suavidad (Condición 1) en el caso de la tomografía (Sección 6.4.2).

En la Sección 4.6.1 se habla sobre un ajuste que permite controlar el error entre valor de la función objetivo obtenida con el algoritmo SGLD y el óptimo de esta función, en la que el rol de los parámetros usuales de los algoritmos del tipo SGD se sigue cumpliendo, esto es: el paso del algoritmo, $\eta$, siempre debe tender a cero y el número final de iteraciones $K$, siempre será favorable de aumentar. Mientras que el nuevo parámetro, el término de temperatura $\beta$, cuya función es controlar la varianza del ruido gaussiano, debe ser lo suficientemente grande, del orden de una potencia del "tiempo de simulación" del algoritmo $\eta \times \textit{iteración}$, para que el ruido sea lo suficientemente pequeño en cada paso. Sin embargo, aunque esta elección de parámetros de control sobre el error, la cantidad de ruido inicial parece ser importante en la práctica. En la Sección 6.7.2 se menciona que al parámetro de temperatura se le pondera un elevado valor con el propósito de partir con ruido suficientemente pequeño desde la primera iteración porque, de lo contrario, el gradiente no logra influir lo suficiente para mostrar un decrecimiento de la función objetivo. No se muestra de forma implícita la forma de este valor ponderado en este trabajo, pero se sospecha que dependerá de la función objetivo y de cómo varía solamente por efecto del ruido.

En el marco del problema de tomografía sísmica, se ocupa la teoría formulada en Delplancke y col., 2023 para construir el algoritmo SGLD. Se recurre a técnicas de cálculo y del

resultados de Delplancke y col., 2023 y Alberti y Ambrosio, 1999, para probar la existencia del diferencial de segundo orden para la función de tiempo mínimo de viaje entre dos puntos $\mathcal{T}(s)$, además de una caracterización necesaria para $D_s^2\mathcal{T}$.

Entre los resultados numéricos para resolución del problema de tomografía se muestra que el algoritmo SGLD logra recuperar mejores detalles que el campo estimado en Delplancke y col., 2020. Sin embargo, las funciones objetivos de estas estimaciones no son comparables por la elección de parámetros que influyen en estos valores. Por otro lado, fijando los parámetros de ambos algoritmos, podemos notar que el mismo efecto se puede encontrar en una comparación "más justa" entre el algoritmo SGD y SGLD, como la mostrada en la Figura 6.5, en que el algoritmo SGD presenta bloques de mayor tamaño impidiendo definir formas más detalladas, para el mismo rendimiento de función objetivo (Figura 6.6).

En cuanto a los comportamientos de la función objetivo notamos que, bajo una buena elección de temperatura, el algoritmo SGLD logra mejores rendimientos que el algoritmo SGD. Si ambos algoritmos se usan paralelamente, se muestra la heurística que le dé, al algoritmo SGLD, la opción de mejorar siempre el valor final de la función objetivo.

Se analizan los efectos del modelo de sensores y el parámetro de temperatura. En cuanto a este último, se logra observar que un leve aumento del ruido favorecía aún más la disminución del valor objetivo con respecto al algoritmo SGD. Al mismo tiempo, mayores niveles de ruido perjudicaban esta mejora, incluso llegaban a perder la dirección de decrecimiento en el loop más fino de discretización.

# Bibliografía


Alberti, G. S. & Ambrosio, L. (1999). A geometrical approach to monotone functions in $\mathbb{R}^n$. *Mathematische Zeitschrift*, *230*, 259-316.

Bakry, D., Gentil, I. & Ledoux, M. (2013). *Analysis and Geometry of Markov Diffusion Operators*. Springer International Publishing. https://books.google.cl/books?id=gU3ABAAAQBAJ

Barthe, F. & Strzelecki, M. (2021). Functional Inequalities for Two-Level Concentration. *Potential Analysis*, *56*(4), 669-696. https://doi.org/10.1007/s11118-021-09900-9

Billingsley, P., Collection, K. M. R., Shewhart, W., series in probability, W., mathematical statistics & Wilks, S. (1999). *Convergence of Probability Measures*. Wiley. https://onlinelibrary.wiley.com/doi/book/10.1002/9780470316962

Borkar, V. & Mitter, S. (1999). A Strong Approximation Theorem for Stochastic Recursive Algorithms. *Journal of Optimization Theory and Applications*, *100*, 499-513. https://doi.org/10.1023/A:1022630321574

