UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS MATEMÁTICAS Y DE LA ESPECIALIZACIÓN EN ESTADÍSTICA APLICADA

ESTUDIANDO LA CAPACIDAD DE GENERALIZACIÓN DEL GRADIENTE ESTOCÁSTICO DE LA DINÁMICA DE LANGEVIN BASÁNDOSE EN TEORÍA DE LA INFORMACIÓN

TESIS
QUE PARA OPTAR POR EL GRADO DE:
MAESTRO (A) EN CIENCIAS

PRESENTA:
CHRISTIAN RODRIGO CRUZ FLORES

DIRECTOR
DR. MARIO ALBERTO DIAZ TORRES
INSTITUTO DE INVESTIGACIONES EN MATEMÁTICAS APLICADAS Y EN SISTEMAS

CIUDAD UNIVERSITARIA, CIUDAD DE MÉXICO, 7 DE JUNIO DEL 2022.

Agradecimientos

Primeramente, quisiera agradecer al Dr. Mario Diaz no sólo por su guía durante la creación de este texto sino por la confianza que desde un principio depositó en mi trabajo.

También le debo mi agradecimiento al Dr. Eduardo Gutiérrez por su consejo a lo largo de estos dos años.

Asimismo, deseo agradecer de la manera más sincera a mis padres, Lupita y Felipe, por su apoyo incondicional ante todas mis decisiones. A Mariela, al señor Gerardo, a la señora María y a la familia Rodríguez por siempre brindarme un segundo hogar.

Finalmente, le debo un agradecimiento especial a Eduardo Gomezcaña y al Dr. Arturo Erdely por sus atinadas recomendaciones, las cuales me llevaron a tomar una de las mejores decisiones de mi vida: estudiar una maestría.

Índice general

Símbolos

Símbolo	Descripción
$\mathbb{R}^+$	El conjunto de los números reales positivos
$\lvert\mathcal{X}\rvert$	La cardinalidad del conjunto $\mathcal{X}$
$X \perp\!\!\!\perp Y$	Denota independencia entre las variables aleatorias X y Y
$\mathcal{M}_{n\times m}(\mathbb{R})$	El conjunto de matrices de dimensión $n \times m$ con entradas en los reales
$\oplus$	La suma de enteros módulo 2
$N_d(\mu, \Sigma)$	La distribución normal d-variada con vector de medias μ y matriz de covarianza Σ
Leb	La medida de Lebesgue
$\lvert\Sigma\rvert$	El determinante de la matriz Σ
$[T]$	El subconjunto de números enteros $\{1, 2, \ldots, T\}$
$\nabla_{\mathbf{w}} f(\mathbf{w}, \mathbf{z})$	El gradiente de $f(\mathbf{w}, \mathbf{z})$ al fijar el vector $\mathbf{z}$
$\mathcal{X}^{\mathbb{N}}$	El conjunto de todas las sucesiones con entradas en $\mathcal{X}$
X^n	Denota al vector $(X_1, X_2, \ldots, X_n)$
2^{Ω}	El conjunto potencia de Ω
$\mu^{\otimes n}$	La medida producto $\underbrace{\mu \otimes \mu \otimes \cdots \otimes \mu}_{n\text{-veces}}$

Capítulo 1

Introducción

Los grandes avances tecnológicos de los últimos años, particularmente el incremento del poder computacional disponible para un individuo y la enorme cantidad de datos que se almacenan día con día, han extendido el uso de técnicas estadísticas a casi todos los sectores del quehacer humano. Más aún, el desarrollo de algoritmos estadísticos sofisticados, como las Redes Neuronales Artificiales (ANN), han convertido esta tendencia en un éxito rotundo. Sin embargo, de manera paradójica, el aprendizaje automático también se ha transformado en una caja negra para muchos; aún para aquellos familiarizados con la matemática y la computación.

Es importante notar que, a pesar de la inmensa cantidad de cursos y tutoriales disponibles en la web que enseñan a implementar algoritmos de aprendizaje automático, falta material, particularmente en nuestro idioma, que brinde las bases teóricas necesarias para comprender de una manera profunda, inclusive elemental, el marco teórico bajo el cual se desarrolla el aprendizaje estadístico supervisado, así como la intuición y las garantías detrás de sus procedimientos.

1.1. Objetivo y alcance

El objetivo de esta tesis es presentar la teoría necesaria para estudiar, formalmente, un aspecto del aprendizaje máquina supervisado (en este caso, el concepto de generalización) con base en un artículo de investigación contemporáneo. Más aún, se busca que la teoría expuesta parta de conceptos que deberían ser familiares para un egresado de una maestría orientada a la probabilidad y/o estadística.

El alcance de este texto es describir el planteamiento teórico de un problema de aprendizaje automático supervisado usando como principal referencia Shalev-Schwarz y Ben-David (2014). Asimismo, presentar de manera formal, pero no

exhaustiva, conceptos provenientes de diversas ramas de las matemáticas (concretamente, teoría de la información, optimización convexa y transporte óptimo) con el objetivo de revisar los resultados presentados en Wang y col. (2021) de la manera más autónoma posible.

1.2. Relevancia del estudio

La principal motivación detrás de esta tesis es contribuir a la bibliografía disponible a los egresados de maestrías en probabilidad y estadística en el país con una introducción al estudio del aprendizaje máquina desde un enfoque teórico y formal; además de proveer referencias concretas para ahondar en los temas presentados. Más aún, el autor busca contribuir con la perspectiva de un alumno que siempre se ha visto fascinado por los resultados empíricos de las técnicas estadísticas contemporáneas pero que, varias veces, se ha encontrado decepcionado con la falta de claridad en los recursos que explican dichos métodos. Adicionalmente, este texto tiene el propósito de presentar conceptos y resultados construyendo sobre teoría bien conocida de probabilidad y estadística y explicarlos de una manera más detallada a la que se presenta en partes de la bibliografía consultada durante el desarrollo de esta tesis; concretamente, en los temas de aprendizaje automático y teoría de la información.

A pesar de que este trabajo no está diseñado para ser leído antes de cursar una maestría en matemáticas, se espera que el presente sirva a algunos estudiantes para notar el gran esfuerzo que implica entender sólidamente algunos algoritmos de aprendizaje automático y que este proceso requiere de estudiar matemáticas más allá de lo que se puede hacer en una licenciatura o una maestría.

1.3. Esquema general

Esta tesis consiste de cinco capítulos, adicionales al presente, divididos en tres partes. En la primera de éstas (Capítulos 2, 3 y 4), se expone la teoría básica correspondiente a la Teoría de la Información, la Optimización Convexa, el Transporte Óptimo y el Aprendizaje Máquina, respectivamente. El Capítulo 2 introduce conceptos fundamentales de la Teoría de la Información: la entropía de Shannon, la divergencia de Kullback-Leibler y la información mutua junto con algunas propiedades de éstas. El objetivo del Capítulo 2 es ayudarnos a comprender mejor la información mutua para entender cómo es utilizada en Wang y col. (2021) para estudiar el concepto de generalización en el aprendizaje automático supervisado.

El Capítulo 3 presenta algunas bases teóricas de la optimización convexa necesarias para resolver problemas de aprendizaje estadístico. Adicionalmente, en este capítulo se introducen, de manera superficial, algunas ideas del Transporte Óptimo. Éstas se utilizan para acotar la información mutua, lo que a su vez nos permitirá describir la capacidad de generalización de un algoritmo en términos de cantidades fáciles de estimar. Finalmente, el Capítulo 4 se ayuda, particularmente, del Capítulo 3 para presentar una teoría enfocada al estudio formal del aprendizaje automático supervisado y la resolución de problemas de predicción/clasificación.

En la segunda parte de este trabajo (Capítulo 5), se presentan algunos resultados de artículos recientes (Xu y Raginsky 2017 y Bu, Zou y Veeravalli 2020) que juegan un papel esencial en Wang y col. (2021) para analizar la capacidad de generalización de un algoritmo de aprendizaje. Concretamente, se prueba una selección de resultados de Xu y Raginsky (2017) y Bu, Zou y Veeravalli (2020) de los cuales se deducen los lemas necesarios para demostrar el Teorema 1 en Wang y col. (2021).

Capítulo 2

Preliminares de Teoría de la Información

Como se mencionó en la Introducción, el objetivo de esta tesis es presentar las bases teóricas necesarias para estudiar la capacidad de generalización de un algoritmo de aprendizaje automático supervisado conforme se detalla en Wang y col. (2021); concretamente, del **descenso por gradiente de la dinámica de Langevin**. Dos de los recursos teóricos que nos permitirán describir esta capacidad de generalización provienen de la teoría de la información: la **información mutua** y las **desigualdades fuertes de procesamiento de la información**.

El objetivo de este capítulo es brindar una introducción breve y formal a la Teoría de la Información. Se revisarán desde conceptos fundamentales de esta rama de las matemáticas, como la entropía de Shannon, hasta aquellos que aún conciernen artículos de investigación recientes; específicamente, coeficientes de contracción asociados a f-divergencias. Este capítulo permite presentar los resultados de Wang y col. (2021) en el Capítulo 5 de la manera más autosuficiente posible dentro del alcance de este texto.

La primera sección de este capítulo intenta motivar el uso de métodos provenientes de la teoría de la información en el análisis estadístico. La segunda sección se centra en el concepto clave detrás de la teoría de la información: la **entropía de Shannon**. A pesar de que ésta no juega un papel dentro de los resultados en Wang y col. (2021), su estudio es la introducción, por excelencia, a los objetivos generales de la teoría de la información. Posteriormente, con ayuda de la derivada de Radon-Nykodim, se introduce la Divergencia de Kullback-Leibler en un contexto general. Ésta funge como una pseudodistancia entre distribuciones de probabilidad por lo que es ideal para construir una herramienta que intenta cuantificar la

dependencia entre variables aleatorias: la **información mutua**. Finalmente, en la última sección se generaliza la noción detrás de la divergencia de Kullback-Leibler mediante el estudio de la familia de f-divergencias y las desigualdades fuertes de procesamiento de la información.

2.1. Introducción a la teoría de la información y su conexión con la inferencia estadística

Uno de los principales objetivos de la teoría de la información es: « [...] mejorar el diseño de señales, y de los canales mediante los cuales éstas se transmiten, para comunicar y almacenar información de manera óptima, además de permitir la decodificación más efectiva» (Duchi 2019, p. 6). En este sentido, la teoría de la información proporciona un marco dentro del cual se pueden plantear y resolver problemas de naturaleza estadística.

Por ejemplo, aunque al realizar inferencias estadísticas los canales ya existen de manera natural, a través de éstos se reciben señales (muestras aleatorias) las cuales se desean codificar (resumir) con ayuda de una transformación adecuada (e.g. un estadístico que estime algún parámetro de interés, medidas de tendencia central y/o de variación). Posteriormente, se desea decodificar este resumen para entender mejor el fenómeno o para predecir futuras observaciones.

En particular, de acuerdo con el texto introductorio (Duchi 2019), existen dos preguntas que competen a la teoría de la información que son análogamente relevantes para la disciplina estadística:

1. ¿Cúanta información contiene una señal?
2. ¿Cuánta información puede transmitir de manera confiable un canal con ruido?

La primera de éstas, naturalmente, puede relacionarse con la información que provee una muestra acerca del fenómeno que la genera. La segunda es otra manera de preguntarse cuánta información se puede extraer de una variable Y si se observa X y éstas están relacionadas por una distribución (canal) $P_{Y \mid X}$. Nuestra introducción a la teoría de la información comienza definiendo una función conocida como la entropía de Shannon la cual pretende contestar la primera interrogante.

2.2. Entropía

En esta sección se presenta la definición de la entropía de Shannon, su versión condicional y algunos ejemplos de éstas. Asimismo, se muestran algunas propiedades que complementan la intuición detrás de este concepto.

2.2.1. Definición

Definición 2.1 (Polyanskiy y Wu 2019, Definición 1.1)**.** Sea X una variable aleatoria que toma valores en un conjunto contable $\mathcal{X}$ según la medida de probabilidad P_X. Se define la entropía de X como

$$H(X) \coloneqq \mathbb{E}_X\left(\log \frac{1}{P_X(X)}\right) = \sum_{x\in\mathcal{X}} P_X(x) \log \frac{1}{P_X(x)}. \tag{2.1}$$

Caben recalcar dos puntos importantes de la definición anterior:

- Ésta no se limita al caso univariado.
- $H(X)$ es una función de P_X por lo que es común denotar por $H(P_X)$ a la entropía de X.

Ejemplo 2.1 (Bernoulli)**.** Sea X una variable aleatoria Bernoulli(p). Con base en (2.1) se tiene

$$H(X) = p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p}. \tag{2.2}$$

A la función de p que define (2.2) se le conoce como la entropía binaria y se denota usualmente por $h_b(p)$.

Ejemplo 2.2 (Uniforme)**.** Sea X una variable aleatoria uniforme en un conjunto finito $\mathcal{X}$. Luego,

$$H(X) = \sum_{x\in\mathcal{X}} \frac{1}{|\mathcal{X}|} \log |\mathcal{X}| = \log |\mathcal{X}|. \tag{2.3}$$

2.2.2. Entropía condicional

Inspirándose en la Definición 2.1, se puede construir un instrumento que busca cuantificar la información restante en una variable aleatoria X después de observar otra variable Y.

Definición 2.2 (Polyanskiy y Wu 2019, Definición 1.3)**.** Sean X y Y dos variables aleatorias discretas. Se define la entropía condicional de X dado Y como

$$H(X \,|\, Y) \coloneqq \mathbb{E}_{Y\sim P_y}[H(P_{X\,|\,Y}(\cdot \,|\, Y))] = \mathbb{E}\left(\log \frac{1}{P_{X\,|\,Y}(X \,|\, Y)}\right). \tag{2.4}$$

Ejemplo 2.3 (Canal Simétrico Binario)**.** Sean X y Z dos variables Bernoulli independientes de parámetros $p, \alpha \in [0,1]$ respectivamente. Defínase $Y \coloneqq X \oplus Z$, donde $\oplus$ representa la suma en $\mathbb{Z}_2$. A partir de (2.4) es directo calcular

$$H(Y \mid X) = \alpha \log \frac{1}{\alpha} + (1-\alpha) \log \frac{1}{1-\alpha} = h_b(\alpha).$$

Con el objetivo de probar algunas propiedades de la entropía y la entropía diferencial, así como otros resultados a lo largo de esta sección, se recurre a la desigualdad de Jensen para funciones estrictamente convexas, la cual se enuncia y prueba a continuación.

Proposición 2.1 (Desigualdad de Jensen)**.** *Sea $f : \mathbb{R} \to \mathbb{R}$ una función estrictamente convexa. Si X y $f(X)$ son variables aleatorias integrables, entonces*

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)],$$

con igualdad si y sólo si X es una constante casi seguramente; i.e. $X = \mathbb{E}(X)$.

Demostración. De acuerdo con la Observación 1.6.4 en Niculescu y Persson (2018), f es una función estrictamente convexa si y sólo si para todo $a \in \mathbb{R}$, existe $\lambda \in \mathbb{R}$ tal que

$$f(x) > f(a) + \lambda(x-a), \ \ \forall x \in \mathbb{R} \setminus \{a\}. \tag{2.5}$$

Así, para $x_0 = \mathbb{E}[X]$, existe $\lambda_0 \in \mathbb{R}$ tal que

$$f(X) \geq f(x_0) + \lambda_0(X - x_0),$$

donde la igualdad se da solamente si $X = x_0$. Debido a la monotonicidad de la esperanza, lo anterior implica que

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X]). \tag{2.6}$$

Adicionalmente, si se asume que $\mathbb{E}[f(X)] = f(\mathbb{E}[X])$, entonces

$$\mathbb{E}\left[f(X) - (f(x_0) + \lambda_0(X - x_0))\right] = 0. \tag{2.7}$$

Sin embargo, nótese que $f(X) - (f(x_0) + \lambda_0(X - x_0)$ es una variable aleatoria no-negativa por lo que (2.7) se cumple solamente si $f(X) = f(x_0) + \lambda_0(X - x_0)$ casi seguramente. Luego, como f es estrictamente convexa, se sigue que $X = x_0 = \mathbb{E}[X]$ casi seguramente. Finalmente, suponer que X es una constante c.s. implica directamente que $\mathbb{E}[f(X)] = f(\mathbb{E}[X])$ c.s. □

La siguiente proposición nos brinda algunas propiedades de la entropía y la entropía condicional. El lector puede notar fácilmente lo útil que resulta la Proposición 2.1 en las siguientes demostraciones.

Proposición 2.2 (Polyanskiy y Wu 2019, Teorema 1.1)**.** *Algunas propiedades de H son:*

1. *$H(X) \geq 0$ con igualdad solamente si X es una constante.*
2. *Si $|\mathcal{X}| < \infty$, entonces $H(X) \leq \log |\mathcal{X}|$, con igualdad si y sólo si X tiene distribución uniforme.*
3. *$H(X) = H(f(X))$ para toda función biyectiva f.*
4. *Dos desigualdades de la entropía condicional son:*
   a) *$0 \leq H(X \,|\, Y)$, con igualdad si y sólo si $X = f(Y)$ para alguna función f.*
   b) *$H(X \,|\, Y) \leq H(X)$, con igualdad si y sólo si X y Y son independientes.*
5. *$H(X, Y) = H(X) + H(Y \,|\, X) \leq H(X) + H(Y)$.*
6. *$H(X) = H(X, f(X)) \geq H(f(X))$ con igualdad si y sólo si f es uno a uno en el soporte de X.*
7. *Si se define $X^i \coloneqq (X_1, \ldots, X_i)$, entonces $H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i \,|\, X^{i-1}) \leq \sum_{i=1}^{n} H(X_i)$, donde se cumple la igualdad si y sólo si $X_1, \ldots, X_n$ son independientes a pares.*

Demostración.

1. Como $P_X(x) \leq 1$ para todo $x \in \mathcal{X}$, entonces $-\log P_X(X) \geq 0$ casi seguramente. Más aún, la esperanza de una variable aleatoria no-negativa es cero si y sólo si ésta es cero casi seguramente; i.e., $H(X) = \mathbb{E}(-\log P_X(X)) = 0$ si y sólo si $P_X(X) = 1$ casi seguramente (X es determinista).
2. Dado que $x \mapsto \log x$ es estrictamente cóncava en $(0, \infty)$, la desigualdad de Jensen (Proposición 2.1) implica que
$$\log \mathbb{E}\left(\frac{1}{P_X(X)}\right) \geq \mathbb{E}\left(\log \frac{1}{P_X(X)}\right), \tag{2.8}$$
con igualdad si y sólo si $\frac{1}{P_X(X)} = \mathbb{E}\left(\frac{1}{P_X(X)}\right)$ casi seguramente. Luego, como $\mathbb{E}\left(\frac{1}{P_X(X)}\right) = \sum_{x \in \mathcal{X}} P_X(x) \cdot \frac{1}{P_X(x)} = |\mathcal{X}|$, de la ecuación (2.8) se sigue lo deseado.
3. Sea $f(\mathcal{X}) = \{f(x) : x \in \mathcal{X}\}$. Nótese que $P_{f(X)}(y) = \mathbb{P}(f(X) = y)$. En virtud de que f es biyectiva, para toda $y \in f(\mathcal{X})$ existe un único $x \in \mathcal{X}$ tal que

$f(x) = y$. Así, para todo $x \in \mathcal{X}$, se tiene $P_{f(X)}(f(x)) = P_X(x)$. De tal suerte que

$$\begin{aligned} H(f(X)) &= \sum_{x \in \mathcal{X}} P_{f(X)}(f(x)) \log \frac{1}{P_{f(X)}(f(x))} \\ &= \sum_{x \in \mathcal{X}} P_X(x) \log \frac{1}{P_X(x)} \\ &= H(X). \end{aligned}$$

4.

a) Del primer resultado de esta proposición se sigue que $H(P_{X \mid Y}(\cdot \mid y)) \geq 0$ para todo $y \in \mathcal{Y}$. Es decir, $H(P_{X \mid Y}(\cdot \mid Y)) \geq 0$. La desigualdad buscada se obtiene apelando a la monotonicidad de la esperanza y la Definición 2.2. Adicionalmente, $\mathbb{E}[H(P_{X \mid Y}(\cdot \mid Y))] = 0$ si y sólo si $H(P_{X \mid Y}(\cdot \mid Y)) = 0$ casi seguramente. Es decir, $P_{X \mid Y}$ es una distribución determinista. Por lo tanto, $H(X \mid Y) = 0$ si y sólo si $X = f(Y)$ para alguna función f.

b) Nótese que

$$\log \frac{1}{P_{X \mid Y}(X \mid Y)} = \log \frac{P_X(X)P_Y(Y)}{P_{X,Y}(X,Y)} + \log \frac{1}{P_X(X)}. \tag{2.9}$$

Al tomar la esperanza de ambos lados de la igualdad con respecto a la medida $P_{X,Y}$ se obtiene

$$H(X \mid Y) = H(X) + \mathbb{E}\left[\log \frac{P_X(X)P_Y(Y)}{P_{X,Y}(X,Y)}\right]. \tag{2.10}$$

Dado que $x \mapsto \log x$ es estrictamente cóncava, al aplicar la desigualdad de Jensen a (2.10),

$$H(X \mid Y) \leq H(X) + \log \mathbb{E}\left[\frac{P_X(X)P_Y(Y)}{P_{X,Y}(X,Y)}\right] = H(X), \tag{2.11}$$

con igualdad si y sólo si $\frac{P_X(X)P_Y(Y)}{P_{X,Y}(X,Y)} = 1$ casi seguramente. Es decir, $X \perp\!\!\!\perp Y$.

5. Es directo deducir

$$\log \frac{1}{P_{Y \mid X}(Y \mid X)} = \log \frac{1}{P_{X,Y}(X,Y)} - \log \frac{1}{P_X(X)}. \tag{2.12}$$

Al integrar respecto a la medida $P_{X,Y}$ se obtiene

$$H(Y \mid X) = H(X,Y) - H(X). \tag{2.13}$$

Luego, al reacomodar la ecuación (2.13) y aplicar el resultado 4, se concluye

$$H(X,Y) = H(X) + H(Y \mid X) \leq H(X) + H(Y). \tag{2.14}$$

6. Nótese que para toda función f, la relación $x \mapsto (x, f(x))$ es biyectiva. Así, debido a 3, 4a y 5 se sigue

$$H(X) = H(X, f(X)) = H(f(X)) + H(X \mid f(X)) \geq H(f(X)). \tag{2.15}$$

Nótese que $H(X \mid f(X)) = 0$ si y sólo si X es una función determinista de $f(X)$. Es decir, $H(X) = H(f(X))$ si y sólo si f tiene inversa.

7. Dado que $P_{X_1,X_2,\ldots,X_n} = P_{X_1} P_{X_2 \mid X_1} \cdots P_{X_n \mid X^{n-1}}$, entonces

$$\log \frac{1}{P_{X_1,\ldots,X_n}(X_1,\ldots,X_n)} = \sum_{i=1}^{n} \log \frac{1}{P_{X_i \mid X^{i-1}}(X_i \mid X^{i-1})}. \tag{2.16}$$

Tomando la esperanza de ambos lados de la igualdad y aplicando 4b se obtiene el resultado buscado.

□

Hasta el momento se ha definido la entropía de Shannon y su versión condicional. Además, hemos enunciado algunas propiedades de éstas que concuerdan con hipótesis intuitivas de cómo se comporta la información en la realidad. Por ejemplo, la Proposición 2.2.3 nos dice que una función biyectiva preserva la información que nos brinda una variable aleatoria mientras que la Proposición 2.2.6 nos asegura que no existe una transformación que se pueda aplicar a una variable aleatoria para obtener más información de ésta. Sin embargo, la entropía presenta algunas limitantes; la más evidente es que sólo está definida en contextos discretos. La siguiente sección busca construir un instrumento con usos similares pero útil en situaciones generales.

2.3. Divergencia de Kullback-Leibler

En esta sección se define una herramienta aplicable en contextos más generales que la entropía. Particularmente, esta última sólo está definida para variables aleatorias discretas. Motivado en lo anterior, en esta sección se define una especie de distancia entre medidas de probabilidad: la divergencia de Kullback-Leibler. Asimismo, se revisa su versión condicional y algunas propiedades de éstas.

Debido a que se desea construir un utensilio que sea aplicable en un espacio medible arbitrario, éste se debe definir con ayuda de un concepto suficientemente general. En nuestro caso, se recurrirá a la derivada de Radon-Nikodym, la cual es el enfoque principal de la siguiente subsección.

2.3.1. Derivada de Radon-Nikodym

Definición 2.3 (Kallenberg 2002)**.** Sea $f \geq 0$ una función medible en un espacio de medida $(\Omega, \mathcal{F}, \mu)$. Se define

$$(f \cdot \mu)(A) \coloneqq \int_A f \, d\mu, \;\; A \in \mathcal{F}.$$

Nótese que la función $\nu \coloneqq f \cdot \mu$ vuelve a ser una medida en $(\Omega, \mathcal{F})$. En este caso, f es conocida como la μ-densidad de ν.

El siguiente resultado es una herramienta útil cuando se trabaja con este tipo de densidades. La prueba de éste es directa mediante el método estándar.

Proposición 2.3 (Kallenberg 2002, Lema 1.23)**.** *En un espacio de medida* $(\Omega, \mathcal{F}, \mu)$*, dadas dos funciones medibles* $f : \Omega \to \mathbb{R}^+$ *y* $g : \Omega \to \mathbb{R}$*, se tiene* $\int fg \, d\mu = \int g \, d(f \cdot \mu)$ *siempre que cualquier lado de la igualdad exista.*

Es claro que la existencia de esta clase de densidades puede resultar conveniente en muchas situaciones por lo que es natural preguntarse cuándo esto se puede asegurar. A esta interrogante responde el Teorema de Radon-Nikodym donde las siguientes dos definiciones juegan un papel central.

Definición 2.4 (Kallenberg 2002, p. 12)**.** Dado un espacio de medida $(\Omega, \mathcal{F}, \mu)$, se dice que una relación entre funciones se cumple casi en todas partes con respecto a μ (abreviado μ-*a.e.*) si ésta se cumple para todo $\omega \in \Omega$ excepto en un conjunto $A \subset \Omega$ tal que $\mu(A) = 0$.

Definición 2.5 (Jacod y Protter 2004, Definición 28.1)**.** Sean μ y ν dos medidas finitas del espacio medible $(\Omega, \mathcal{F})$. Se dice que ν es absolutamente continua con respecto de μ si para todo $A \in \mathcal{F}$ tal que $\mu(A) = 0$ se tiene que $\nu(A) = 0$. Lo anterior se denota por $\nu \ll \mu$.

Demostrar el Teorema de Radon-Nikodym va más allá del alcance de este trabajo. Sin embargo, con base en el Teorema 2.10 en Kallenberg (2002) se enuncia dicho resultado.

Teorema 2.1 (Teorema de Radon-Nikodym)**.** *Dadas dos medidas finitas* μ *y* ν *en* $\mathcal{X}$ *tales que* $\nu \ll \mu$ *existe una* μ*-a.e. única función medible* $f \geq 0$ *con dominio en* $\mathcal{X}$ *tal que* $\nu = f \cdot \mu$*.*

Es común denotar a f como $\frac{d\nu}{d\mu}$ y referirse a ésta como la derivada de Radon-Nikodym. Nótese que

$$\int_A d\nu = \nu(A) = (f \cdot \mu)(A) = \int_A f \, d\mu = \int_A \frac{d\nu}{d\mu} \, d\mu;$$

es decir, esta notación permite escribir:

$$\int_A \frac{d\nu}{d\mu}\, d\mu = \int_A d\nu.$$

Más aún, en virtud de la Proposición 2.3, si $\nu \ll \mu$, entonces para cualquier función medible g se cumple

$$\int \left(\frac{d\nu}{d\mu} \cdot g \right) d\mu = \int g\, d\nu,$$

siempre que cualquiera de los dos lados de la igualdad exista. Cabe mencionar que si ν y μ son distribuciones de variables aleatorias discretas, $\frac{d\nu}{d\mu}$ es el cociente de las funciones de masa de probabilidades y, análogamente, si ν y μ son distribuciones de variables absolutamente continuas, entonces $\frac{d\nu}{d\mu}$ es el cociente de las funciones de densidad de probabilidades.

La derivada de Radon-Nikodym es la piedra angular en la definición de la Divergencia de Kullback-Leibler; y, en general, de cualquier f-divergencia. Debido a esto, el desarrollo anterior será de gran utilidad en la prueba de algunas proposiciones en un contexto general.

2.3.2. Definición

Con base en la teoría previa, se define la Divergencia de Kullback-Leibler en su versión más general.

Definición 2.6 (Polyanskiy y Wu 2019, Definición 1.4)**.** Sea $(\Omega, \mathcal{F})$ un espacio medible y sean P y Q dos medidas de probabilidad en éste. Definimos la divergencia de Kullback-Leibler como

$$D(P\,||\,Q) = \begin{cases} \mathbb{E}_P\left(\log \frac{dP}{dQ}\right) = \mathbb{E}_Q\left(\frac{dP}{dQ} \log \frac{dP}{dQ}\right), & P \ll Q; \\ \infty, & \text{cualquier otro caso.} \end{cases} \tag{2.17}$$

Adicionalmente, se establece, por convención, que

$$0 \log \frac{0}{0} = 0. \tag{2.18}$$

En particular, si $\Omega = \mathbb{R}$, $\mathcal{F} = \mathcal{B}(\mathbb{R})$, Leb denota a la medida de Lebesgue en $\mathbb{R}$, tanto P como Q son absolutamente continuas con respecto a la medida de Lebesgue y las funciones de densidad de probabilidades respectivas se denotan por p y q, entonces

$$D(P\,||\,Q) = \begin{cases} \int p(x) \log \frac{p(x)}{q(x)} dx, & \text{Leb}(p > 0, q = 0) = 0; \\ \infty, & \text{cualquier otro caso.} \end{cases} \tag{2.19}$$

A continuación se muestran algunos ejemplos que ilustran cómo utlizar (2.17) y (2.19).

Ejemplo 2.4 (Bernoulli)**.** Sean P y Q dos medidas de probabilidad correspondientes a un modelo Bernoulli de parámetros p y q respectivamente, donde $p, q \in [0,1]$. La definición establecida en (2.17) dicta que

$$D(P \,||\, Q) = \begin{cases} p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}, & p, q \in (0,1) \text{ ó } p = q; \\ \infty, & \text{cualquier otro caso.} \end{cases} \tag{2.20}$$

Es común referirise a la función de $(p, q) \in [0,1]^2$ inducida por (2.20) como la divergencia binaria y denotarla por $d(p \,||\, q)$.

Ejemplo 2.5 (Normal multivariada)**.** Sean X_1 y X_2 variables aleatorias normales n-variadas con media y varianza $\mu_1 \in \mathbb{R}^n$, $\Sigma_1 \in \mathcal{M}_{n\times n}(\mathbb{R})$ y $\mu_2 \in \mathbb{R}^n$, $\Sigma_2 \in \mathcal{M}_{n\times n}(\mathbb{R})$ respectivamente, donde Σ_1 y Σ_2 son matrices invertibles. En virtud de lo anterior, es claro que las distribuciones P_{X_1} y P_{X_2} son absolutamente continuas con respecto a la medida de Lebesgue. Denótese por p_{X_1} y p_{X_2} a las funciones de densidad correspondientes. Así, con base en (2.19),

$$D(P_{X_1} \,||\, P_{X_2}) = \int p_{X_1}(\mathbf{x}) \log \left(\frac{p_{X_1}(\mathbf{x})}{p_{X_2}(\mathbf{x})} \right) d\mathbf{x}. \tag{2.21}$$

Nótese que

$$\log \left(\frac{p_{X_1}(\mathbf{x})}{p_{X_2}(\mathbf{x})} \right) = \frac{1}{2} \left[\log \frac{|\Sigma_2|}{|\Sigma_1|} - (\mathbf{x}-\mu_1)^\top \Sigma_1^{-1} (\mathbf{x}-\mu_1) + (\mathbf{x}-\mu_2)^\top \Sigma_2^{-1} (\mathbf{x}-\mu_2) \right], \tag{2.22}$$

donde $|\Sigma_i|$ denota el determinante de la matriz Σ_i. Es fácil corroborar que $X_1 - \mu_1$ y $X_1 - \mu_2$ son variables aleatorias normales n-variadas con matriz de covarianza Σ_1 y vectores de medias $\mathbf{0}$ y $(\mu_1 - \mu_2)$ respectivamente. En virtud de (Seber y Lee 2003, Teorema 5.1) se tiene:

$$\mathbb{E}[(X_1-\mu_1)^\top \Sigma_1^{-1} (X_1-\mu_1)] = \operatorname{tr}(\Sigma_1^{-1}\Sigma_1) = \operatorname{tr}(I_n) = n. \tag{2.23}$$

$$\mathbb{E}[(X_1-\mu_2)^\top \Sigma_2^{-1} (X_1-\mu_2)] = \operatorname{tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_1-\mu_2)^\top \Sigma_2^{-1} (\mu_1-\mu_2). \tag{2.24}$$

Sustituyendo la ecuación (2.22) en (2.21) y recurriendo a los resultados (2.23) y (2.24) se concluye que

$$D(P_{X_1} \,||\, P_{X_2}) = \frac{1}{2} \left(\log \frac{|\Sigma_2|}{|\Sigma_1|} + \operatorname{tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_1-\mu_2)^\top \Sigma_2^{-1} (\mu_1-\mu_2) - n \right). \tag{2.25}$$

Como ya se mencionó, la divergencia de K-L puede interpretarse como una especie de distancia entre medidas de probabilidad. Naturalmente, una de las primeras características que uno se cuestiona acerca de la divergencia es la positividad. La siguiente proposición responde afirmativamente a esa pregunta.

Proposición 2.4 (Polyanskiy y Wu 2019, Teorema 2.1)**.** *Sean P y Q dos medidas de probabilidad sobre un mismo espacio medible $(\Omega, \mathcal{F})$. Se tiene que*

$$D(P \,||\, Q) \geq 0, \tag{2.26}$$

con igualdad si y sólo si $P = Q$.

Demostración. Sea $\phi : \mathbb{R}^+ \to \mathbb{R}$ dada por $\phi(x) = x \log x$. Claramente, $\phi''(x) > 0$. Así, ϕ es estrictamente convexa. En virtud de la Desigualdad de Jensen (Proposición 2.1) y la Proposición 2.3

$$D(P \,||\, Q) = \mathbb{E}_Q \left(\frac{dP}{dQ} \log \frac{dP}{dQ} \right) \geq \phi \left(\mathbb{E}_Q \left(\frac{dP}{dQ} \right) \right) = \phi(\mathbb{E}_P(1)) = \phi(1) = 0.$$

con igualdad si y sólo si $\frac{dP}{dQ} = 1$ Q-c.s.; es decir, $P = Q$. □

Hasta el momento se ha definido la divergencia de K-L y se han presentando algunos ejemplos de ésta. Sin embargo, para continuar estudiando a fondo algunas de sus características es necesario involucrar uno de los componentes más importantes y útiles dentro de la probabilidad contemporánea: la probabilidad condicional.

2.3.3. Divergencia condicional

De acuerdo con Polyanskiy y Wu (2019), una de la maneras más comúnes para crear nuevas variables aleatorias de alguna ya existente o definir relaciones entre éstas es mediante transformaciones aleatorias. Dichas transformaciones se especifican mediante un kernel de transición:

Definición 2.7 (Kallenberg 2002, p. 20)**.** Dados dos espacios de probabilidad $(F, \mathcal{F})$ y $(G, \mathcal{G})$, una función $k : \mathcal{G} \times F \to [0, 1]$ será llamada un kernel de transición de F a G si $k(\cdot \,|\, x)$ es una medida de probabilidad en $(G, \mathcal{G})$ para todo $x \in F$ y $k(B \,|\, \cdot)$ es una función $\mathcal{F}$-medible para todo $B \in \mathcal{G}$.

Nota. A lo largo de este texto se denota por $P_{Y \,|\, X}$ a un kernel de $\mathcal{X}$ a $\mathcal{Y}$ con el objetivo de que sean claros los espacios en los que éste opera.