Bottou, L. Online Algorithms and Stochastic Approximations (D. Saad, Ed.) [revised, may 2018]. En: En *Online Learning and Neural Networks* (D. Saad, Ed.). Ed. por Saad, D. revised, may 2018. Cambridge, UK: Cambridge University Press, 1998. https://leon.bottou.org/publications/pdf/online-1998.pdf

Cattiaux, P., Guillin, A. & Wu, L. (2008). A note on Talagrand's transportation inequality and logarithmic Sobolev inequality. https://doi.org/10.48550/ARXIV.0810.5435

Chen, T., Fox, E. B. & Guestrin, C. (2014). Stochastic Gradient Hamiltonian Monte Carlo. https://doi.org/10.48550/ARXIV.1402.4102

Chiang, T, Hwang, C. & Sheu, S.-J. (1987). Diffusion for Global Optimization in $\mathbb{R}^n$. *Siam Journal on Control and Optimization - SIAM.*

Delplancke, C., Fontbona, J., Prado, J. & Brault, A. (2023). A stochasticalgorithm for the inverse problem of first travel-time passive seismic tomography. (Artículo en preparación).

Delplancke, C., Fontbona, J. & Prado, J. (2020). A scalable online algorithm for passive seismic tomography in underground mines. *Geophysics*, *85*(4), WA201-WA211. https://doi.org/10.1190/geo2019-0440.1

Dembo, A. & Zeitouni, O. (2009). *Large Deviations Techniques and Applications*. Springer Berlin Heidelberg. https://books.google.cl/books?id=iT9JRlGPx5gC

Freidlin, M., Szücs, J. & Wentzell, A. (2012). *Random Perturbations of Dynamical Systems*. Springer. http://books.google.de/books?id=p8LFMILAiMEC

Ghadimi, S. & Lan, G. (2012). Optimal Stochastic Approximation Algorithms for Strongly Convex Stochastic Composite Optimization I: A Generic Algorithmic Framework. *SIAM Journal on Optimization*, *22*(4), 1469-1492. https://doi.org/10.1137/110848864

Ghadimi, S. & Lan, G. (2013). Stochastic First- and Zeroth-order Methods for Nonconvex Stochastic Programming. https://doi.org/10.48550/ARXIV.1309.5549

Gross, L. (1993). Logarithmic Sobolev inequalities and contractivity properties of semigroups. *Dirichlet Forms: Lectures given at the 1st Session of the Centro Internazionale Matematico Estivo (C.I.M.E.) held in Varenna, Italy, June 8–19, 1992* (pp. 54-88). Springer Berlin Heidelberg. https://doi.org/10.1007/BFb0074091

Held, L. & Bové, D. (2013). *Applied Statistical Inference: Likelihood and Bayes*. Springer Berlin Heidelberg. https://books.google.cl/books?id=Xv3FBAAAQBAJ

Hu, W., Li, C. J., Li, L. & Liu, J.-G. (2017). On the diffusion approximation of nonconvex stochastic gradient descent. https://doi.org/10.48550/ARXIV.1705.07562

Hwang, C.-R. (1980). Laplace's Method Revisited: Weak Convergence of Probability Measures. *The Annals of Probability*, *8*(6), 1177 -1182. https://doi.org/10.1214/aop/1176994579

Karatzas, I., Shreve, I., Shreve, S. & Shreve, S. (1991). *Brownian Motion and Stochastic Calculus*. Springer New York. https://books.google.cl/books?id=ATNy_Zg3PSsC

Lee, W., Lee, W., Lee, X., Stewart, S. & Stewart, S. (1981). *Principles and Applications of Microearthquake Networks*. Academic Press. https://books.google.cl/books?id=NjW_CVz7NSoC

Li, Q., Tai, C. & E, W. (2017). Stochastic Modified Equations and Adaptive Stochastic Gradient Algorithms. http://proceedings.mlr.press/v70/li17f/li17f.pdf

Ma, Y.-A., Chen, T. & Fox, E. (2015). A Complete Recipe for Stochastic Gradient MCMC.