Definición 2.8 (Polyanskiy y Wu 2019, Definición 2.2)**.** Sean $P_{Y\,|\,X}$, $Q_{Y\,|\,X}$ kernels de transición y P_X una distribución de probabilidad. La divergencia condicional de Kullback-Leibler se define como

$$D(P_{Y\,|\,X}\,||\,Q_{Y\,|\,X}\,|\,P_X) = \mathbb{E}_{X\sim P_X}[D(P_{Y\,|\,X}(\cdot\,|\,X)\,||\,Q_{Y\,|\,X}(\cdot\,|\,X))]. \tag{2.27}$$

Ejemplo 2.6 (Canal Simétrico Binario)**.** Sean $\alpha_1, \alpha_2, p \in [0,1]$. Tómense $X \sim$ Bernoulli(p), $Z_1 \sim$ Bernoulli(α_1) y $Z_2 \sim$ Bernoulli(α_2) independientes entre sí. Luego, defínase $Y_1 \coloneqq X \oplus Z_1$ y $Y_2 \coloneqq X \oplus Z_2$, donde $\oplus$ representa la suma de enteros módulo 2. Nótese que

$$\begin{aligned} &D(P_{Y_1\,|\,X}(\cdot\,|\,X)\;||\;P_{Y_2\,|\,X}(\cdot\,|\,X)) \\ &\quad= [(1-\alpha_1)\mathbb{1}_{\{X=0\}} + \alpha_1\mathbb{1}_{\{X=1\}}]\log\left(\frac{(1-\alpha_1)\mathbb{1}_{\{X=0\}} + \alpha_1\mathbb{1}_{\{X=1\}}}{(1-\alpha_2)\mathbb{1}_{\{X=0\}} + \alpha_2\mathbb{1}_{\{X=1\}}}\right) \\ &\quad+ [\alpha_1\mathbb{1}_{\{X=0\}} + (1-\alpha_1)\mathbb{1}_{\{X=1\}}]\log\left(\frac{\alpha_1\mathbb{1}_{\{X=0\}} + (1-\alpha_1)\mathbb{1}_{\{X=1\}}}{\alpha_2\mathbb{1}_{\{X=0\}} + (1-\alpha_2)\mathbb{1}_{\{X=1\}}}\right). \end{aligned} \tag{2.28}$$

Así, con base en (2.28) y la Definición 2.8, se obtiene

$$D(\mathrm{BSC}(\alpha_1)\;||\;\mathrm{BSC}(\alpha_2)\,|\,\mathrm{Bern}(p)) = d(\alpha_1\,||\,\alpha_2), \tag{2.29}$$

donde d está definida por (2.20) y $\mathrm{BSC}(\alpha_i)$ denota el kernel de transición $P_{Y_i\,|\,X}$.

Una vez definidas la divergencia de Kullback-Leibler y su versión condicional, la Proposición 2.6 nos provee una lista de propiedades y relaciones entre éstas. Sin embargo, antes de enunciar dichas propiedades se especifica alguna notación y se prueba un resultado fundamental en la prueba de esta proposición.

Definición 2.9. Sean $(F,\mathcal{F})$ y $(G,\mathcal{G})$ dos espacios medibles. Supóngase que μ es una medida de probabilidad sobre $(F,\mathcal{F})$ y k un kernel de transición de F a G. Con base en el Teorema 6.4 en Kallenberg (2002), se define la medida $\mu\otimes k$ sobre $\mathcal{F}\otimes\mathcal{G}$ mediante

$$\mu\otimes k(C) \coloneqq \int \mu(dx)\int \mathbb{1}_C(x,y)\,k(dy\,|\,x);\;\; C\in\mathcal{F}\otimes\mathcal{G}. \tag{2.30}$$

En general, el Teorema 6.4 en Kallenberg (2002) asegura que para una función integrable f con dominio en $F\times G$

$$\int f\,d(\mu\otimes k) = \int \mu(dx)\int f(x,y)\,k(dy\,|\,x). \tag{2.31}$$

Definición 2.10. En el contexto de la Definición 2.9, se define la medida *push-forward* $k\circ\mu$ sobre $(G,\mathcal{G})$ como

$$k\circ\mu(B) \coloneqq \mu\otimes k(F\times B),\;\; B\in\mathcal{G}.$$

En particular, si se tiene una variable aleatoria X que toma valores en $\mathcal{X}$ de acuerdo con una medida P_X y se tiene un kernel de transición $P_{Y\,|\,X}$ de $\mathcal{X}$ a un espacio $\mathcal{Y}$, uno se refiere a la distribución *pushforward* de Y inducida por X y el kernel $P_{Y\,|\,X}$ como $P_{Y\,|\,X} \circ P_X$. Usualmente, debido a la notación adoptada, se puede evitar especificar los espacios en los que operan las medidas y los kernels de transición.

Ejemplo 2.7. Supóngase que se tienen dos variables aleatorias X y Z que toman valores en $\{0,1\}$ de acuerdo a una distribución Bernoulli(p) y Bernoulli(α) respectivamente. Luego, se define la variable Y, como en el Ejemplo 2.6, mediante

$$Y \coloneqq X \oplus Z.$$

En este ejemplo se busca encontrar la distribución de probabilidad conjunta de X y Y con ayuda de la Definición 2.9. Concretamente, sea μ la medida de probabilidad asociada a una distribución Bernoulli(p) y sea $k : 2^{\{0,1\}} \times \{0,1\}$ el kernel de transición que cumple

$$k(0\,|\,1) = \alpha = k(1\,|\,0),$$
$$k(0\,|\,0) = 1 - \alpha = k(1\,|\,1)$$

(nótese que el kernel k ha aparecido en otros ejemplos y se le ha denotado por BSC(α)). En este caso, si $C \in 2^{\{0,1\}} \otimes 2^{\{0,1\}}$, entonces $C = A \times B$ para algunos $A, B \in 2^{\{0,1\}}$. Así,

$$P_{XY}(C) = \mu \otimes k(C) = \sum_{x \in A} \sum_{y \in B} \mu(x) k(y\,|\,x).$$

Asimismo, se tiene que

$$\begin{aligned} k \circ \mu(1) &= \mu \otimes k(\{0,1\} \times \{1\}) \\ &= (1-p)k(1\,|\,0) + pk(1\,|\,1) \\ &= (1-p)\alpha + p(1-\alpha). \end{aligned}$$

De igual manera,

$$\begin{aligned} k \circ \mu(0) &= \mu \otimes k(\{0,1\} \times \{0\}) \\ &= (1-p)k(0\,|\,0) + pk(0\,|\,1) \\ &= (1-p)(1-\alpha) + p\alpha. \end{aligned}$$

De tal manera que la distribución *pushforward* de Y, en este caso, es Bernoulli de parámetro $[1-p]\alpha + p[1-\alpha]$.

Proposición 2.5. *Sean Q_X y P_X dos medidas de probabilidad en un espacio medible $(F, \mathcal{F})$ y $P_{Y\,|\,X}$, $Q_{Y\,|\,X}$ dos kernels de transición de F a un espacio medible $(G, \mathcal{G})$. Defínase $Q_{XY} \coloneqq Q_X \otimes Q_{Y\,|\,X}$ y $P_{XY} \coloneqq P_X \otimes P_{Y\,|\,X}$. Si $P_{Y\,|\,X}(\cdot\,|\,x) \ll Q_{Y\,|\,X}(\cdot\,|\,x)$ para todo $x \in \mathcal{X}$ y $P_X \ll Q_X$, entonces $P_{XY} \ll Q_{XY}$ y, Q_{XY}-casi-seguramente,*

$$\frac{dP_{XY}}{dQ_{XY}} = \frac{dP_{Y\,|\,X}}{dQ_{Y\,|\,X}} \cdot \frac{dP_X}{dQ_X},$$

donde $\frac{dP_{Y\,|\,X}}{dQ_{Y\,|\,X}} \cdot \frac{dP_X}{dQ_X}(x, y) \coloneqq \frac{dP_{Y\,|\,X}(\cdot\,|\,x)}{dQ_{Y\,|\,X}(\cdot\,|\,x)}(y) \cdot \frac{dP_X}{dQ_X}(x)$.

Demostración. Nótese que para $C \in \mathcal{F} \otimes \mathcal{G}$

$$P_{XY}(C) = \int Q_X(dx) \int \mathbb{1}_C(x, y) \cdot \frac{dP_{Y\,|\,X}(\cdot\,|\,x)}{dQ_{Y\,|\,X}(\cdot\,|\,x)}(y) \cdot \frac{dP_X}{dQ_X}(x)\, Q_{Y\,|\,X}(dy\,|\,x). \quad (2.32)$$

Así, para todo $C \in \mathcal{F} \otimes \mathcal{G}$,

$$P_{XY}(C) = \int \mathbb{1}_C(x, y) \cdot \frac{dP_{Y\,|\,X}}{dQ_{Y\,|\,X}} \cdot \frac{dP_X}{dQ_X}(x, y)\, dQ_{XY}. \quad (2.33)$$

De (2.33), se sigue que si $Q_{XY}(C) = 0$, entonces $P_{XY}(C) = 0$, $C \in \mathcal{F} \otimes \mathcal{G}$; es decir, $P_{XY} \ll Q_{XY}$. Más aún, dado que $\frac{dP_{XY}}{dQ_{XY}}$ es Q_{XY}-c.s. la única función que cumple (2.33), se concluye que

$$\frac{dP_{XY}}{dQ_{XY}} = \frac{dP_{Y\,|\,X}}{dQ_{Y\,|\,X}} \cdot \frac{dP_X}{dQ_X}, Q_{XY} - c.s.$$

□

Nótese que si se toma $P_{Y\,|\,X} = Q_{Y\,|\,X}$ en el contexto de la Proposición 2.5, entonces

$$\frac{dP_X \otimes P_{Y\,|\,X}}{dQ_X \otimes P_{Y\,|\,X}} = \frac{dP_X}{dQ_X}. \quad (2.34)$$

Proposición 2.6 (Polyanskiy y Wu 2019, Teorema 2.2)**.** *La divergencia de Kullback-Leibler cumple las siguientes propiedades siempre que existan las probabilidades condicionales regulares correspondientes y las derivadas de Radon-Nikodym respectivas.*

1. $D(P_{Y\,|\,X} \,||\, Q_{Y\,|\,X} \,|\, P_X) = D(P_X \otimes P_{Y\,|\,X} \,||\, P_X \otimes Q_{Y\,|\,X})$.
2. $D(P_{XY} \,||\, Q_{XY}) = D(P_{Y\,|\,X} \,||\, Q_{Y\,|\,X} \,|\, P_X) + D(P_X \,||Q_X)$.
3. $D(P_{XY} \,||\, Q_{XY}) \geq D(P_Y \,||\, Q_Y)$.

4. *Sean $P_{Y\,|\,X}$ y $Q_{Y\,|\,X}$ dos kernel de transición. Tómense $P_Y = P_{Y\,|\,X} \circ P_X$ y $Q_Y = Q_{Y\,|\,X} \circ P_X$. Se cumple que*

$$D(P_Y \,||\, Q_Y) \leq D(P_{Y\,|\,X} \,||\, Q_{Y\,|\,X} \,|\, P_X),$$

con igualdad si y sólo si $D(P_{X\,|\,Y} \,||\, Q_{X\,|\,Y} \,|\, P_Y) = 0$.

5. *Sea $P_{Y\,|\,X}$ un kernel de transición. Tómense $P_Y = P_{Y\,|\,X} \circ P_X$ y $Q_Y = P_{Y\,|\,X} \circ Q_X$. Entonces*

$$D(P_Y \,||Q_Y) \leq D(P_X \,||Q_X).$$

Demostración.

1. En virtud de la Proposición 2.5,

$$\begin{aligned} D(P_X \otimes P_{Y\,|\,X} \,||\, P_X \otimes Q_{Y\,|\,X}) &= \int \log \frac{dP_X \otimes P_{Y\,|\,X}}{dP_X \otimes Q_{Y\,|\,X}} \, dP_X \otimes P_{Y\,|\,X} \\ &= \int \log \frac{P_{Y\,|\,X}}{Q_{Y\,|\,X}} \, dP_X \otimes P_{Y\,|\,X} \\ &\overset{(2.31)}{=} D(P_{Y\,|\,X} \,||\, Q_{Y\,|\,X} \,|\, P_X). \end{aligned}$$

2. Apelando a la Proposición 2.5, se tiene

$$\begin{aligned} D(P_{XY} \,||\, Q_{XY}) &= \int \log \frac{dP_X \otimes P_{Y\,|\,X}}{dQ_X \otimes Q_{Y\,|\,X}} \, dP_X \otimes P_{Y\,|\,X} \\ &= \int \log \frac{dP_X}{dQ_X} dP_X \otimes P_{Y\,|\,X} + \int \log \frac{dP_{Y\,|\,X}}{dQ_{Y\,|\,X}} dP_X \otimes P_{Y\,|\,X} \\ &= D(P_X \,||\, Q_X) + D(P_{Y\,|\,X} \,||\, Q_{Y\,|\,X} \,|\, P_X). \end{aligned}$$

3. Siguiendo el procedimiento del inciso anterior, se deduce la igualdad análoga

$$D(P_{XY} \,||\, Q_{XY}) = D(P_Y \,||\, Q_Y) + D(P_{X\,|\,Y} \,||\, Q_{X\,|\,Y} \,|\, P_Y). \tag{2.35}$$

Luego, la desigualdad de interés se sigue de la positividad de la divergencia.

4. De la Proposición 2.6.2 y de (2.35) se tiene

$$\begin{aligned} D(P_{XY} \,||\, Q_{XY}) &= D(P_{Y\,|\,X} \,||\, Q_{Y\,|\,X} \,|\, P_X) + D(P_X \,||\, P_X) \\ &= D(P_{X\,|\,Y} \,||\, Q_{X\,|\,Y} \,|\, P_Y) + D(P_Y \,||\, Q_Y). \end{aligned}$$

Dado que $D(P_X \,||\, P_X) = 0$, entonces

$$D(P_{Y\,|\,X} \,||\, Q_{Y\,|\,X} \,|\, P_X) \geq D(P_Y \,||\, Q_Y),$$

con igualdad si y solamente si $D(P_{X\,|\,Y} \,||\, Q_{X\,|\,Y} \,|\, P_Y) = 0$.

5. Debido a los resultados en 2.6.2 y 2.6.3 se tiene

$$\begin{aligned} D(P_Y \,||Q_Y) &\leq D(P_{XY} \,||Q_{XY}) \\ &= D(P_{Y\,|\,X} \,||\, P_{Y\,|\,X} \,|\, P_X) + D(P_X \,||Q_X) \\ &= D(P_X \,||\, Q_X). \end{aligned}$$

□

Como ya se mencionó anteriormente, con la ayuda de la divergencia de K-L, se busca desarrollar un concepto análogo a la entropía de Shannon para todo tipo de variables aleatorias. Lamentablemente, éste resulta ser exclusivamente útil cuando la variable en cuestión es absolutamente continua con respecto a la medida de Lebesgue.

Definición 2.11 (Polyanskiy y Wu 2019, Definición 1.5)**.** La entropía diferencial de un vector aleatorio continuo X está dada por

$$h(X) \coloneqq -D(P_X \,||\, \mathrm{Leb}). \tag{2.36}$$

Concretamente, si X tiene función de densidad de probabilidades p_X, se tiene que $h(X) = \mathbb{E}\left(\log \frac{1}{p_X(X)}\right)$; en cualquier otro caso $h(X) = -\infty$.

Es importante mencionar que la entropía diferencial no es una generalización de la entropía de Shannon y no comparte las mismas propiedades. Por ejemplo, h puede tomar valores positivos, negativos o en $\{-\infty, \infty\}$. Para una lista ilustrativa de diferencias se recomienda revisar la advertencia en la página 21 de Polyanskiy y Wu (2019). No obstante, la idéntica relación que éstas comparten con la información mutua (Proposición 2.8) le otorga relevancia a la entropía diferencial dentro de este texto.

Ejemplo 2.8 (Gaussiano)**.** Si $X \sim N_d(\mu, \Sigma)$ con $\mu \in \mathbb{R}^d$ y $\det \Sigma \neq 0$, entonces $h(X) = \frac{1}{2}\log(\det(2\pi e \Sigma))$.

Demostración. Dado que $\det(\Sigma) \neq 0$, existe la función de densidad de probabilidades

$$p_X(\mathbf{x}) = \frac{\exp(-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu))}{((2\pi)^d |\Sigma|)^{\frac{1}{2}}}. \tag{2.37}$$

Luego,

$$h(X) = \frac{1}{2}\log((2\pi)^d|\Sigma|) + \frac{1}{2}\mathbb{E}\left((X-\mu)^\top \Sigma^{-1}(X-\mu)\right). \tag{2.38}$$

En virtud del Teorema 5.1 en Seber y Lee (2003), se deduce

$$\mathbb{E}\left((X-\mu)^\top \Sigma^{-1}(X-\mu)\right) = \operatorname{tr}(\Sigma^{-1}\Sigma) = \operatorname{tr}(I_d) = d. \tag{2.39}$$

Sustituyendo (2.39) en (2.38), se obtiene

$$\begin{aligned} h(X^n) &= \frac{1}{2}\log((2\pi)^d|\Sigma|) + \frac{1}{2}d \\ &= \frac{1}{2}(\log(\det(2\pi e\Sigma)). \end{aligned}$$

□

Hasta el momento se ha evitado ahondar en algunos problemas técnicos que acarrea trabajar con kernels de transición. La siguiente subsección es fundamental para mantener un nivel de formalidad adecuado en este trabajo. Ésta se basa en el Apartado 2.6 de Polyanskiy y Wu (2019).

2.3.4. Algunos problemas de medibilidad

La Definición 2.8 conlleva una interrogante que quizás haya molestado a los lectores más rigurosos. Concretamente, para poder calcular la divergencia condicional en el caso donde $P_{Y\,|\,X}(\cdot\,|\,x) \ll Q_{Y\,|\,X}(\cdot\,|\,x)$ para todo $x \in \mathcal{X}$ es necesario que

$$\frac{dP_{Y\,|\,X}(\cdot\,|\,x)}{dQ_{Y\,|\,X}(\cdot\,|\,x)}(y)$$

sea una función medible en $\mathcal{X}\times\mathcal{Y}$. Esto queda asegurado en virtud de la versión de Doob del Teorema de Radon-Nikodym (Çinlar 2011, Teorema 4.44) siempre que se trabaje con una σ-álgebra separable de $\mathcal{Y}$. Sin embargo, para deducir algunas propiedades de la información mutua se vuelve necesario laborar bajo el siguiente acuerdo; que también soluciona este problema:

Acuerdo 1 (Polyanskiy y Wu 2019, Acuerdo A2)**.** Todas las distribuciones conjuntas P_{XY} están especificadas por medio de datos. Es decir, existen medidas σ-finitas μ_1 y μ_2 en $\mathcal{X}$ y $\mathcal{Y}$, respectivamente, y una función medible $\lambda(x,y)$ en $\mathcal{X}\times\mathcal{Y}$ tales que

$$P_{XY}(C) \coloneqq \int_C \lambda(x,y)\mu_1(dx)\mu_2(dy). \tag{2.40}$$

De lo anterior es posible desintegrar y obtener las marginales y condicionales dadas por

$$P_{Y\,|\,X}(A\,|\,x) = \int_A \rho_{Y\,|\,X}(y\,|\,x)\mu_2(dy). \tag{2.41}$$

$$P_X(B) = \int_B p_X(x)\mu_1(dx). \tag{2.42}$$

donde $\rho_{Y\,|\,X}(y\,|\,x) \coloneqq \frac{\lambda(x,y)}{p(x)}$ y $p_X(x) = \int_{\mathcal{Y}} \lambda(x,y)\mu_2(dy)$. Nótese que dadas dos medidas P_{XY} y Q_{XY} caracterizadas de esta manera, apelando al teorema de Radon-Nikodym, se puede asumir que están definidas bajo una misma medida dominante. Más aún, se tiene que si

$$P_{Y\,|\,X}(A\,|\,x) = \int_A \rho_{Y\,|\,X}(y\,|\,x)\mu_2(dy), \quad Q_{Y\,|\,X}(A\,|\,x) = \int_A \rho'_{Y\,|\,X}(y\,|\,x)\mu_2(dy)$$

y $P_{Y\,|\,X}(\cdot\,|\,x) \ll P_{Y\,|\,X}(\cdot\,|\,x)$, entonces $\frac{\rho_{Y\,|\,X}(y\,|\,x)}{\rho'_{Y\,|\,X}(y\,|\,x)}$ es una versión de la derivada de Radon-Nikodym. Por lo tanto, ésta es medible en $\mathcal{X} \times \mathcal{Y}$.