Newey, W. K. & McFadden, D. Chapter 36 Large sample estimation and hypothesis testing. En: Vol. 4. Handbook of Econometrics. Elsevier, 1994, pp. 2111-2245. https://www.sciencedirect.com/science/article/pii/S1573441205800054

Noble, M., Gesret, A. & Belayouni, N. (2014). Accurate 3-D finite difference computation of traveltimes in strongly heterogeneous media. *Geophysical Journal International*, *199*. https://doi.org/10.1093/gji/ggu358

Nolet, G. (2008). *A breviary of seismic tomography : Imaging hte Interior of the Earth and Sun*. Cambridge University Press.

Nowack, R. & Li, C. (2009, enero). Seismic Tomography. https://doi.org/10.1007/978-0-387-30441-0_91

Raginsky, M., Rakhlin, A. & Telgarsky, M. (2017). Non-convex learning via Stochastic Gradient Langevin Dynamics: a nonasymptotic analysis. *CoRR*, *abs/1702.03849*. http://arxiv.org/abs/1702.03849

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*. Wiley. https://books.google.cl/books?id=lPhQAAAAMAAJ

San Martín, J. (2015). *Teoría de la Medida*. Editorial universitaria.

Simmons, G. (1983). *Introduction to Topology and Modern Analysis*. R.E. Krieger Publishing Company. https://books.google.cl/books?id=tEFAAQAAIAAJ

Tarantola, A. (2004). *Inverse Problem Theory and Methods for Model Parameter Estimation*. http://www.ipgp.fr/~tarantola/Files/Professional/SIAM/index.html

Tarantola, A. & Valette, B. (1982). Inverse problems = Quest for information. *Journal of geophysics*, *50*, 159-170.

Vapnik, V. Principles of Risk Minimization for Learning Theory. En: *Proceedings of the 4th International Conference on Neural Information Processing Systems*. NIPS'91. Denver, Colorado: Morgan Kaufmann Publishers Inc., 1991, 831–838. ISBN: 1558602224.

Vapnik, V. (2000). *The Nature of Statistical Learning Theory*. Springer: New York.

Vapnik, V. N. Estimation of Dependences Based on Empirical Data (in Russian). En: 1979.

Versteeg, R. (1994). The Marmousi experience: Velocity model determination on a synthetic complex data set. *The Leading Edge*, *13*(9), 927-936. https://doi.org/10.1190/1.1437051

Villani, C. (2003). *Topics in Optimal Transportation*. American Mathematical Society. https://books.google.cl/books?id=idyFAwAAQBAJ

Welling, M. & Teh, Y. Bayesian Learning via Stochastic Gradient Langevin Dynamics. En: 2011, enero, 681-688.

# Anexos:
# Demostraciones complementarias

## A. Demostración Teorema 6.1

Previo a la demostración del Teorema 6.1, los siguientes resultados y análisis están extraídos de Delplancke y col., 2023, que enmarcan el contexto de parametrizaciones de campos y son elementales en los resultados ocupados en este trabajo.

Recordando que $\mathcal{P}_{a,b}$ es el conjunto de caminos Lipschitz continuos entre los puntos $a$ y $b$ en $\mathbb{R}^3$, definidos en la Sección 6.1. Si $p \in \mathcal{P}_{a,b}$ y $\gamma(t)$ es su parametrización continua, entonces podemos definir el largo del camino $p$ como

$$Len(p) := \int_0^{t_\gamma} \|\gamma'\| dt,$$

donde $t_\gamma$ es el tiempo en que la curva pasa por primera vez el punto $b$. Además se cumple que $|\langle \Gamma_p, S\rangle| \leq \|S\|_\infty Len(p)$, donde $\langle \Gamma_p, S\rangle$ es la integral de camino definida en (6.1).

Partiremos enunciando algunas propiedades del tiempo mínimo de viaje entre dos puntos fijos, $a$ y $b$, cuando el tiempo se ve como una función de un campo escalar arbitrario $S \in \mathcal{B}(\mathbb{R}^3, \mathbb{R})$. Definimos

$$T : \mathcal{B}(\mathbb{R}^3, \mathbb{R}) \cap \{-\infty\}, \text{ por } T(S) := d_S(a, b)$$

cuyo dominio $Dom\, T$ contiene a $\mathcal{B}(\mathbb{R}^3, [0, \infty))$.

**Proposición 7.1** (Proposición 2.3 en Delplancke y col., 2023) *El mapeo $S \in \mathcal{B}(\mathbb{R}, \mathbb{R}) \mapsto T(S)$ es propio, cóncavo, semi-continuo superior y $IntDom\, T$ contiene a $\bigcup_{m>0} \mathcal{B}(\mathbb{R}^3, [m, \infty))$, el cono abierto de campos positivos y acotados lejos de 0. En particular, en cada $S \in \mathcal{B}(\mathbb{R}^3, [m, \infty))$ con $m > 0$, $T$ es continuo y su super-diferencial $\partial_S T$ es no vacío. Por último, si para $S \in Dom\, T$ existe $p \in \mathcal{P}_{a,b}$ tal que $T(S) = \int_p S$, entonces $\Gamma_p \in \partial_S T$.*

Demostración. (Proposición 7.1) Como ínfimo de una familia de funciones lineales y continuas, $T$ es cóncava y semi-continua superior. Claramente, si $S \in \mathcal{B}(\mathbb{R}^3, [m, \infty))$ entonces $T(S) \geq 0$. Para el resto de la demostración, consideramos $S \in \mathcal{B}(\mathbb{R}^3, [m, \infty))$ fijo con $m > 0$. Para todo $S' \in \mathcal{B})\mathbb{R}^3, \mathbb{R})$ tal que $\|S'(x) - S(x)\|_\infty \leq m/2$, tenemos que $S'(x) \geq S'(x) - S(x) + m \geq m/2$, esto es $S' \in \mathcal{B}(\mathbb{R}^3, [m', \infty))$ con $m' = m/2$. Tomando $(p_n)_{n\in\mathbb{N}} \subseteq \mathcal{P}_{a,b}$ una sucesión tal que $T(S) \leq \int_{p_n} S \leq T(S) + 2^{-n}$, tenemos que $T(S') - T(S) \leq \int_{p_n} S' - \int_{p_n} S + 2^{-n}$, por lo tanto $T(S') - T(S) \leq \|S' - S\| Len(p_n) + 2^{-n}$. Pero, de la cota $\int_{p_n} S \leq T(S) + 2^{-n}$ también tenemos que $Len(p_n) \leq m^{-1}(T(S) + 2^{-n})$. Por lo tanto, dado

que $d_S(a,b) \leq \|S\|_\infty \|a-b\|$, haciendo tender $n \to \infty$ encontramos que

$$T(S') - T(S) \leq \|S - S'\|_\infty m^{-1} \|S\|_\infty \|a-b\|. \tag{7.1}$$

Procediendo de forma similar, intercambiando los roles de $S$ y $S'$, también tenemos que

$$\begin{aligned} T(S) - T(S') &\leq \|S' - S\|_\infty (m')^{-1} \|S'\|_\infty \|a-b\| \\ &\leq \|S' - S\|_\infty (m')^{-1} \left(\|S\|_\infty + m/2\right) \|a-b\|. \end{aligned} \tag{7.2}$$

De esto se sigue que $|T(S') - T(S)| \leq C_m(S) \|S' - S\|_\infty$ para alguna constante $C_m(S) > 0$ dependiente de $m$ y de $S$. Finalmente, si para $S \in Dom\, T$ se cumpliera que $T(S) = \int_p S$ para algún camino $p \in \mathcal{P}_{a,b}$, entonces para cada $S'' \in Dom\, T$ tendríamos que

$$T(S'') - T(S) \leq \int_p S'' - \int_p S = \langle \Gamma_p, S'' - S \rangle.$$

□

Sea $J : \mathbb{R}^d \to \mathcal{B}(\mathbb{R}^3, \mathbb{R})$ un operador lineal inyectivo y continuo, interpretado como la parametrización lineal de algún subespacio de campos acotados. Denotamos por $J^* : \mathcal{B}(\mathbb{R}^3, \mathbb{R})^* \to (\mathbb{R}^d)^* = \mathbb{R}^d$, el operador adjunto de $J$, esto es, el mapeo lineal continuo que cumple que

$$\langle J^*\Gamma, r \rangle_{\mathbb{R}^d} = \langle \Gamma, Jr \rangle, \quad \forall\, \Gamma \in \mathcal{B}(\mathbb{R}^3, \mathbb{R})^*,\ r \in \mathbb{R}^d.$$