Bajo el Acuerdo 1, es posible asegurar que las distribuciones marginales y condicionales se comportan de manera análoga al caso discreto. Aunque se pierde generalidad, esto sirve para probar la Identidad de Kolmogorov, la cual es utilizada en las demostraciones del Capítulo 5.

2.4. Información mutua

Una vez contando con una pseudométrica para medidas de probabilidad como la divergencia de Kullback-Leibler, es natural utilizar ésta para cuantificar la dependencia entre variables aleatorias al comparar la distribución conjunta con la distribución correspondiente al supuesto de independencia. En particular, debido a las propiedades de la divergencia, ésta cantidad es cero si y sólo si las variables son independientes por lo que, en principio, es una herramienta ideal para este propósito.

2.4.1. Definición

Definición 2.12 (Polyanskiy y Wu 2019, Definición 2.3)**.** Sean X y Y dos variables aleatorias con distribución conjunta P_{XY} y distribuciones marginales P_X y P_Y. Se define la información mutua entre X y Y como

$$I(X;Y) \coloneqq D(P_{XY} \,||\, P_X \otimes P_Y), \tag{2.43}$$

donde $P_X \otimes P_Y$ denota a la medida de producto de P_X y P_Y.

La siguiente proposición enuncia algunas características de la información mutua.

Proposición 2.7 (Polyanskiy y Wu 2019, Teorema 2.3)**.** *La información mutua posee las siguientes propiedades:*

1. $I(X;Y) = D(P_{XY} \,||\, P_X \otimes P_Y) = D(P_{Y\,|\,X} \,||\, P_Y \,|\, P_X) = D(P_{X\,|\,Y} \,||\, P_X \,|\, P_Y)$.
2. *$I(X;Y) = I(Y;X)$.*
3. $I(X;Y) \geq 0$, *con igualdad si y sólo si* $X \perp\!\!\!\perp Y$.

Demostración.

1. Si existen las probabilidades condicionales regulares $P_{Y\,|\,X}$ y $P_{X\,|\,Y}$ tales que $P_{XY} = P_X \otimes P_{Y\,|\,X} = P_X \otimes P_{X\,|\,Y}$, entonces la Definición 2.12 implica

$$I(X;Y) = D(P_Y \otimes P_{X\,|\,Y} \,||\, P_X \otimes P_Y) = D(P_X \otimes P_{Y\,|\,X} \,||\, P_X \otimes P_Y). \quad (2.44)$$

   Aplicar la Proposición 2.6.1 a (2.44) concluye la prueba.

2. Sean $\bar{X}$ y $\bar{Y}$ copias independientes de X y Y e independientes entre sí. Defínase $\phi : (x, y) \mapsto (y, x)$. Así, si $Z \coloneqq \phi(X, Y)$ y $\bar{Z} \coloneqq \phi(\bar{X}, \bar{Y})$, entonces $P_Z = k \circ P_{XY}$, $P_{\bar{Z}} = k \circ P_{\bar{X}\bar{Y}}$, $P_{XY} = k \circ P_Z$ y $P_{\bar{X}\bar{Y}} = k \circ P_{\bar{Z}}$, donde $k(A \,|\, x, y) = \mathbb{1}_A(y, x)$. Nótese que k es un kernel de transición de $\mathcal{X} \times \mathcal{Y}$ a $\mathcal{Y} \times \mathcal{X}$. Así, en virtud de la Proposición 2.6.5 (desigualdad de procesamiento de la información para la divergencia), se conluye que

$$D(P_Z \,||\, P_{\bar{Z}}) \leq I(X;Y) = D(P_{XY} \,||\, P_{\bar{X}\bar{Y}}) \leq D(P_Z \,||\, P_{\bar{Z}}). \quad (2.45)$$

3. El resultado se sigue directamente de la Definición 2.12 y la Proposición 2.4.

□

Para el caso discreto y el absolutamente continuo, la información mutua tiene una representación intuitiva con respecto a la entropía y la entropía diferencial respectivamente.

Proposición 2.8 (Polyanskiy y Wu 2019, Teorema 2.4)**.**

1. *Si X y Y son ambas variables aleatorias discretas*

$$I(X;Y) = H(X) - H(X \,|\, Y).$$

2. *Si X y Y son ambas variables aleatorias discretas*

$$I(X;Y) = H(X) + H(Y) - H(X, Y)$$

3. *Si X y Y son vectores aleatorios con densidad conjunta y las tres entropías diferenciales son finitas, entonces*

$$I(X;Y) = h(X) + h(Y) - h(X, Y).$$

Más aún, si X tiene densidad marginal y existe la densidad condicional $p_{X\,|\,Y}$ entonces

$$I(X;Y) = h(X) - h(X\,|\,Y).$$

Demostración.

1. Si $(X,Y) \sim P_{XY}$, se tiene que

$$\begin{aligned} I(X;Y) &= \mathbb{E}\left(\log \frac{P_Y(Y)\cdot P_{X\,|\,Y}(X\,|\,Y)}{P_X(X)\cdot P_Y(Y)}\right) \\ &= \mathbb{E}\left(\log \frac{1}{P_X(X)}\right) - E\left(\log \frac{1}{P_{X\,|\,Y}(X\,|\,Y)}\right) \\ &= H(X) - H(X\,|\,Y). \end{aligned}$$

2. Nuevamente, si $(X,Y) \sim P_{XY}$, entonces

$$\begin{aligned} I(X;Y) &= \mathbb{E}\left(\log \frac{P_{XY}(X,Y)}{P_X(X)\cdot P_Y(Y)}\right) \\ &= \mathbb{E}\left(\log \frac{1}{P_X(X)}\right) + \mathbb{E}\left(\log \frac{1}{P_Y(Y)}\right) - E\left(\log \frac{1}{P_{XY}(X,Y)}\right) \\ &= H(X) + H(Y) - H(X,Y). \end{aligned}$$

3. Reemplazando las funciones de masa de probabilidades por las funciones de densidad respectivas en las demostraciones anteriores prueba este caso.

□

Ejemplo 2.9 (Gaussiano)**.** Sean $X \in \mathbb{R}^m$ y $Y \in \mathbb{R}^n$ vectores conjuntamente gaussianos. Denotemos por $[X,Y] \in \mathbb{R}^{n+m}$ a la concatenación de X y Y. Si las matrices de correlación Σ_X, Σ_Y y $\Sigma_{[X,Y]}$ son invertibles, entonces

$$I(X;Y) = \frac{1}{2}\log\left(\frac{\det\Sigma_X \det\Sigma_Y}{\det\Sigma_{[X,Y]}}\right)$$

Demostración. Se sabe que para un vector aleatorio gaussiano $[X,Y] \in \mathbb{R}^{n+m}$, donde $X \in \mathbb{R}^m$ y $Y \in \mathbb{R}^m$, los subvectores X y Y son también vectores gaussianos, es decir, $X \sim N_m(\mu_X, \Sigma_X)$ y $Y \sim N_n(\mu_Y, \Sigma_Y)$; donde $\mathbb{E}([X,Y]) = [\mu_X, \mu_Y]$. Debido a que Σ_X, Σ_Y y $\Sigma_{[X,Y]}$ son invertibles, entonces existen las funciones de densidad

respectivas. Luego, en virtud del Ejemplo 2.8 y la Proposición 2.8.3,

$$\begin{aligned} I(X;Y) &= \frac{1}{2}\left(\log[(2\pi e)^m \det \Sigma_X] + \log[(2\pi e)^n \det \Sigma_Y] - \log[(2\pi e)^{n+m} \det \Sigma_{[X,Y]}]\right) \\ &= \frac{1}{2}\log\left(\frac{\det \Sigma_X \det \Sigma_Y}{\det \Sigma_{[X,Y]}}\right). \end{aligned}$$

□

2.4.2. Información Mutua Condicional

Definición 2.13. Dadas tres variables aleatorias X, Y y Z, se dice que éstas forman una cadena de Markov, denotado por $X - Y - Z$, si y sólo si $X \perp\!\!\!\perp Z \,|\, Y$. Si existen las probabilidades condicionales regulares correspondientes, se tiene que $X - Y - Z$ si y sólo si

$$P_{ZX\,|\,Y}(\cdot\,|\,y) = P_{Z\,|\,X}(\cdot\,|\,y) \otimes P_{X\,|\,Y}(\cdot\,|\,y), \;\; \forall y \in \mathcal{Y}.$$

Definición 2.14. Sean X, Y y Z variables aleatorias con distribución conjunta P_{XYZ} y condicionales $P_{XY\,|\,Z}$, $P_{X\,|\,Z}$ y $P_{Y\,|\,Z}$. Se define la información mutua condicional como

$$I(X;Y\,|\,Z) \coloneqq \int I(X;Y\,|\,Z=z) dP_Z(dz), \tag{2.46}$$

donde $I(X;Y\,|\,Z=z) \coloneqq D(P_{XY\,|\,Z}(\cdot\,|\,z)\,||\,P_{X\,|\,Z}(\cdot\,|\,z) \otimes P_{Y\,|\,Z}(\cdot\,|\,z))$.

Proposición 2.9 (Polyanskiy y Wu 2019, Teorema 2.5)**.** *Se tienen las siguientes propiedades de la información mutua y su versión condicional:*

1. $I(X;Z\,|\,Y) \geq 0$, *con igualdad si y sólo si* $X - Y - Z$.
2. *(Identidad de Kolmogorov)*

$$\begin{aligned} I(X,Y;Z) &= I(X;Z) + I(Y;Z\,|\,X) \\ &= I(Y;Z) + I(X;Z\,|\,Y). \end{aligned}$$

3. *(Desigualdad de procesamiento de la información) Si* $X - Y - Z$, *entonces*
   a) $I(X;Z) \leq I(X;Y)$, *con igualdad si y sólo si* $X - Z - Y$.
   b) $I(X;Y\,|\,Z) \leq I(X;Y)$.

Demostración.

1. De la Proposición 2.4 se asegura que $I_Z(X;Y \mid Z)$ es una variable aleatoria no-negativa. Así, $I(X;Y \mid Z) \geq 0$. Además, $I(X;Y \mid Z) = 0$ si y sólo si $I_Z(X;Y \mid Z) = 0$ casi seguramente. Finalmente, la Proposición 2.4 también implica que $I_Z(X;Y \mid Z) = 0$ casi seguramente si y sólo si $P_{XY \mid Z} = P_{X \mid Z} \otimes P_{Y \mid Z}$.

2. Ajustando el Acuerdo 1 al contexto de tres variables aleatorias, es directo deducir que $P_{Y \mid XZ} = \frac{P_{XYZ}}{P_{XZ}} = \frac{P_{YZ \mid X}}{P_{Z \mid X}}$. Así,

$$\log \frac{P_{XYZ}}{P_{XY}P_Z} = \log \frac{P_{Y \mid XZ} \cdot P_{XZ}}{P_{Y \mid X} \cdot (P_X \cdot P_Z)} = \log \frac{P_{XZ}}{P_X P_Z} + \log \frac{P_{YZ \mid X}}{P_{Y \mid X} P_{Z \mid X}}. \tag{2.47}$$

3. Aplicando la Identidad de Kolmogorov, se obtiene

$$I(Y,Z;X) = I(X;Y) + I(X;Z \mid Y) = I(X;Z) + I(X;Y \mid Z). \tag{2.48}$$

   Dado que $X - Y - Z$, (2.48) y la Proposición 2.9.1 implica que

$$I(X;Y) = I(X;Z) + I(X;Y \mid Z). \tag{2.49}$$

   En virtud de que $I(X;Z), I(X;Y \mid Z) \geq 0$, se concluye que $I(X;Z) \leq I(X;Y)$ y $(X;Y \mid Z) \leq I(X;Y)$. Finalmente, la igualdad en (2.49) asegura que $I(X;Z) = I(X;Y)$ si y sólo si $I(X;Y \mid Z) = 0$; es decir, apelando a la Proposición 2.9.1, si y sólo si $X - Z - Y$.

□

2.5. Desigualdades fuertes de procesamiento de la información

En la Proposición 2.6.6 se presentó la Desigualdad de Procesamiento de la Información para la Divergencia de Kullbak-Leilbler. Intuitivamente, esta propiedad de la información mutua nos dice que es imposible procesar la variable Y para extraer más información respecto a X. Como se menciona en Wang y col. (2021), bajo ciertas condiciones es posible mejorar dicha desigualdad. De allí surgen las desigualdades fuertes de procesamiento de la información; las cuales, usualmente, están relacionadas con un coeficiente de contracción.

Para estudiar estas desigualdades es necesario generalizar el concepto de la divergencia de Kullback-Leibler. La definición siguiente tiene este objetivo.

Definición 2.15. Sean $f : (0, \infty) \to \mathbb{R}$ una función convexa que es estrictamente convexa en 1 con $f(1) = 0$ y P, Q dos distribuciones de probabilidad sobre el conjunto $\mathcal{X} \subseteq \mathbb{R}^d$ tales que $P \ll Q$. La f-divergencia entre P y Q se define como

$$D_f(P \,||\, Q) \coloneqq \int_{\mathcal{X}} f\left(\frac{dP}{dQ}\right) dQ. \tag{2.50}$$

Algunos ejemplos de f-divergencias populares son la divergencia de Kullback-Leibler (D_{KL}) donde $f(x) = x \log x$ (Definición 2.6) y la distancia de variación total (TV) donde $f(x) = \frac{|x-1|}{2}$. La siguiente proposición nos brinda otra representación de la variación total. Esta caracterización será utilizada para calcular explícitamente la distancia de variación total entre dos variables aleatorias gaussianas cuya varianza es un múltiplo de la matriz identidad.

Proposición 2.10. *Sean P y Q dos medidas de probabilidad en $\mathcal{X}$ con $P \ll Q$, entonces*

$$TV(P, Q) = \sup_{E \subset \mathcal{X}} [P(E) - Q(E)]. \tag{2.51}$$

Demostración. Claramente existe una medida de probabilidad μ tal que $P \ll \mu$ y $Q \ll \mu$; por ejemplo, $\mu = \frac{1}{2}(P + Q)$. Apelando al Teorema de Radon-Nikodym (Kallenberg 2002, Teorema 2.10), denótese $p \coloneqq \frac{dP}{d\mu}$ y $q \coloneqq \frac{dQ}{d\mu}$. Luego, en virtud de la Proposición 2.3,

$$\mathrm{TV}(P, Q) = \frac{1}{2} \int \left| \frac{dP}{dQ} - 1 \right| dQ = \frac{1}{2} \int \left| \frac{dP}{dQ} \cdot q - q \right| d\mu. \tag{2.52}$$

Nótese que p es única $\mu - c.s.$, Además, para todo $A \subset \mathcal{X}$

$$\int_A \frac{dP}{dQ} \cdot q \, d\mu = \int_A \frac{dP}{dQ} d(q \cdot \mu) = \int_A \frac{dP}{dQ} dQ = \int_A dP.$$

De tal suerte que $p = \frac{dP}{dQ} \cdot q$ μ-casi-seguramente. Así, de (2.52) se sigue

$$\mathrm{TV}(P, Q) = \frac{1}{2} \int |p - q| \, d\mu. \tag{2.53}$$

Por otra parte, para cualquier $E \subset \mathcal{X}$,

$$P(E) + P(E^{\mathsf{c}}) = 1 = Q(E) + Q(E^{\mathsf{c}}).$$

Así,

$$P(E) - Q(E) = Q(E^{\mathsf{c}}) - P(E^{\mathsf{c}}). \tag{2.54}$$

De tal suerte que

$$\begin{aligned} P(E) - Q(E) &= \frac{1}{2}\left(\int_E (p-q)d\mu + \int_{E^c}(q-p)d\mu\right) \\ &\leq \frac{1}{2}\left(\int_E |p-q|d\mu + \int_{E^c}|q-p|d\mu\right) \\ &= \frac{1}{2}\int |p-q|d\mu. \end{aligned}$$

Debido a que la desigualdad anterior es válida para todo $E \subset X$, entonces de (2.53) se deduce

$$\sup_{E\subset\mathcal{X}}[P(E) - Q(E)] \leq TV(P,Q). \tag{2.55}$$

Por otra parte, defínase $E_0 = \{x \in \mathcal{X} : p(x) \geq q(x)\}$, de tal suerte que al aplicar, de nuevo, (2.54) se tiene

$$TV(P,Q) = \frac{1}{2}\left[\int_{E_0}(p-q)d\mu + \int_{E_0^c}(q-p)d\mu\right] \tag{2.56}$$

$$= P(E_0) - Q(E_0). \tag{2.57}$$

Por lo tanto,

$$\mathrm{TV}(P,Q) \leq \sup_{E\subset\mathcal{X}}[P(E) - Q(E)]. \tag{2.58}$$

De (2.55) y (2.58) se concluye lo deseado. □

Ejemplo 2.10.

$$\mathrm{TV}(N_d(\mu_1,\sigma^2 I_d), N_d(\mu_2,\sigma^2 I_d)) = 1 - 2\bar{\Phi}\left(\frac{\|\mu_1-\mu_2\|_2}{2\sigma}\right), \tag{2.59}$$

donde $\bar{\Phi}$ es la función de supervivencia de una variable normal estándar univariada.