En particular, $J^* : \mathcal{B}(\mathbb{R}^3, \mathbb{R}) \to \mathbb{R}^d$ incrusta el conjunto de los caminos Lipschitz entre los puntos $a$ y $b$, $\{\Gamma_p : p \in \mathcal{P}_{a,b}\}$, en algún conjunto de "caminos finito-dimensionales representativos"

$$J^*\left(\{\Gamma_p : p \in \mathcal{P}_{a,b}\}\right) \subset \mathbb{R}^d.$$

Dado un campo de lentitud, $S$, acotado lejos de 0, siempre un "camino representativo minimizante" entre dos puntos en esta configuración de parametrizaciones finitas, independiente de la regularidad de $S$:

**Proposición 7.2** (Proposición 4.2 de Delplancke y col., 2023) *Sea $m > 0$ y $S = Js \in J(\mathbb{R}^d) \cap \mathcal{B}(\mathbb{R}^3, [m, \infty))$ para algún $s \in \mathbb{R}^d$. Entonces, existe $w \in Cl\left(J^*\left(\{\Gamma_p : p \in \mathcal{P}_{a,b}\}\right)\right) \subseteq \mathbb{R}^d$ tal que*[6]

$$d_S(a,b) = \langle w, s \rangle_{\mathbb{R}^d},$$

*donde $\langle \cdot, \cdot \rangle_{\mathbb{R}^d}$ denota el producto escalar en $\mathbb{R}^d$.*

Demostración. (Proposición 7.2) Por simplicidad escribimos $d_S(a,b) = T(S)$. Sea $(p_n)_{n \in \mathbb{N}}$ una sucesión en $\mathcal{P}_{a,b}$ tal que

$$T(S) \leq \langle \Gamma_{p_n}, S \rangle \leq T(S) - 2^{-n} \tag{7.3}$$

para todo $n \in \mathbb{N}$, y definimos los vectores $w_n := J^*\Gamma_{p_n} \in \mathbb{R}^d$. Entonces, tenemos que $\langle w_n, s \rangle_{\mathbb{R}^d} \to T(S)$ cuando $n \to \infty$, y deducimos que

$$\|w_n\| = \sup_{\|r\| \leq 1} |\langle \Gamma_{p_n}, Jr \rangle| \leq \sup_{\|r\| \leq 1} \|Jr\|_\infty Len(p_n) \leq \|J\|_{op} Len(p_n),$$

[6] La notación $Cl(A)$ indica la clausura del conjunto $A$.

donde $\|\cdot\|_{op}$ denota la norma del operador. Dado que $S \in \mathcal{B}(\mathbb{R}^3, [m, \infty))$ obtenemos de (7.3) que $Len(p) \leq m^{-1}(T(S) + 1)$ para todo $n \in \mathbb{N}$. Por lo tanto, $\|w_n\|$ es acotado y existe una subsucesión de $(w_n)_{n\in\mathbb{N}}$ que converge a un vector $w \in \mathbb{R}^d$ que cumple que $\langle w, s\rangle_{\mathbb{R}^d} = T(S)$. □

Desde ahora fijaremos los puntos $a$ y $b$ en $\mathbb{R}^3$, y la notación $T(S) := d_S(a, b)$. Denotamos por $\mathcal{T}$ a la función cóncava

$$\mathcal{T} := T \circ J : \mathbb{R}^d \to \mathbb{R} \cup \{-\infty\}$$

y escribimos $D_s\mathcal{T}(h)$ para la derivada direccional en $s \in Dom\mathcal{T}$ en dirección $h \in \mathbb{R}^d$, cuando exista.

Con lo aclarado anteriormente, podemos iniciar la demostración del Teorema 6.1.

Demostración. (Teorema 6.1) Sea $s \in \mathcal{S}_J$. Por la Proposición 7.2 existe una sucesión de caminos $(p_n)_{n\in\mathbb{N}}$ en $\mathcal{P}_{a,b}$ y $w \in \mathbb{R}^d$ tal que $J^*\Gamma_{p_n} \to w$ en $\mathbb{R}^d$ cuando $n \to \infty$, y $\langle w, s\rangle_{\mathbb{R}^d} = \mathcal{T}(s)$. En particular, para alguna sucesión $\varepsilon_n \to 0$ y cualquier $s' \in \mathcal{S}_J$ tenemos

$$\mathcal{T}(s') - \mathcal{T}(s) \leq \langle J^*\Gamma_{p_n}, s'\rangle - \langle J^*\Gamma_{p_n}.s\rangle + \varepsilon_n.$$

Tomando $n \to \infty$ tenemos que $\mathcal{T}(s') - \mathcal{T}(s) \leq \langle w, s' - s\rangle_{\mathbb{R}^d}$, esto es, $w \in \partial_s\mathcal{T}$ es supergradiente.