Demostración. Con base en (2.57), se tiene

$$\mathrm{TV}(N_d(\mu_1,\sigma^2\mathbf{I}_d), N_d(\mu_2,\sigma^2\mathbf{I}_d)) = \mathbb{P}_1(E_0) - \mathbb{P}_2(E_0), \tag{2.60}$$

donde $\mathbb{P}_1$ y $\mathbb{P}_2$ son las medidas imagen de las variables gaussianas en cuestión, $E_0 = \{\mathbf{x} \in \mathbb{R}^d : f_1(\mathbf{x}) \geq f_2(\mathbf{x})\}$ y f_1, f_2 denotan las funciones de densidad de probabilidades respectivas. Defínase $\phi : \mathbb{R}^d \to \mathbb{R}$ como

$$\phi(\mathbf{x}) = 2(\mu_1-\mu_2)^\top\mathbf{x} + \|\mu_2\|_2^2 - \|\mu_1\|_2^2.$$

Posteriormente, nótese que $E_0 = \{\mathbf{x} : \phi(\mathbf{x}) \geq 0\}$. Así, en virtud del Teorema 2.2 y el Ejemplo 2.5 en Seber y Lee (2003), si X_1 tiene distribución $N_d(\mu_1, \sigma^2\mathbf{I}_d)$, entonces $\phi(X_1) \sim N_1(\|\mu_1 - \mu_2\|_2^2, 4\sigma^2\|\mu_1 - \mu_2\|_2^2)$. De tal suerte que

$$\begin{aligned}\mathbb{P}_1(E_0) &= \mathbb{P}(\phi(X_1) \geq 0) \\ &= \mathbb{P}\left(\frac{\phi(X_1) - \|\mu_1 - \mu_2\|_2^2}{2\sigma\|\mu_1 - \mu_2\|_2} \geq -\frac{\|\mu_1 - \mu_2\|_2}{2\sigma}\right) \\ &= 1 - \bar{\Phi}\left(\frac{\|\mu_1 - \mu_2\|_2}{2\sigma}\right).\end{aligned}$$

De manera análoga, si $X_2 \sim N_d(\mu_2, \sigma^2\mathbf{I}_d)$, entonces $\phi(X_2)$ tiene distribución dada por $N_1(-\|\mu_1 - \mu_2\|^2, 4\sigma^2\|\mu_1 - \mu_2\|_2^2)$. Por lo que

$$\begin{aligned}\mathbb{P}_2(E_0) &= \mathbb{P}\left(\frac{\phi(X_1) + \|\mu_1 - \mu_2\|_2^2}{2\sigma\|\mu_1 - \mu_2\|_2} \geq \frac{\|\mu_1 - \mu_2\|_2}{2\sigma}\right) \\ &= \bar{\Phi}\left(\frac{\|\mu_1 - \mu_2\|_2}{2\sigma}\right).\end{aligned}$$

De lo anterior y (2.60) se deduce (2.59). □

Usualmente, las desigualdades de procesamiento de la información están ligadas a un coeficiente de contracción. Si éste último se encuentra en $(0, 1)$, se obtiene una desigualdad fuerte de procesamiento de la información. Motivado en lo anterior, se tiene la siguiente definición.

Definición 2.16. Para un kernel de transición $P_{Y\,|\,X}$, se define el coeficiente de contracción de $P_{Y\,|\,X}$ para D_f como

$$\eta_f(P_{Y\,|\,X}) \coloneqq \sup_{\substack{P,Q\in\mathcal{P}(\mathcal{X}):\\ 0<D_f(P\,||\,Q)<\infty}} \frac{D_f(P_{Y\,|\,X}\circ P\,||\,P_{Y\,|\,X}\circ Q)}{D_f(P\,||\,Q)}. \tag{2.61}$$

En general, el supremo se toma sobre un conjunto no-vacío para toda f-divergencia (una prueba puede encontrarse en el Apéndice A de Polyanskiy y Wu (2017)). Además, es importante notar que la desigualdad de procesamiento de la información para la divegencia de Kullback-Leibler se generaliza para toda f-divergencia (véase Polyanskiy y Wu 2019, Teorema 6.2); es decir,

$$D_f(P_{Y\,|\,X}\circ P\,||\,P_{Y\,|\,X}\circ Q) \leq D_f(P\,||\,Q),\ \ \forall P, Q \in \mathcal{P}(\mathcal{X}). \tag{2.62}$$

De tal manera que se puede asegurar la existencia de η_f para toda f que cumpla con la Definición 2.15. Con base en esto, se denotará por $\eta_{\mathrm{KL}}(P_{Y\,|\,X})$ y $\eta_{\mathrm{TV}}(P_{Y\,|\,X})$ al coeficiente de contracción del kernel $P_{Y\,|\,X}$ para la divergencia de K-L y la variación total respectivamente.

Ejemplo 2.11. Si $Y = X + m \cdot N$ donde X es una variable que toma valores en un conjunto acotado $\mathcal{X} \subset \mathbb{R}^d$ y N es un vector gaussiano estándar e independiente del resto de las variables, entonces

$$\eta_{\mathrm{TV}}(P_{Y\,|\,X}) = 1 - 2\bar{\Phi}\left(\frac{\operatorname{diam}(\mathcal{X})}{2m}\right), \tag{2.63}$$

donde $\operatorname{diam}(\mathcal{X}) = \sup_{x,x' \in \mathcal{X}} \|x' - x\|_2$.

Demostración. En virtud de (Raginsky 2016, Teorema III.2)

$$\eta_{\mathrm{TV}}(P_{Y\,|\,X}) = \sup_{x,x' \in \mathcal{X}} \mathrm{TV}(P_{Y\,|\,X}(\cdot\,|\,x), P_{Y\,|\,X}(\cdot\,|\,x')). \tag{2.64}$$

Nótese que $P_{Y\,|\,X}(\cdot\,|\,x)$ es la distribución de una variable aleatoria $N_d(x, m\mathbf{I}_d)$. Luego, aplicando el Ejemplo 2.10 se obtiene que

$$\eta_{\mathrm{TV}}(P_{Y\,|\,X}) = \sup_{x,x' \in \mathcal{X}} \left(1 - 2\bar{\Phi}\left(\frac{\|x - x'\|_2}{2m}\right)\right). \tag{2.65}$$

La ecuación (2.63) se obtiene al notar que $\bar{\Phi}$ es una función decreciente. □

La siguiente proposición muestra que η_{TV}, conocido como el coeficiente de Dobrushin, acota al resto de coeficientes de contracción. La prueba se puede encontrar en Cohen, Kempermann y Zbaganu (1998).

Proposición 2.11 (Cohen, Kempermann y Zbaganu 1998, Proposición II.4.10)**.** *Para toda f-divergencia*

$$\eta_f(P_{Y\,|\,X}) \leq \eta_{TV}(P_{Y\,|\,X}). \tag{2.66}$$

Por otra parte, η_{KL} está ligado fuertemente a la información mutua. En particular, se cuenta con el siguiente resultado cuya demostración puede revisarse en el Apéndice B de Polyanskiy y Wu (2016). Esta conexión resulta esencial en el desarrollo del resultado principal de Wang y col. (2021) y explica la necesidad de estudiar las desigualdades fuertes de procesamiento de la información en este trabajo.

Proposición 2.12. *El coeficiente de contracción para la divergencia de K-L cumple*

$$\eta_{KL}(P_{Y\,|\,X}) = \sup_{\substack{U,X:\\ U-X-Y\\ 0<I(U,X)<\infty}} \frac{I(U;Y)}{I(U;X)}. \tag{2.67}$$

Así, para toda cadena de Markov $U - X - Y$,

$$I(U;Y) \leq \eta_{KL}(P_{Y\,|\,X}) \cdot I(U;X). \tag{2.68}$$

Discusión

Comprender las motivaciones e interpretaciones detrás de la entropía, la divergencia de Kullback-Leibler y, particularmente, la información mutua es uno de los efectos esperados de este capítulo en el lector. Como puede notarse, muchos de estos conceptos tienen una naturaleza práctica. Sin embargo, esto dificulta la tarea de generalizar los resultados a espacios de probabilidad arbitrarios. En este trabajo se ha puesto especial empeño en probar las proposiciones con un grado de generalidad alto. No obstante, esto lleva a preguntas profundamente teóricas relacionadas con la medibilidad de funciones y conjuntos que requieren mucho más tiempo para responder que el que se dedicó en la Subsección 2.3.4. Para más información respecto a este último punto, uno puede consultar la sección *How to avoid measurability problems?* en Polyanskiy y Wu (2019).

La teoría de la información presenta un conjunto de herramientas ideales para estudiar conceptos provenientes de otras áreas de las matemáticas y las matemáticas aplicadas. Particularmente, se logró un avance importante en este capítulo al deducir ecuaciones y desigualdades relacionadas con la divergencia de Kullback-Leibler y la información mutua. Adicionalmente, debido a que calcular, o estimar, la información mutua entre dos variables puede ser un trabajo bastante complicado, la última sección es la clave para deducir el Lema 2 en Wang y col. (2021), el cual nos permite, eventualmente, acotar esta cantidad por una más sencilla de estimar.

A pesar de lo anterior, aún quedan varios conceptos por estudiar para poder plantear la teoría detrás de un problema de aprendizaje máquina supervisado y deducir los resultados de Wang y col. (2021). Concretamente, es fundamental tener una noción básica de la optimización convexa para comenzar a estudiar problemas de aprendizaje automático ya que éstos, esencialmente, son problemas de minimización. Asimismo, se necesita obtener una herramienta para poder acotar definitivamente la información mutua por un factor más sencillo de estimar. En nuestro caso, se recurrirá a la teoría de transporte óptimo para lograr este objetivo.

Capítulo 3

Preliminares de Optimización Convexa y Transporte Óptimo

Hasta el momento se han introducido las ideas relacionadas con la teoría de la información necesarias para alcanzar los objetivos de este trabajo. Ahora, debido a la fuerte conexión que existe entre la optimización convexa y el aprendizaje automático, es necesario proveer algunas nociones básicas de esta primera con el propósito de mantener las pruebas y explicaciones presentadas en este texto tan auto-contenidas como sea posible. Asimismo, resulta imprescindible introducir una herramienta que nos permita acotar la información mutua cuando se trabaja con ruido gaussiano; como es el caso del descenso por gradiente de la dinámica de Langevin.

El primer objetivo de este capítulo es presentar una introducción sucinta a la optimización convexa que facilite la presentación de ciertos conceptos en el siguiente capítulo. El segundo objetivo es acotar superiormente la divergencia de Kullback-Leibler con ayuda de la distancia de Wasserstein cuando se trabaja con dos variables con ruido Gaussiano. Para lograr lo anterior, en la segunda sección se definirán las distancias de Wasserstein y se probará una desigualdad que conecta la divergencia con una de éstas. Dicha desigualdad juega un papel fundamental en las demostraciones de los resultados más importantes que se revisan en este trabajo.

Estructuralmente, la primera mitad revisa algunas ideas básicas de la optimización convexa: funciones convexas y un algoritmo de optimización para éstas. La segunda mitad introduce el concepto de distancia de Wasserstein basándose, principalmente, en el Capítulo 6 de Villani (2009). La última parte de este capítulo consiste en enunciar y probar el Lema 3.2; el cual es una extensión al caso multivariado del

Lema 3.4.2 en Ranginsky y Sason (2012).

3.1. Optimización convexa

Uno de los problemas más comunes de encontrar en la vida cotidiana es optimizar: minimizar los gastos del mes, maximizar la ingestión de vitaminas en una dieta, reducir el tiempo de traslado a la universidad, entre muchos otros. Sin embargo, como lo menciona Nesterov (2004), la mayoría de las veces el planteamiento teórico es mucho más sencillo que el proceso necesario para obtener una solución. Más aún, en casi todos los contextos, los problemas de optimización no tienen una solución (Nesterov 2004, p. xv). Particularmente, el aprendizaje estadístico supervisado se reduce a seleccionar un predictor que minimice el riesgo de equivocarse. No obstante, uno de los principales obstáculos para resolver el asunto es dar con el predictor que cumple dicha condición.

A pesar de lo anterior, se cuenta con una familia de funciones bien comportadas en ambientes de optimización: las funciones convexas. En esta sección se presenta una teoría básica con el objetivo de responder algunas preguntas importantes que el lector no familiarizado con este tema puede tener mientras se estudia la teoría del aprendizaje automático supervisado presentada en el Capítulo 4. Para pronfundizar más en el tema se recomienda consultar Nesterov (2004) y Hazan (2016).

3.1.1. Conjuntos y funciones convexas

Para entender mejor aquello que caracteriza a la funciones convexas, es de ayuda comprender el papel que juegan los conjuntos convexos en la definición de éstas. Debido a esto, se enuncia la siguiente definición.

Definición 3.1. Se dice que un conjunto $K \subset \mathbb{R}^d$ es convexo si para todo $\mathbf{x}, \mathbf{y} \in K$ se tiene

$$\alpha\mathbf{x} + (1-\alpha)\mathbf{y} \in K, \ \ \forall \alpha \in [0,1].$$

En virtud de la Definición 3.1 se dice que una función $f : \mathbb{R}^d \to \mathbb{R}$ es convexa si su epigrafo es un conjunto convexo de $\mathbb{R}^{d+1}$ (véase Rockafellar 1997, p.23). Sin embargo, la siguiente definición resulta más apropiada para nuestro contexto y ambas son equivalentes de acuerdo con el Teorema 4.1 en Rockafellar (1997).

Definición 3.2. Sea $K \subset \mathbb{R}^d$. Se dice que una función $f : K \to \mathbb{R}$ es convexa si para cualesquiera $\mathbf{x}, \mathbf{y} \in K$

$$f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}), \ \ \forall \alpha \in [0,1].$$

En particular, la siguiente proposición nos brinda una útil caracterización de las funciones convexas cuando éstas son diferenciables.

Proposición 3.1. *Sean $K \subset \mathbb{R}^d$, $d > 1$, convexo y $f : K \to \mathbb{R}$ diferenciable, entonces f es convexa si y sólo si para todo $\boldsymbol{x}, \boldsymbol{y} \in K$*

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle.$$

Demostración. Primero, supóngase que $f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$ para toda $\mathbf{x}, \mathbf{y} \in K$. Así, se tiene que

$$\begin{aligned} f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle &\leq f(\mathbf{y}), \\ f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle &\leq f(\mathbf{x}). \end{aligned}$$

Sumar las desigualdades anteriores resulta en

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq 0. \tag{3.1}$$

Luego, tómense $0 \leq \alpha < \beta \leq 1$ y defínanse $\mathbf{y}_\alpha = \mathbf{y} + \alpha(\mathbf{x} - \mathbf{y})$, $\mathbf{y}_\beta = \mathbf{y} + \beta(\mathbf{x} - \mathbf{y})$ y $\phi : [0, 1] \to \mathbb{R}$ como

$$\phi(\alpha) = f(\mathbf{y} + \alpha(\mathbf{x} - \mathbf{y})). \tag{3.2}$$

Nótese que ϕ está bien definida para cualesquiera $\mathbf{x}, \mathbf{y} \in K$ pues K es convexo. Posteriormente, la regla de la cadena para campos escalares (Apostol 2014, Teorema 8.8) asegura que ϕ es diferenciable y que

$$\phi'(\alpha) = \langle \nabla f(\mathbf{y} + \alpha(\mathbf{x} - \mathbf{y})), \mathbf{x} - \mathbf{y} \rangle.$$

Operaciones aritméticas elementales resultan en

$$\phi'(\beta) - \phi'(\alpha) = \frac{1}{\beta - \alpha} \left\langle \nabla f(\mathbf{y}_\beta) - f(\mathbf{y}_\alpha), \mathbf{y}_\beta - \mathbf{y}_\alpha \right\rangle \overset{(3.1)}{\geq} 0. \tag{3.3}$$

Así, ϕ' es no-decreciente y, por ello, ϕ es convexa (véase Spivak 2014, Teorema 2). En particular, si $\alpha \in [0, 1]$,

$$\begin{aligned} f(\alpha\mathbf{x} + (1 - \alpha)\mathbf{y}) &= \phi(\alpha) = \phi(\alpha + (1 - \alpha)\, 0) \\ &\leq \alpha\phi(1) + (1 - \alpha)\phi(0) \\ &= \alpha f(\mathbf{x}) + (1 - \alpha) f(\mathbf{y}). \end{aligned}$$

De tal suerte que f es convexa según la Definición 3.2. Por otra parte, supóngase $f(\lambda\mathbf{x} + (1 - \lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{y})$ para toda $\lambda \in [0, 1]$ y $\mathbf{x}, \mathbf{y} \in K$. Es bien sabido que

$$\langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle = f'(\mathbf{x}; \mathbf{y} - \mathbf{x}) = \lim_{h \to 0} \frac{f(\mathbf{x} + h(\mathbf{y} - \mathbf{x})) - f(\mathbf{x})}{h}. \tag{3.4}$$

Además, en virtud de que f y K son convexos, para cualesquiera $\mathbf{x}, \mathbf{y} \in K$, la función $\psi : h \mapsto \frac{f(\mathbf{x}+h(\mathbf{y}-\mathbf{x}))-f(\mathbf{x})}{h}$ es no-decreciente y está bien definida en $(0, 1]$. Así, de la ecuación (3.4) se concluye que para toda $\mathbf{x}, \mathbf{y} \in K$,

$$\langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leq \psi(1) = f(\mathbf{y}) - f(\mathbf{x}).$$

Por lo tanto,

$$f(\mathbf{y}) \geq \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + f(\mathbf{x}), \quad \forall \mathbf{x}, \mathbf{y} \in K.$$

□

La Proposición 3.1 provee una manera de minimizar funciones convexas y diferenciables en varias variables de una manera análoga a como se enseña en cursos básicos de cálculo diferencial en $\mathbb{R}$. Esto se ve reflejado en la siguiente afirmación.

Proposición 3.2 (Nesterov 2004, Teorema 2.1.1)**.** *Sea $K \subset \mathbb{R}^d$ un conjunto convexo y sea $f : K \to \mathbb{R}$ una función convexa y diferenciable. Si $\nabla f(\boldsymbol{x}^*) = 0$ para algún $\boldsymbol{x}^* \in K$, entonces $\boldsymbol{x}^* \in \arg\min_{\boldsymbol{x} \in K} f(\boldsymbol{x})$.*

Demostración. En virtud de la Proposición 3.1, para todo $\mathbf{x} \in K$

$$f(\mathbf{x}) \geq f(\mathbf{x}^*) + \langle \nabla f(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle = f(\mathbf{x}^*).$$

□

Es decir, si se tiene una función convexa y diferenciable que va de un conjunto convexo de $\mathbb{R}^d$ a $\mathbb{R}$, encontrar las raíces del gradiente es equivalente a minimizar la función sobre todo su dominio. A raíz de esto, se presenta un algoritmo con el objetivo de optimizar funciones convexas y diferenciables.

3.1.2. Descenso por Gradiente

No está de más mencionar que encontrar las raíces de una función arbitraria es imposible en casi cualquier situación. En virtud de lo anterior, lo más común es recurrir a métodos numéricos para resolver el problema de optimización (¡ya sabemos que la solución existe cuando se trabaja con funciones convexas!). El descenso por gradiente (*GD*) es un algoritmo bastante utilizado en estas situaciones. Debido a esto, conocerlo es fundamental para entender el porqué se introducen nuevos algoritmos en la Sección 4.3. Con este propósito en mente, se describe a grandes rasgos el contenido del apartado 14.1 en Shalev-Schwarz y Ben-David (2014).

Supóngase que se busca minimizar una función convexa y diferenciable $f : \mathbb{R}^d \to \mathbb{R}$. El descenso por gradiente es un algoritmo iterativo, que se inicializa con un punto

$\mathbf{w}_0$, donde se busca encontrar el mínimo de f al seguir la dirección contraria a la que apunta el gradiente mediante el siguiente procedimiento:

Para $t = 0, 1, \ldots, T-1$ realizar:
- $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla f(\mathbf{w}_t),\ \eta > 0.$

Finalmente, regresar $\bar{\mathbf{w}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{w}_t$.

En general, el algoritmo puede regresar $\mathbf{w}_T$ o el vector con mejor desempeño, es decir, $\arg\min_{t \in [T]} f(\mathbf{w}_t)$; donde $[T] \coloneqq \{1, 2, \times T\}$. Sin embargo, como se plantea en Shalev-Schwarz y Ben-David (2014), regresar el promedio aritmético resulta congruente con el planteamiento del caso estocástico. De acuerdo con (Shalev-Schwarz y Ben-David (2014), p. 151), la intuición detrás de este procedimiento es que el gradiente dirige hacia el punto donde se encuentra la mayor tasa incremental de cambio de f alrededor de $\mathbf{w}_t$. Luego, se sigue la dirección contraria para disminuir el valor de la función. A continuación, se da una definición esencial para obtener algunas garantías respecto a la convergencia del algoritmo en un contexto particular.

Definición 3.3 (Shalev-Schwarz y Ben-David 2014, Definición 12.6)**.** Una función $f : \mathbb{R}^d \to \mathbb{R}^k$ es ρ-Lipschitz si para todo $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^d$ se cumple

$$\|f(\mathbf{w}_1) - f(\mathbf{w}_2)\| \leq \|\mathbf{w}_1 - \mathbf{w}_2\|. \tag{3.5}$$

En el caso de funciones convexas y ρ-Lipschitz, se puede garantizar la convergencia del algoritmo al mínimo de f. Estas garantías se presentan con ayuda de la siguiente proposición.

Proposición 3.3 (Shalev-Schwarz y Ben-David 2014, Lemma 14.1)**.** *Sean $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_T$ una secuencia arbitraria de vectores. Cualquier algoritmo con inicialización $\boldsymbol{w}_1 = 0$ y regla de actualización*

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \boldsymbol{v}_t \tag{3.6}$$

satisface que

$$\sum_{t=1}^{T} \langle \boldsymbol{w}_t - \boldsymbol{w}^*, \boldsymbol{v}_t \rangle \leq \frac{\|\boldsymbol{w}^*\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|\boldsymbol{v}_t\|^2.$$

En particular, para todo $B, \rho > 0$, si para todo t se satisface $\|\boldsymbol{v}_t\| \leq \rho$ y si se escoge $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$, entonces para toda $\boldsymbol{w}^*$ con $|\boldsymbol{w}^*\| \leq B$ se tiene

$$\frac{1}{T} \sum_{t=1}^{T} \langle \boldsymbol{w}_t - \boldsymbol{w}^*, \boldsymbol{v}_t \rangle \leq \frac{B\rho}{\sqrt{T}}.$$

Demostración. Nótese que

$$\|\mathbf{w}_t - \mathbf{w}^* - \eta\mathbf{v}_t\|^2 = \|\mathbf{w}_t - \mathbf{w}^*\|^2 - 2\langle \mathbf{w}_t - \mathbf{w}^*, \eta\mathbf{v}_t\rangle + \eta^2\|\mathbf{v}_t\|^2. \tag{3.7}$$

Luego, (3.7) implica,

$$\begin{aligned}\langle \mathbf{w}_t - \mathbf{w}^*, \mathbf{v}_t\rangle &= \frac{1}{2\eta}(-\|\mathbf{w}_t - \mathbf{w}^* - \eta\mathbf{v}_t\|^2 + \|\mathbf{w}_t - \mathbf{w}^*\|^2 + \eta^2\|\mathbf{v}_t\|^2) \\ &= \frac{1}{2\eta}\left(-\|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2 + \|\mathbf{w}_t - \mathbf{w}^*\|^2\right) + \frac{\eta}{2}\|\mathbf{v}_t\|^2,\end{aligned}$$

donde la última igualdad se debe a (3.6). Sumando la última igualdad sobre toda $t \in [T]$ implica que

$$\sum_{t=1}^{T}\langle \mathbf{w}_t - \mathbf{w}^*, \mathbf{v}_t\rangle = \frac{1}{2\eta}\sum_{t=1}^{T}\left(-\|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2 + \|\mathbf{w}_t - \mathbf{w}^*\|^2\right) + \frac{\eta}{2}\sum_{t=1}^{T}\|\mathbf{v}_t\|^2. \tag{3.8}$$

Dado que $\sum_{t=1}^{T}(-\|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2 + \|\mathbf{w}_t - \mathbf{w}^*\|^2)$ es una suma telescópica y, además, se ha escogido $\mathbf{w}_1 = 0$, entonces

$$\sum_{t=1}^{T}\left(-\|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2 + \|\mathbf{w}_t - \mathbf{w}^*\|^2\right) = \|\mathbf{w}^*\|^2 - \|\mathbf{w}_{T+1} - \mathbf{w}^*\|^2. \tag{3.9}$$

Sustituyendo (3.9) en (3.8) es posible concluir que

$$\sum_{t=1}^{T}\langle \mathbf{w}_t - \mathbf{w}^*, \mathbf{v}_t\rangle \leq \frac{1}{2\eta}\|\mathbf{w}^*\|^2 + \frac{\eta}{2}\sum_{t=1}^{T}\|\mathbf{v}_t\|^2. \tag{3.10}$$

Adicionalmente, si $\mathbf{w}^* \leq B$, $\|\mathbf{v}_t\| \leq \rho$ y se toma $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$, entonces

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\langle \mathbf{w}_t - \mathbf{w}^*, \mathbf{v}_t\rangle &\leq \frac{1}{2T}\left(\frac{B^2}{\eta} + T\eta\rho^2\right) \\ &= \frac{1}{2T}\left(B\rho\sqrt{T} + B\rho\sqrt{T}\right) \\ &= \frac{B\rho}{\sqrt{T}}.\end{aligned}$$

□

Si f es diferenciable y ρ-Lipschitz, se tiene que $\nabla f(\mathbf{w}) \leq \rho$ (véase, por ejemplo, el Lema 14.7 en Shalev-Schwarz y Ben-David (2014)). Así, se puede aplicar la Proposición 3.3 con $\mathbf{v}_t = \nabla f(\mathbf{w}_t)$. Concretamente, para el caso convexo y diferenciable se obtiene el siguiente corolario que nos habla de la convergencia del *GD*.

Corolario 3.1 (Shalev-Schwarz y Ben-David 2014, Corolario 14.2)**.** *Sea f una función convexa, diferenciable y ρ-Lipschitz. Tómese $\boldsymbol{w}^* \in \arg\min_{\boldsymbol{w}:\|\boldsymbol{w}\|\leq B} f(\boldsymbol{w})$. Si se corre el algoritmo GD durante T iteraciones con $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$, entonces el vector resultante $\bar{\boldsymbol{w}}$ satisface*

$$f(\bar{\boldsymbol{w}}) - f(\boldsymbol{w}^*) \leq \frac{B\rho}{\sqrt{T}}.$$

Demostración. Recordando que $\bar{w} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}_t$, al aplicar la desigualdad de Jensen se obtiene

$$f(\bar{\mathbf{w}}) - f(\mathbf{w}^*) \leq \frac{1}{T}\sum_{t=1}^{T} \left(f(\mathbf{w}_t) - f(\mathbf{w}^*)\right). \tag{3.11}$$

Luego, como f es convexa, entonces $f(\mathbf{w}_t) - f(\mathbf{w}^*) \leq \langle \mathbf{w}_t - \mathbf{w}^*, \nabla f(\mathbf{w}_t)\rangle$ para todo $t \in [T]$. Así,

$$\frac{1}{T}\sum_{t=1}^{T} f(\mathbf{w}_t) - f(\mathbf{w}^*) \leq \frac{1}{T}\sum_{t=1}^{T} \langle \mathbf{w}_t - \mathbf{w}^*, \nabla f(\mathbf{w}_t)\rangle. \tag{3.12}$$

Por otra parte, apelando a la Proposición 3.3 con $\mathbf{v}_t = \nabla f(\mathbf{w}_t)$, se deduce

$$\frac{1}{T}\sum_{t=1}^{T} \langle \mathbf{w}_t - \mathbf{w}^*, \nabla f(\mathbf{w}_t)\rangle \leq \frac{B\rho}{\sqrt{T}}. \tag{3.13}$$

Combinando las desigualdades (3.11)-(3.13) se concluye lo deseado. □

Nótese que la teoría presentada para el *GD* se limita a cuando la función a analizar tiene como dominio a $\mathbb{R}^d$. Sin embargo, la primera parte de este capítulo nos sugiere que este método debe funcionar aún cuando $\operatorname{dom} f \subset \mathbb{R}^d$. Para trabajar con algoritmos en este nuevo nivel de abstracción es necesario apoyarse de la siguiente definición y el subsecuente lema.

Definición 3.4. Para un conjunto $K \subseteq \mathbb{R}^d$ cerrado y un vector $\mathbf{x} \in \mathbb{R}^d$ se define la proyección de $\mathbf{x}$ sobre K como

$$\operatorname{Proj}_K(\mathbf{x}) \coloneqq \arg\min_{\mathbf{y}\in K} \|\mathbf{x} - \mathbf{y}\|_2.$$

Lema 3.1 (Lema de Proyección, Shalev-Schwarz y Ben-David 2014, Lema 14.9)**.** *Sea $K \subseteq \mathbb{R}^d$ un conjunto cerrado y convexo. Tómese $\boldsymbol{v} \coloneqq Proj_K(\boldsymbol{w})$ para algún $\boldsymbol{w} \in \mathbb{R}^d$, entonces*

$$\|\boldsymbol{w} - \boldsymbol{x}\| \geq \|\boldsymbol{v} - \boldsymbol{x}\|, \ \ \forall \boldsymbol{x} \in K.$$

Demostración. Dado que K es convexo, entonces para $\alpha \in (0,1)$ y $\mathbf{x} \in K$, $\mathbf{v} - \alpha(\mathbf{x} - \mathbf{v}) \in K$. Luego, de la definición de $\mathbf{v}$ se sigue

$$\begin{aligned}\|\mathbf{v} - \mathbf{w}\|^2 &\leq \|\mathbf{v} - \alpha(\mathbf{x} - \mathbf{v}) - \mathbf{w}\| \\ &= \|\mathbf{v} - \mathbf{w}\|^2 - 2\alpha\langle \mathbf{v} - \mathbf{w}, \mathbf{x} - \mathbf{v}\rangle + \alpha^2\|\mathbf{x} - \mathbf{v}\|^2.\end{aligned}$$

Así,

$$-\alpha\|\mathbf{x} - \mathbf{v}\|^2 \leq 2\langle \mathbf{v} - \mathbf{w}, \mathbf{x} - \mathbf{v}\rangle. \tag{3.14}$$

Tomando el límite cuando $\alpha \to 0$ en (3.14) se sigue

$$\langle \mathbf{v} - \mathbf{w}, \mathbf{x} - \mathbf{v}\rangle \geq 0. \tag{3.15}$$

De tal forma que

$$\begin{aligned}\|\mathbf{w} - \mathbf{x}\| &= \|\mathbf{w} - \mathbf{v}\| + 2\langle \mathbf{v} - \mathbf{w}, \mathbf{x} - \mathbf{v}\rangle + \|\mathbf{v} - \mathbf{x}\| \\ &\geq \|\mathbf{v} - \mathbf{x}\|.\end{aligned}$$

□

Con ayuda del Lema 3.1, es posible extender el resultado del Corolario 3.1 a funciones cuyo dominio es un subconjunto convexo y cerrado. Este procedimiento resultará más claro cuando se visite la Sección 4.3 donde se prueba un resultado similar para un algoritmo que involucra un paso de proyección.

A pesar de que se ha profundizado muy poco en la teoría de optimización convexa, el autor espera que aquellos lectores poco familiarizados con el tema hayan comprendido mejor la caracterización de las funciones convexas y las propiedades que las vuelven tan útiles en contextos de optimización; así como el papel que juegan los métodos numéricos en estos problemas. Para leer una revisión minuciosa acerca de la optimización convexa se puede consultar Nesterov (2004) y para estudiar una evolución de ésta, con uso extendido en problemas contemporáneos, Hazan (2016) es una perfecta introducción. De esta manera, el enfoque de este capítulo cambia a la teoría del transporte óptimo. Con ayuda de ésta, se busca superar la limitante de calcular/estimar la información mutua para acotar el error de generalización.

3.2. Transporte óptimo

Supóngase que se es el encargado de entregar los productos que compran consumidores en una tienda en línea que cuenta con diversos almacenes. La dirección de un comprador y la ubicación del producto que adquirió están modeladas en el espacio tridimensional por las medidas de probabilidad μ y ν respectivamente. Es

intuitivo pensar que en cuanto más lejos se encuentren los bienes de la ubicación del adquisidor, más costoso deberá ser realizar la entrega.

La situación anterior se puede modelar teóricamente de la siguiente manera. Primero, defínase $\mathcal{P}(\mathcal{X})$ como el espacio de medidas de probabilidad cuyo dominio es una σ-álgebra de $\mathcal{X}$. Sean μ y ν distribuciones de probabilidad en $\mathcal{P}(\mathcal{X})$ y $\mathcal{P}(\mathcal{Y})$ respectivamente. Se define el espacio $\Pi(\mu, \nu)$ como la colección de todas las distribuciones conjuntas en $\mathcal{X} \times \mathcal{Y}$ con marginales μ y ν. Así, una manera de cuantificar el costo esperado de una entrega es el **costo de transporte óptimo** entre dos medidas; el cual está dado por:

$$C(\mu, \nu) \coloneqq \inf_{\pi \in \Pi(\mu,\nu)} \int c(x, y) d\pi(x, y), \tag{3.16}$$

donde $c(x, y)$ representa el costo de transportar una unidad de masa del punto x al punto y.

3.2.1. Distancias de Wasserstein

Como lo menciona Villani (2009), puede parecer natural pensar en C como una distancia entre la medidas de probabilidad μ y ν. Sin embargo, C no necesariamente posee las propiedades características de una distancia. No obstante, es sencillo asegurar que C cumpla con éstas si se escoge c adecuadamente. Con base en esto, se da la siguiente definición y la subsecuente proposición.

Definición 3.5. Sea $(\mathcal{X}, d)$ un espacio métrico Polaco y sea $p \in [1, \infty)$. Para cualesquiera dos medidas de probabilidad $\mu, \nu \in \mathcal{P}(\mathcal{X})$, se define la distancia de Wasserstein de orden p entre μ y ν como

$$W_p(\mu, \nu) \coloneqq \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}} d(x, y)^p d\pi(x, y) \right)^{1/p}. \tag{3.17}$$

Proposición 3.4. *La distancia de Wassserstein de orden p tiene las características de una métrica; i.e.,*

1. *Para cualesquiera* $\mu, \nu \in \mathcal{P}(\mathcal{X})$

$$W_p(\mu, \nu) = W_p(\nu, \mu).$$

2. *Si* $\mu_1, \mu_2, \mu_3 \in \mathcal{P}(\mathcal{X})$*, entonces*

$$W_p(\mu_1, \mu_3) \leq W_p(\mu_1, \mu_2) + W_p(\mu_2, \mu_3).$$

3. *Para* $\mu, \nu \in \mathcal{P}(\mathcal{X})$ *se tiene que* $W_p(\mu, \nu) \geq 0$ *con igualdad si y sólo si* $\mu = \nu$.

Demostración.

1. Dado que d es una métrica, entonces

$$\begin{aligned}\inf_{\pi\in\Pi(\mu,\nu)}\int_{\mathcal{X}} d(x,y)^p d\pi(x,y) &= \inf_{\pi\in\Pi(\mu,\nu)}\int_{\mathcal{X}} d(y,x)^p d\pi(x,y)\\ &= \inf_{\pi\in\Pi(\nu,\mu)}\int_{\mathcal{X}} d(x,y)^p d\pi(x,y)\\ &= W_p(\nu,\mu).\end{aligned}$$

2. Sean $\mu_1,\mu_2,\mu_3 \in \mathcal{P}(\mathcal{X})$. Tómese a (X_1,X_2) como un acoplamiento óptimo de (μ_1,μ_2) y a (Y_2,Y_3) como un acoplamiento óptimo de (μ_2,μ_3) con respecto a la función de costo d^p. En virtud del Lema 1.1.10 en Dudley (1999), existe un vector aleatorio (X_1',X_2',X_3') tal que $(X_1,X_2)\overset{d}{=}(X_1',X_2')$ y $(Y_2,Y_2)\overset{d}{=}(X_2',X_3')$. Nótese que, particularmente, (X_1',X_3') es un acoplamiento de (μ_1,μ_3), de tal forma que al apelar a la desigualdad del triángulo se obtiene

$$W_p(\mu_1,\mu_3) \leq \left(\mathbb{E}\left[d(X_1',X_3')^p\right]\right)^{\frac{1}{p}} \leq \left(\mathbb{E}\left[\left(d(X_1',X_2')+d(X_2',X_3')\right)^p\right]\right)^{\frac{1}{p}}. \quad (3.18)$$

Luego, debido a la desigualdad de Minkowski en L^p (véase, por ejemplo, el Lema 1.29 en Kallenberg (2002))

$$\begin{aligned}&\left(\mathbb{E}\left[\left(d(X_1',X_2')+d(X_2',X_3')\right)^p\right]\right)^{\frac{1}{p}}\\ &\qquad\leq \left(\mathbb{E}\left[d(X_1',X_2')^p\right]\right)^{\frac{1}{p}} + \left(\mathbb{E}\left[d(X_2',X_3')^p\right]\right)^{\frac{1}{p}}. \quad (3.19)\end{aligned}$$

Así, al juntar (3.18)-(3.19) y recordar que (X_1',X_2') y (X_2',X_3') son acoplamientos óptimos se concluye

$$\begin{aligned}W_p(\mu_1,\mu_3) &\leq \left(\mathbb{E}\left[d(X_1',X_2')^p\right]\right)^{\frac{1}{p}} + \left(\mathbb{E}\left[d(X_2',X_3')^p\right]\right)^{\frac{1}{p}}\\ &= W_p(\mu_1,\mu_2)+W_p(\mu_2,\mu_3).\end{aligned}$$

3. Debido a la monotonicidad de la integral, es claro que W_p es no-negativa. Luego, supóngase que $W_p(\mu,\nu)=0$, entonces existe una medida de probabilidad $\pi\in\Pi(\mu,\nu)$ concentrada completamente en $\{(x,x) : x\in\mathcal{X}\}$. Sin embargo, lo anterior sólo puede ocurrir si $\mu=\nu$.

□

Al haber definido las distancias de Wasserstein, estamos preparados para lograr el objetivo principal de la segunda mitad de este capítulo: enunciar y probar el Lema 3.2. La siguiente subsección se enfoca completamente en cumplir lo anterior.