Ahora, como $\bigcup_{m>0} \mathcal{B}(\mathbb{R}^3, [m, \infty)) \subseteq IntDom\, T$, por lo visto en la Proposición 7.1, junto a la continuidad de $J : \mathbb{R}^d \to \mathcal{B}(\mathbb{R}^3, \mathbb{R})$, implica que $\mathcal{S}_J \subseteq IntDom\, \mathcal{T}$. En efecto, si $s \in \mathcal{S}_J$ con $Js \in \mathcal{B}(\mathbb{R}^3, [0, m))$ para algún $m > 0$, para todo $s' \in \mathcal{S}_J$ con $\|s - s'\| \leq m/(2\|J\|_{op})$ se tiene que

$$Js' \geq Js - \|Js' - Js\|_\infty \geq Js - \|J\|_{op}\|s - s'\| \geq m/2,$$

de donde se infiere que $s' \in Dom(\mathcal{T})$ para cada $s'$ de esa forma. En consecuencia, $\mathcal{T}$ es localmente Lipschitz en subconjuntos acotados de $\mathcal{S}_J$ y, por el teorema de Rademacher, es diferenciable c.t.p. en $\mathcal{S}_J$. En particular, para $s \in \mathcal{S}_J$-c.t.p. el conjunto $\partial_s\mathcal{T} = \{\nabla\mathcal{T}(s)\}$ es un singleton, con lo que existe un único $w = w(s) \in Cl\,(J^*(\{\Gamma_p : p \in \mathcal{P}_{a,b}\}))$ tal que $\mathcal{T}(s) = \langle w, s\rangle_{\mathbb{R}^d}$ y tenemos que $D_s\mathcal{T}(h) = \langle\nabla\mathcal{T}(s), h\rangle_{\mathbb{R}^d} = \langle w, h\rangle_{\mathbb{R}^d}$, para todo $h \in \mathbb{R}^d$. □

# B. Demostración Lema 6.1

Demostración. (Lema 6.1)

Por simplicidad se tomará $N = N^i$ durante la demostración. Sea $\{u_1, \cdots, u_{N-1}\}$ la base ortonormal de $\mathcal{H}$. Definimos $u_N := N^{-1/2}(1, \cdots, 1)^\top$. La matriz $P := [u_1, \cdots, u_{N-1}, u_N]$ es ortogonal y sus columnas forman una base ortonormal de vectores propios de $(I_N - \frac{1}{N}\mathbf{1}_N)$. Deducimos de (6.9) que

$$\Sigma^i = \sigma^2 P\,\mathrm{diag}(1, \cdots, 1, 0)P^\top = P\,\mathrm{diag}(\sigma^2, \cdots, \sigma^2, 0)P^\top. \tag{7.4}$$

Escribiendo ahora $\bar{P} := [u_1, \cdots, u_{N-1}]$, tenemos que $\bar{\epsilon}^i = \bar{P}^\top\epsilon^i$, es un vector Gaussiano centrado con matriz de covarianzas

$$\begin{aligned}\bar{\Sigma}^i &= \bar{P}^\top\Sigma^i\bar{P}\\ &= \sigma^2 I_{N-1},\end{aligned}$$

donde la última igualdad sale de (7.4). El primer lado de la expresión de la densidad de $\bar{\epsilon}^i$ es

inmediata. El segundo lado surge de que $\|\bar{\epsilon}^i\|^2 = (\epsilon^i)^\top \bar{P}\bar{P}^\top \epsilon^i$, notando que $\bar{P}\bar{P}^\top$ es la matriz de proyección ortogonal $(I_N - \frac{1}{N}\mathbf{1}_N)$ y por lo tanto $\bar{P}\bar{P}^\top \epsilon^i = \epsilon^i$ ya que $\epsilon^i \in \mathcal{H}$. □

# C. Demostración Teorema 6.2

Previo a presentar la demostración del Teorema 6.2, extraída de Delplancke y col., 2023, es necesario enunciar el Lema 2.1 del mismo artículo:

**Lema 7.1** (Lema 2.1 de Delplancke y col., 2023) *Asumiendo que $S \in \mathcal{B}(\mathbb{R}^3, [m, \infty))$ con $m > 0$. Entonces $d_S : (\mathbb{R}^3)^2 \to [0, \infty)$ define una distancia equivalente a la métrica Euclideana.*

A continuación, procederemos con la demostración del Teorema 6.2, extraída de Delplancke y col., 2023.