3.2.2. Distancia de Wasserstein de segundo orden y la divergencia de Kullback-Leibler

Al intentar acotar el error de generalización con ayuda de la información mutua, es indispensable acotar esta última. El siguiente lema nos permite encontrar una cota para la divergencia de K-L con ayuda de la distancia de Wasserstein. Dada la estrecha relación entre la divergencia y la información mutua, este resultado es ideal para nuestros intereses.

Lema 3.2. *Sea (X, Y) un par de vectores aleatorios que toman valores en $\mathbb{R}^d$ y tómese $N \sim N_d(\boldsymbol{0}, \boldsymbol{I}_d)$ independiente de (X, Y). Para todo $t > 0$ se cumple*

$$D\left(P_{X+\sqrt{t}N} \,||\, P_{Y+\sqrt{t}N}\right) \leq \frac{1}{2t} W_2^2(P_X, P_Y). \tag{3.20}$$

Demostración. De la Proposición 2.6.3 se tiene

$$D(P_{X,Y,X+\sqrt{t}N} \,||\, P_{X,Y,Y+\sqrt{t}N}) \geq D(P_{X+\sqrt{t}N}, P_{Y+\sqrt{t}N}). \tag{3.21}$$

Luego, de (2.27) se deduce

$$\begin{aligned} &D(P_{X,Y,X+\sqrt{t}N}, P_{X,Y,Y+\sqrt{t}N}) \\ &\qquad = \mathbb{E}\left[D(P_{X+\sqrt{t}N\,|\,X,Y}(\cdot\,|\,X,Y) \,||\, P_{Y+\sqrt{t}N\,|\,X,Y}(\cdot\,|\,X,Y))\right]. \end{aligned} \tag{3.22}$$

Es claro que $(X + \sqrt{t}N \,|\, X, Y) \sim N_d(X, t\,\mathbf{I}_d)$ y $(Y + \sqrt{t}N \,|\, X, Y) \sim N_d(Y, t\,\mathbf{I}_d)$. Así,

$$D(P_{X,Y,X+\sqrt{t}N}, P_{X,Y,Y+\sqrt{t}N}) = \mathbb{E}\left[D\left(N_d(X, t\,\mathbf{I}_d) \,||\, N_d(Y, t\,\mathbf{I}_d)\right)\right]. \tag{3.23}$$

En virtud del Ejemplo 2.5 se tiene

$$D\left(N_d(X, t^{-1}\mathbf{I}_d) \,||\, N_d(Y, t^{-1}\mathbf{I}_d)\right) = \frac{1}{2t}(X-Y)^\top(X-Y) = \frac{1}{2t}\|X-Y\|_2^2. \tag{3.24}$$

De tal manera que combinar (3.21), (3.23) y (3.24) resulta en

$$\frac{1}{2t}\mathbb{E}\left[\|X-Y\|_2^2\right] \geq D(P_{X+\sqrt{t}N}, P_{Y+\sqrt{t}N}). \tag{3.25}$$

Dado que la desigualdad anterior es válida independientemente de la distribución conjunta de (X, Y) y el lado derecho no depende de ésta, entonces

$$D(P_{X+\sqrt{t}N} \,||\, P_{Y+\sqrt{t}N}) \leq \frac{1}{2t} \inf_{\substack{\mu \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d): \\ (X,Y) \sim \mu}} \mathbb{E}\left[\|X-Y\|_2^2\right] = W_2^2(P_X, P_Y). \tag{3.26}$$

□

Discusión

Optimizar está profundamente ligado a la vida de seres racionales como nosotros, por lo que es vital preguntarse cómo resolver este tipo de problemas y, más importante aún, preguntarse cuáles de éstos tienen solución. Como se presentó en este capítulo, la optimización convexa intenta responder a estas preguntas. Es importante mencionar que la teoría presentada en la primera mitad de esta sección se puede extender a funciones convexas, sin importar si éstas son diferenciables, mediante el concepto de **sub-gradientes**. Sin embargo, este nivel de generalización no es necesario para estudiar el algoritmo que nos concierne en este trabajo: el descenso por gradiente de la dinámica de Langevin. Para ahondar en la teoría clásica de la optimización convexa, se refiere al lector a Nesterov (2004). Más aún, uno puede revisar Hazan (2016) para estudiar una vertiente más moderna de esta teoría: la optimizaicón convexa en línea.

En lo que corresponde a la segunda mitad de este capítulo, una simple consulta de Villani (2009) es suficiente para entender que la teoría de transporte óptimo es densa y extensa. Sin embargo, como se presentó en la Sección 3.2, tiene un planteamiento intuitivo y su alcance se entrelaza con diversas ramas de las matemáticas por lo que ésta se vuelve una herramienta para atacar distintos tipos de problemas matemáticos.

Una vez establecido el marco teórico para estudiar problemas de aprendizaje de máquina supervisado, la Sección 3.1 servirá para establecer una metodología para resolver este tipo de cuestiones. El segundo objetivo de este capítulo fue deducir el Lema 3.2. Éste permite, dentro del contexto establecido en Wang y col. (2021), acotar la información mutua superiormente y evitar tener que estimarla a partir de una muestra (el cual es un trabajo bastante complicado).

Así, nos encontramos en la situación adecuada para presentar una teoría orientada a plantear y resolver problemas de aprendizaje automático supervisado. Lo cual, a su vez, nos permitirá estudiar características de éstos; por ejemplo, como se hace en Wang y col. (2021), la capacidad de generalización de un algoritmo.

Capítulo 4

Preliminares de Aprendizaje Máquina

En los capítulos anteriores se revisaron algunos conceptos necesarios para el desarrollo de los resultados principales revisados en este texto. Este capítulo tiene el objetivo de proveer al lector con el resto de la teoría necesaria para deducir y comprender éstos.

Concretamente, este capítulo es una introducción al aprendizaje máquina estadístico y tiene el objetivo de brindar al lector las últimas herramientas necesarias para comprender tanto las demostraciones como la relevancia del trabajo presentado en Wang y col. (2021). Es importante mencionar que, dada la naturaleza práctica del aprendizaje máquina, las ideas presentadas en este capítulo son de gran ayuda para entender la importancia de encontrar cotas para el error de generalización esperado y las aportaciones de Wang y col. (2021) a esta iniciativa.

Estructuralmente, la primera sección del presente capítulo intenta motivar el estudio formal del aprendizaje automático. Dentro de la segunda sección se presentan los componentes de un problema de este tipo y algunas posibles soluciones a éste. Desafortunadamente, a pesar de la sencillez teórica de estas soluciones, su implementación en la práctica se convierte en todo un desafío. Debido a esto, en la última sección se presentan algoritmos que prometen ayudar a resolver problemas de aprendizaje estadístico en situaciones específicas (e.g. convexidad y diferenciabilidad).

4.1. Motivación

En muchos contextos científicos, económicos y sociales es usual encontrarse intentando predecir (clasificar) una variable no observada con ayuda de un conjunto de características observadas y (potencialmente) relacionadas con dicha cantidad de interés. Comúnmente, esta variable de interés es llamada **variable dependiente** y a dichas características se les conoce como las **variables independientes** del modelo.

Este tipo de situaciones normalmente se presentan cuando es tanto fundamental conocer el valor de la variable dependiente como es difícil observarla; por cuestiones de costos, de tiempo, de esfuerzo, etc. De tal manera que resulta mucho más conveniente, o como única opción, obtener los valores de las variables independientes e intentar predecir el valor de la cantidad de interés. Como puede deducirse de la descripción anterior, es posible estandarizar los componentes de un problema de aprendizaje automatizado como se presenta en la próxima sección.

4.2. Marco teórico de un problema de aprendizaje estadístico

En general, en todo problema de aprendizaje de máquina se cuenta con:

- Un **conjunto de características** $\mathcal{X}$ que contiene las cualidades observables de interés del fenómeno a estudiar.
- Un **espacio de etiquetas** $\mathcal{Y}$ con todas las posibles categorías a las que podemos asignar a una instancia.
- Una **distribución de probabilidad** μ sobre el espacio $\mathcal{Z} \coloneqq \mathcal{X} \times \mathcal{Y}$.
- Un **conjunto de datos de entrenamiento**, i.e. una muestra aleatoria proveniente de la distribución μ. En este trabajo, una muestra de tamaño n será denotada por S_n. Bajo estas convenciones, se denota

$$S_n = ((X_1, Y_1), \ldots, (X_n, Y_n)) = (Z_1, \ldots, Z_n).$$

- Una **clase de hipótesis**. Es decir, un subconjunto no-vacío de funciones de $\mathcal{X}$ a $\mathcal{Y}$ del cual elegir un **predictor/clasificador** h. El trabajo de h es etiquetar una instancia con base en sus características observadas. Nótese que si la clase de hipótesis es un conjunto de funciones parametrizadas, entonces ésta induce un espacio parametral $\mathcal{W}$. Debido a esto, será común utilizar $\mathcal{H}$ y $\mathcal{W}$ de manera intercambiable.

En la práctica, al contar con una muestra aleatoria observada de (X, Y), denotada por $s_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, se busca encontrar un predictor $h \in \mathcal{H}$ que aproxime *correctamente* a Y con base en la información que proporciona X. Es particularmente importante notar la vaguedad que conlleva la palabra *correctamente* en la oración anterior; esto se debe a que la manera de calificar el desempeño de un predictor queda en manos de quien estudia el fenómeno de interés. Es por esto que se vuelve necesario añadir un último elemento a la lista anterior:

- Una **función de pérdida** $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$ que cuantifique el error en el que incurre un clasificador en un punto de características y etiquetas $(x, y) \in \mathcal{Z}$.

Ejemplo 4.1. Supóngase que se trabaja con 8 imágenes en blanco y negro con una resolución de 15 píxeles donde cada una de éstas representa un dígito dentro del conjunto $\{1, 7\}$; las cuales se muestran en la Figura 4.1. En este caso, se puede plantear un problema de aprendizaje estadístico de la siguiente manera: sea $\mathcal{X}$ el conjunto de matrices de 5×3 con entradas en $\{0, 1\}$; i.e. $\mathcal{X} = \mathbb{M}_{5\times 3}(\{0, 1\})$ y sea $\mathcal{Y} = \{1, 7\}$. Finalmente, se ha decidido trabajar con la clase de hipótesis $\mathcal{H} = \{h_t : t \in \mathbb{R}\}$ donde

$$h_t(x) = \begin{cases} 1, & \sum_j x_j \leq t; \\ 7, & \sum_j x_j > t. \end{cases}$$

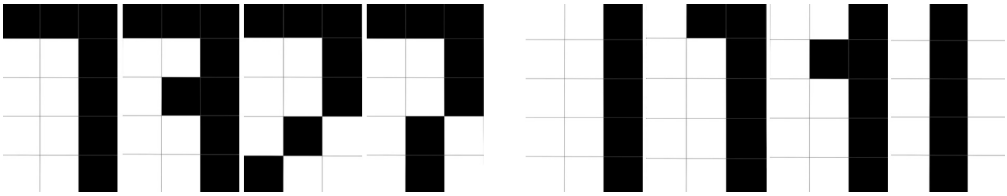

Figura 4.1: Conjunto de entrenamiento de 8 imágenes de 15 píxeles (S_8) que representan al dígito 7 (izquierda) y al dígito 1 (derecha).

Definición 4.1. En el contexto de un problema de aprendizaje automático, se define el riesgo real de un clasificador $h : \mathcal{X} \to \mathcal{Y}$ como

$$L_\mu(h) \coloneqq \mathbb{E}[l(h, (X, Y))], \tag{4.1}$$

donde el vector (X, Y) tiene distribución μ.

La Definición 4.1 parece proporcionarnos una manera sensata de seleccionar un clasificador de $\mathcal{H}$; simplemente, elegir una hipótesis que minimice el riesgo real. No obstante, es importante tener en mente que, usualmente, la distribución μ es desconocida, por lo que es imposible calcular $L_\mu(h)$ para cualquier $h \in \mathcal{H}$. Una posible solución a esto, es utilizar un estimador insesgado de L_μ que sí se pueda computar.

Definición 4.2. Dada una muestra $S_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, se define el riesgo empírico de un clasificador $h \in \mathcal{H}$ como

$$L_{S_n}(h) = \frac{1}{n} \sum_{i=1}^{n} l(h, (X_i, Y_i)). \tag{4.2}$$

Como se aprecia en la ecuación (4.2), dada una muestra aleatoria S_n, es directo calcular el riesgo empírico. Más aún, es fácil corroborar, apelando a la linealidad de la esperanza, que el riesgo empírico es un estimador insesgado del riesgo real. Debido a lo anterior, es una práctica común seleccionar el predictor con el menor riesgo empírico para la muestra dada.

Por otra parte, dado que el marco teórico detallado antes no establece cómo seleccionar un clasificador adecuado, otra estrategia podría ser tomar predictores con riesgo empírico *bajo* y seleccionar uno de éstos de acuerdo con alguna distribución de probabilidad. En general, cualquier manera, determinista o aleatoria, de seleccionar un clasificador con base en la muestra es una opción válida desde un punto de vista teórico. Esto se representa en la siguiente definición.

Definición 4.3. Se le llama *algoritmo de aprendizaje* a una variable aleatoria W que toma una muestra S_n y devuelve una hipótesis en $\mathcal{H}$ de acuerdo con una distribución $P_{W \mid S_n}$.

La próxima definición no es más que un ejemplo de un algoritmo de aprendizaje como una función determinista de la muestra; es decir, $P_{W \mid S_n}(\cdot \mid s_n)$ es la distribución de una variable aleatoria constante.

Definición 4.4. Dada una una función de pérdida ℓ y una clase de hipótesis $\mathcal{H}$, el algoritmo de minimización del riesgo empírico se define como

$$\mathrm{ERM}_{\mathcal{H}}(S_n) \in \arg\min_{h \in \mathcal{H}} L_{S_n}(h), \tag{4.3}$$

donde los empates se rompen de manera arbitraria y

$$\arg\min_{h \in \mathcal{H}} L_{S_n}(h) \coloneqq \{h \in \mathcal{H} : L_{S_n}(h) \leq L_{S_n}(h^*), \ \ \forall h^* \in \mathcal{H}\}$$

Ejemplo 4.2. Al seleccionar la función de pérdida 0-1, es decir, al usar $l(h, (x, y)) = \mathbb{1}_{h(x) \neq y}$ para trabajar con el Ejemplo 4.1, el $ERM_{\mathcal{H}}(S_8)$ está dado por cualquier h_t tal que $t \in [6, 7)$.

En pocas palabras, desde un punto de vista teórico, lo ideal es seleccionar al clasificador $h \in \mathcal{H}$ con el menor riesgo real posible. Sin embargo, la distribución μ raramente es conocida por lo que este método resulta inaplicable. No obstante, se

puede recurrir a minimizar un estimador insesgado del riesgo real, el cual hemos llamado riesgo empírico, y de esta manera obtener un predictor $\mathrm{ERM}_{\mathcal{H}}(S_n) \in \mathcal{H}$.

Al concentrarse en minimizar el riesgo empírico es importante no perder de vista que solamente el riesgo real permite asegurar, desde el punto de vista de la estadística clásica, que, en promedio y a largo plazo, se obtendrán predicciones con precisión acorde a éste. Dicho de otra manera, bajo este acercamiento $\mathrm{ERM}_{\mathcal{H}}(S_n)$ podría tener un desempeño excepcional en nuestra muestra observada pero producir resultados mediocres una vez implementado en la práctica. Motivado de la anterior, surge la siguiente definición.

Definición 4.5. Dado un problema de aprendizaje estadístico y un algoritmo de aprendizaje caracterizado por $P_{W \mid S_n}$, el error de generalización es la diferencia $L_\mu(W) - L_{S_n}(W)$ y su esperanza se denota como

$$\mathrm{gen}(\mu, P_{W \mid S_n}) \coloneqq \mathbb{E}\left[L_\mu(W) - L_{S_n}(W)\right], \tag{4.4}$$

donde la esperanza se toma con respecto a la medida de probabilidad $\mu^{\otimes n} \otimes P_{W \mid S_n}$.

4.3. Algoritmos para minimizar el riesgo en un problema de aprendizaje máquina

Ya se ha visto que una solución sensata a un problema de aprendizaje estadístico es minimizar $L_\mu(h)$; o en su defecto $L_{S_n}(h)$. No obstante, tal problema de optimización puede llegar a ser bastante complicado, por lo que resulta indispensable recurrir a métodos numéricos para resolverlo. En esta sección se presentan dos algoritmos de optimización adecuados, en sus respectivos contextos, para esta tarea.

Primeramente, es importante notar que, en virtud del Corolario 3.1, para una muestra observada s_n, es posible minimizar el riesgo empírico $L_{s_n}(h)$ (bajo las condiciones adecuadas) utilizando el método del descenso por gradiente (*GD*). Sin embargo, un objetivo más ambicioso es intentar minimizar $L_\mu(h)$, el riesgo real, mediante *GD*. Desafortunadamente, la distribución μ es desconocida por lo que resulta imposible calcular el gradiente de $L_\mu(h)$. No obstante, el siguiente algoritmo supera esta limitante al seguir una dirección aleatoria en cada paso; una dirección aleatoria cuyo valor esperado debe ser el inverso aditivo del gradiente.

4.3.1. Descenso por gradiente estocástico

Como se menciona en Shalev-Schwarz y Ben-David (2014), el descenso por gradiente estocástico (*SGD*) presenta una metodología para minimizar el riesgo real $L_\mu(h)$. Su planteamiento es análogo al del *GD* excepto que en cada paso se sigue una dirección aleatoria cuyo valor esperado es un vector que apunta al lado opuesto del gradiente de la función a minimizar.

En particular, en el contexto del aprendizaje máquina con una función de pérdida acotada y diferenciable l, dada una muestra aleatoria $S_n = (Z_1, \ldots, Z_n)$, el $SGD(\mathbf{w}_0, \eta, T)$ se implementa mediante

> Para $t = 0, 1, \ldots, T-1$, realizar
> 1. $\mathbf{v}_{t+1} = \mathbf{w}_t - \eta \nabla_{\mathbf{w}} l(\mathbf{w}_t, Z_{t+1})$,
> 2. $\mathbf{w}_{t+1} = \text{Proj}_{\mathcal{H}}(\mathbf{v}_{t+1})$
>
> Finalmente, regresar $\bar{\mathbf{w}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{w}_t$.

Nótese que si $l(\cdot, z)$ es diferenciable y acotada para todo $z \in \mathcal{Z}$, entonces el Teorema de Convergencia Dominada de Lebesgue y el Teorema del Valor Medio implican que

$$\mathbb{E}[\nabla_{\mathbf{w}} l(\mathbf{w}, Z_{t+1})] = \nabla_{\mathbf{w}} \mathbb{E}[l(\mathbf{w}, Z_{t+1})] = \nabla L_\mu(\mathbf{w}), \ \ \forall t \in [T]. \tag{4.5}$$

La siguiente proposición nos brinda algunas garantías acerca de la eficacia del *SGD*.