Demostración. (Teorema 6.2)

Por el Teorema 6.1, para cada conjunto dado $\{r^i, r^i_j, j = 1, \cdots, N^i\} \subset \mathbb{R}^3$, $s \mapsto \mathbf{F}(r^i, s) = F(r^i, Js)$ (donde $Js$ es el campo parametrizado visto como elemento de $\mathcal{B}(\mathbb{R}^3, (0, \infty))$, como se presenta en el Anexo 7) es diferenciable en $\mathcal{S}_J$-c.t.p., con cierta derivada dado por el Teorema. Se sigue de (6.11) y del teorema de Fubini que para todo $s$ en un conjunto medible $Q \subseteq \mathcal{S}_J$, con medida de Lebesgue completa, se cumple que

$$\nabla_s p(\mathbf{t}^i | a^i, s) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{N^i - 1} \exp\left(-\frac{\|\mathbf{t}^i - \mathbf{F}(r^i, s)\|^2}{2\sigma^2}\right) \sigma^{-2} [\nabla_s \mathbf{F}(r^i, s)]^\top (\mathbf{t}^i - \mathbf{F}(r^i, s)) \quad (7.5)$$

para cada $r^i \in \mathbb{R}^3$ c.t.p. Mostremos ahora que para cada $s \in Q$, tenemos

$$\int_{\mathbb{R}^3} |\nabla_s p(t^i | r, s)| p_{prior}(r) dr < \infty,$$

lo que implicaría la derivabilidad de $s \mapsto p(t^i | s)$ en ese conjunto, y que

$$\nabla_s p(t^i | s) = \int_{\mathbb{R}^3} \nabla_s p(t^i | r^i, s) p_{prior}(r^i) dr^i. \quad (7.6)$$

Por (7.5), y dado que el soporte de $p_{prior}(r)$ es el conjunto compacto $G \subset \mathbb{R}^3$, es suficiente con mostrar que la función $r^i \mapsto [\nabla_s \mathbf{F}(r^i, s)]^\top (\mathbf{t}^i - \mathbf{F}(r^i, s))$ (donde $\mathbf{t}^i$ es constante) es acotada en $G$. Por el Lema 7.1, el factor $(\mathbf{t}^i - \mathbf{F}(r^i, s))$ es acotado cuando $r^i \in G$. Más aún, gracias a los resultados y argumentos de la demostración de la Proposición 7.2 (Ver Anexo 7), obtenemos la siguiente cota

$$\|w^i_j(s)\| \leq \|J\|_{op} m^{-1} (\|S\|_\infty \|r^i - r^i_j\| + 1),$$

donde $m > 0$ es tal que $S \in \mathcal{B}(\mathbb{R}^3, [0, m))$. La última cantidad también es acotada par $r^i \in G$, y la integrabilidad afirmada se obtiene tomando en cuenta la expresión obtenida para $\nabla_s \mathbf{F}(r^i, s)$. Por lo tanto, la expresión (7.6) se cumple. Prosigamos para concluir la expresión

(6.14). Por (7.6), tenemos que

$$\begin{aligned}
\nabla_s \log p(\mathbf{t}^i|s) &= \frac{\nabla_s p(\mathbf{t}^i|s)}{p(\mathbf{t}^i|s)} \\
&= \frac{1}{p(\mathbf{t}^i|s)} \int_{\mathbb{R}^3} \nabla_s p(\mathbf{t}^i|r^i, s) p_{prior}(r^i) dr^i \\
&= \int_{\mathbb{R}^3} \frac{[\nabla_s \log p(\mathbf{t}^i|r^i, s)] p(\mathbf{t}^i|r^i, s) p_{prior}(r^i)}{p(\mathbf{t}^i|s)} dr^i \\
&= \int_{\mathbb{R}^3} (\nabla_s \log p(\mathbf{t}^i|r^i, s)) p_{post}(r^i|\mathbf{t}^i, s) dr^i,
\end{aligned}$$

donde usamos el teorema de Bayes para obtener la igualdad. Esto concluye la demostración. □