Proposición 4.1. *Sea $\mathcal{W} \subset \mathbb{R}^d$ cerrado y convexo. Si $L_\mu : \mathcal{W} \to \mathbb{R}$ es una función diferenciable y convexa, entonces $SGD(\boldsymbol{w}_0, \eta, T)$ satisface*

$$\mathbb{E}\left[L_\mu\left(\frac{1}{T}\sum_{t=0}^{T-1} \boldsymbol{w}_t\right)\right] - L_\mu(\boldsymbol{w}^*) \leq \frac{\|\boldsymbol{w}_0 - \boldsymbol{w}^*\|^2}{2\eta T} + \frac{\eta G^2}{2}, \tag{4.6}$$

donde $\boldsymbol{w}^* = argmin_{\boldsymbol{w}\in\mathcal{W}} L_\mu(\boldsymbol{w})$ y $G := \sup_{z\in\mathcal{Z}} \sup_{\boldsymbol{w}\in\mathcal{W}} \|\nabla l(\boldsymbol{w}, z)\|$.

Demostración. Debido a la convexidad de $\mathcal{W}$, $\frac{1}{T}\sum_{t=1}^{T} \mathbf{w}_t \in \mathcal{W}$. Luego, debido a la convexidad de L_μ y la desigualdad de Jensen, se tiene

$$L_\mu\left(\frac{1}{T}\sum_{t=0}^{T-1} \mathbf{w}_t\right) - L_\mu(\mathbf{w}^*) \leq \frac{1}{T}\sum_{t=0}^{T-1} \langle \nabla L_\mu(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}^* \rangle. \tag{4.7}$$

Al tomar la esperanza de ambos lados de la desigualdad en (4.7) se obtiene

$$\mathbb{E}\left[L_\mu\left(\frac{1}{T}\sum_{t=0}^{T-1} \mathbf{w}_t\right)\right] - L_\mu(\mathbf{w}^*) \leq \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\langle \nabla L_\mu(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}^* \rangle]. \tag{4.8}$$

Por otra parte, en virtud de (4.5),

$$\nabla L_\mu(\mathbf{w}_t) = \mathbb{E}[\nabla l(\mathbf{w}, Z_{t+1})]\Big|_{\mathbf{w}=\mathbf{w}_t}.$$

Luego, dado que $\mathbf{w}_t \perp\!\!\!\perp Z_{t+1}$, entonces, apelando a (Jacod y Protter 2004, ejercicio 23.7),

$$\nabla L_\mu(\mathbf{w}_t) = \mathbb{E}[\nabla l(\mathbf{w}_t, Z_{t+1}) \,|\, \mathbf{w}_t].$$

Así,

$$\mathbb{E}[\langle \nabla L_\mu(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}^* \rangle] = \mathbb{E}[\langle \mathbb{E}[\nabla l(\mathbf{w}_t, Z_{t+1}) \,|\, \mathbf{w}_t], \mathbf{w}_t - \mathbf{w}^* \rangle].$$

Es claro que $\mathbf{w}_t - \mathbf{w}^*$ es una función $\sigma(\mathbf{w}_t)$-medible. De tal manera que, al aplicar la linealidad de la esperanza condicional y la regla de esperanza total, se obtiene

$$\mathbb{E}[\langle \nabla L_\mu(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}^* \rangle] = \mathbb{E}[\nabla l(\mathbf{w}_t, Z_{t+1}), \mathbf{w}_t - \mathbf{w}^* \rangle]. \tag{4.9}$$

Sustituyendo (4.9) en (4.8)

$$\mathbb{E}\left[L_\mu\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathbf{w}_t\right)\right] - L_\mu(\mathbf{w}^*) \leq \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\nabla l(\mathbf{w}_t, Z_{t+1}), \mathbf{w}_t - \mathbf{w}^* \rangle]. \tag{4.10}$$

Adicionalmente, debido al Lema de Proyección (Lema 3.1)

$$\begin{aligned}\|\mathbf{w}^* - \mathbf{w}_{t+1}\|^2 &\leq \|\mathbf{w}^* - \mathbf{w}_t + \eta \nabla l(\mathbf{w}_t, Z_{t+1})\|^2 \\ &= \|\mathbf{w}^* - \mathbf{w}_t\|^2 + 2\eta\langle \nabla l(\mathbf{w}_t, Z_{t+1}), \mathbf{w}^* - \mathbf{w}_t \rangle + \eta^2 \|\nabla l(\mathbf{w}_t, Z_{t+1})\|^2.\end{aligned}$$

Al despejar $\langle \nabla l(\mathbf{w}_t, Z_{t+1}), \mathbf{w}^* - \mathbf{w}_t \rangle$ de la desigualdad anterior, se sigue

$$\begin{aligned}&\langle \nabla l(\mathbf{w}_t, Z_{t+1}), \mathbf{w}^* - \mathbf{w}_t \rangle \\ &\qquad \leq \frac{1}{2\eta}\left(\|\mathbf{w}_t - \mathbf{w}^*\|^2 - \|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2 + \eta^2 \|\nabla l(\mathbf{w}_t, Z_{t+1})\|^2\right).\end{aligned} \tag{4.11}$$

Más aún,

$$\mathbb{E}(\langle \nabla l(\mathbf{w}_t, Z_{t+1}), \mathbf{w}^* - \mathbf{w}_t \rangle) \leq \frac{1}{2\eta}\mathbb{E}\left(\|\mathbf{w}_t - \mathbf{w}^*\|^2 - \|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2\right) + \frac{\eta G^2}{2}. \tag{4.12}$$

Combinando (4.10) y (4.12) y simplificando la suma telescópica se sigue

$$\begin{aligned}\mathbb{E}\left[L_\mu\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathbf{w}_t\right)\right] - L_\mu(\mathbf{w}^*) &\leq \mathbb{E}\left[\frac{1}{2\eta T}\sum_{t=0}^{T-1}\left(\|\mathbf{w}_t - \mathbf{w}^*\|^2 - \|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2\right)\right] + \frac{\eta G^2}{2} \\ &= \frac{1}{2\eta T}\|\mathbf{w}_0 - \mathbf{w}^*\|^2 - \frac{1}{2\eta T}\mathbb{E}(\|\mathbf{w}_T - \mathbf{w}^*\|^2) + \frac{\eta G^2}{2}.\end{aligned}$$

Dado que $\|\mathbf{w}_T - \mathbf{w}^*\|^2 \geq 0$, se concluye la prueba del enunciado. □

El *SGD* puede ser de gran utilidad para optimizar funciones convexas. Sin embargo, en un contexto más general, el algoritmo puede quedarse atrapado en algún mínimo local. Una alternativa que busca solucionar esta cuestión es el descenso por gradiente de la dinámica de Langevin (*SGLD*). La siguiente subsección menciona las características de éste conforme se explican en Wang y col. (2021).

4.3.2. Gradiente estocástico de la dinámica de Langevin

Recordemos que una solución sensata para un problema de aprendizaje automático es encontrar el clasificador que minimiza el riesgo empírico. Un algoritmo de aprendizaje diseñado para lograr este objetivo es el descenso por gradiente de la dinámica de Langevin (*SGLD*) cuya implementación se describe a continuación conforme se detalla en Wang y col. (2021):

Primero, se divide la muestra aleatoria S_n en m *mini-batches* disjuntos:

$$S_n = \bigcup_{i=1}^{m} B_i,$$

donde $|B_i| = M$ y $S_i \cap S_j = \varnothing$ para todo $i \neq j$. Se inicializa el algoritmo con un punto aleatorio $W_0 \in \mathcal{W}$ y se actualiza usando la siguiente regla:

$$W_t = \mathrm{Proj}_{\mathcal{W}} \left(W_{t-1} - \eta_t \nabla_w \hat{\ell}(W_{t-1}, B_{b_t}) + \sqrt{\frac{2\eta_t}{\beta_t}} N \right), \tag{4.13}$$

donde η_t es la tasa de aprendizaje, β_t es el inverso de la temperatura, N es una variable, completamente independiente del resto, con distribución $N_d(\mathbf{0}, \mathbf{I}_d)$, $b_t \in [m]^{\mathbb{N}}$ es la sucesión de índices que determina el *mini-batch* que será utilizado en cada paso de entrenamiento, $\hat{l}$ es una función de pérdida substituta para la función de pérdida l del problema de aprendizaje máquina en cuestión y

$$\nabla_w \hat{\ell}(W_{t-1}, B_{b_t}) \coloneqq \frac{1}{M} \sum_{Z \in B_{b_t}} \nabla_w \hat{l}(W_{t-1}, Z). \tag{4.14}$$

Finalmente, se devuelve la última iteración como resultado del procedimiento; es decir, si el algoritmo corre durante T iteraciones, éste regresa W_T. Particularmente, la mayor parte de la teoría relacionada con el SGLD proviene de un enfoque bayesiano. Con esto en mente, recordemos que si W es la variable aleatoria que representa los "verdaderos" parámetros del modelo, entonces una de las metodologías de la inferencia bayesiana consiste en encontrar el máximo a posteriori (MAP) de la distribución $p_{W \mid S_n}(\mathbf{w} \mid s_n)$ inducida, generalmente, por una función de densidad a priori $p_W(\mathbf{w})$ y una densidad condicional $p_{Z \mid W}(z \mid \mathbf{w})$.

Dado que bajo este paradigma, los elementos de la muestra son independientes cuando se condiciona en W, entonces

$$p_{W \mid S_n}(\mathbf{w} \mid s_n) \propto p_W(\mathbf{w}) \prod_{i=1}^{n} p_{Z \mid W}(z_i \mid \mathbf{w}).$$

De esta forma, un algoritmo de optimización como el SGLD resulta útil también el contexto bayesiano. Este acercamiento se presenta en Welling y Teh (2011) donde definen el SGLD mediante la regla de actualización

$$W_t = W_{t-1} + \eta_t \left(\nabla \log p_W(W_{t-1}) + \frac{n}{M} \sum_{Z \in B_{b_t}} \nabla \log p_{Z \mid W}(Z \mid W_{t-1}) \right) + \sqrt{2\eta_t} N, \quad (4.15)$$

bajo la condición de que $\sum_{t=1}^{\infty} \eta_t = \infty$ y $\sum_{t=1}^{\infty} \eta_t^2 < \infty$. Nótese que lo anterior es equivalente a tomar $\beta_t = 1$ para toda t y tomar como función de pérdida sustituta

$$\hat{l}(\mathbf{w,z}) = -\log p_W(\mathbf{w}) - n \log p_{Z \mid W}(z \mid \mathbf{w})$$

en el planteamiento detallado en Wang y col. (2021) (además de pedir que η_t cumpla las condiciones respectivas). Lo cual resulta en que

$$\nabla_{\mathbf{w}} \hat{\ell}(W_{t-1}, B_{b_t}) \coloneqq -\nabla_{\mathbf{w}} \log p_W(W_{t-1}) - \frac{n}{M} \sum_{Z \in B_{b_t}} \nabla_{\mathbf{w}} \log p_{Z \mid W}(Z \mid W_{t-1}).$$

Más aún, en Welling y Teh (2011) se explica cómo utilizar esta metodología para simular muestras de la distribución posterior. Adicionalemente, existen artículos más recientes, como Li y col. (2016), que utilizan el acercamiento bayesiano para entrenar Redes Neuronales Profundas. En contraste con lo anterior, en el planteamiento del SGLD en Wang y col. (2021) se refiere al lector a Gelfand y Mitter (1991) donde se presenta al algoritmo puramente como la solución de una ecuación diferencial estocástica.

Discusión

En la Sección 4.2 se describieron todos los componentes de un problema de aprendizaje máquina supervisado de tal manera que se pueda estudiar esta rama de las matemáticas aplicadas de una manera estandarizada. Uno de los conceptos claves presentados es el de algoritmo de aprendizaje. Nótese que al seleccionar éste, se está decidiendo cuál función $h \in \mathcal{H}$ aceptamos como el mejor clasificador que

podríamos deducir. Además, a los llamados algoritmos de aprendizaje se les permite ser una función aleatoria de la muestra y en la Subsección 4.3.2 se presenta un algoritmo de aprendizaje el cual se beneficia de no ser una función determinista de los datos de entrenamiento.

Cabe mencionar que una de las carencias de este trabajo es no ahondar en la teoría del *SGLD*. Particularmente, existe una extensa bibliografía sobre análisis bayesiano mediante el *SGLD* y qué garantías existen al utilizar estas técnicas (véase Li y col. 2016). Sin embargo, no se encontró una referencia con un enfoque aplicado para el caso que se describe en la primera mitad de la Subsección 4.3.2. A pesar de que en Wang y col. (2021) se refiere al lector a Gelfand y Mitter (1991), no es trivial deducir garantías para este algoritmo a partir de los resultados de este último texto.

Con este capítulo nos encontramos listos para comprender y apreciar las aportaciones de Wang y col. (2021). Particularmente, conocemos los componentes de un problema de aprendizaje automático supervisado, estamos familiarizados con el error de generalización esperado y la intuición detrás de este concepto. En el siguiente capítulo se explica mejor la importancia del concepto de generalización y, finalmente, se presentan resultados que nos permiten analizar este fenómeno por medio del error de generalización esperado y la información mutua.

Capítulo 5

Acotando el error de generalización esperado

Hasta el momento nuestros esfuerzos se han concentrado en introducir conceptos que no necesariamente se revisan durante una maestría enfocada a la probabilidad y estadística. El trabajo realizado hasta ahora rendirá frutos durante este capítulo al permitirnos exponer algunos de los resultados presentados en Wang y col. (2021). Es importante recalcar que la meta principal de esta tesis es presentar éstos con la mayor claridad posible.

El objetivo de este capítulo es demostrar el teorema principal en Wang y col. (2021) con el mayor detalle posible. Concretamente, dentro de este capítulo se prueban todos los lemas utilizados en Wang y col. (2021) para construir una cota para el error de generalización que dependa del algoritmo de aprendizaje.

Estructuralmente, la primera sección ahonda en el concepto de generalización en el aprendizaje de máquina supervisado. En la parte subsecuente se introducen las variables aleatorias σ-sub-Gaussianas. Luego, en la tercera sección, este tipo de variables se utilizan para enunciar un lema que permite acotar el error de generalización con ayuda de la información mutua. Finalmente, con ayuda de lo anterior, en el último apartado de este capítulo se prueba el resultado principal de Wang y col. (2021).

5.1. La capacidad de generalización de un algoritmo

El aprendizaje máquina puede interpretarse como una evolución natural de la disciplina estadística en un mundo donde las computadoras se encuentran ampliamente disponibles. Lógicamente, esta nueva disciplina comparte con la estadística un objetivo central: encontrar patrones en eventos pasados para poder predecir eventos futuros; en otras palabras, se busca generalizar el comportamiento de un fenómeno a partir de observaciones de éste. De acuerdo con Zhang y col. (2021), en el aprendizaje supervisado se analiza la capacidad de generalización de un algoritmo de la siguiente manera:

1. Se asume que un conjunto de observaciones provienen de un fenómeno fijo que genera datos.
2. En el paso de *entrenamiento* se ajusta un modelo a un conjunto de datos.
3. En el paso de *evaluación* se juzga el desempeño del modelo en un nuevo conjunto de datos generados por el mismo fenómeno.

A pesar de que el concepto descrito anteriormente puede parecer sencillo, la teoría que describe la capacidad de generalización que posee un algoritmo ha eludido por muchos años a aquellos que investigan el aprendizaje estadístico. Concretamente, existen una variedad de teorías que buscan explicar el fenómeno de generalización: convergencia uniforme, estabilidad de algoritmos, entre otros. Sin embargo, la validez de estas teorías para explicar la capacidad que tiene un modelo para generalizar sigue siendo un debate (Zhang y col. 2021).

Actualmente, la eficacia de las Redes Neuronales Artificiales (ANN) para generalizar a partir de lo particular ha orillado a la comunidad científica a replantearse algunas nociones respecto a la capacidad de generalización de un algoritmo. En particular, Zhang y col. (2021) mencionan: «[...] hemos encontrado que la mayoría de las más populares maneras de explicar la generalización fracasan al intentar explicar lo que sucede en los modelos de aprendizaje profundo más recientes.»

Concretamente, en el artículo de Zhang y col. (2021) se prueba que una clase particular de redes neuronales es capaz de memorizar una muestra de tamaño n (donde cada instancia es de dimenisón d) siempre que se cuente con $p = 2n + d$ parámetros. Sin embargo, este nivel de expresividad en los datos de entrenamiento junto con la evidencia empírica respecto a la gran capacidad de generalización de las ANN, desacredita ideas clásicas que aseguraban deficiencias en la generalización si el modelo se ajusta demasiado a la muestra.

En Wang y col. (2021), se usa el **error de generalización esperado** (Defini-

ción 4.5) como una herramienta para cuantificar la capacidad de generalizar de un algoritmo de aprendizaje. En este artículo se deduce una cota para éste, la cual está basada en la relación entre la muestra y el algoritmo implementado para la optimización de los parámetros. Este trabajo es parte de un esfuerzo conjunto por estudiar la generalización de las redes neuronales desde una nueva perspectiva.

5.2. Variables aleatorias σ-sub-Gaussianas

La representación variacional de la divergencia de Kullback-Leibler (Polyanskiy y Wu 2019, Teorema 3.5) permite encontrar desigualdades que involucren a esta cantidad mediante propiedades de las funciones generadoras de momentos. Con esto en mente, nos gustaría trabajar con variables aleatorias cuya función generadora de momentos esté acotada por una expresión sencilla. Esto motiva la siguiente definición.

Definición 5.1. Se dice que una variable aleatoria U es σ-sub-Gaussiana si para todo $\lambda \in \mathbb{R}$

$$\log \mathbb{E}\left(e^{\lambda[U-\mathbb{E}(U)]}\right) \leq \frac{\lambda^2\sigma^2}{2}. \tag{5.1}$$

Ejemplo 5.1. Es bien conocido (véase, por ejemplo, Seber y Lee 2003, p. 20) que si $X \sim N(\mu, \sigma^2)$, entonces

$$\mathbb{E}\left[e^{\lambda(X-\mathbb{E}(X))}\right] = e^{\frac{\lambda^2\sigma^2}{2}}.$$

Así, las variables aleatorias normales univariadas con varianza σ^2 son σ-sub-Gaussianas.

Ejemplo 5.2. Debido al Lema de Hoeffding, es directo corroborar que una variable aleatoria que toma valores en el intervalo $[a, b]$ es $(b-a)/2$-sub-Gaussiana.

En la próxima sección se acotará el error de generalización esperado usando la información mutua bajo un supuesto de σ-sub-Gaussianidad. La principal herramienta que se usa es el teorema de Donsker-Varadhan (Polyanskiy y Wu 2019, Teorema 3.5).

5.3. Acotando el error de generalización utilizando la información mutua

El siguiente resultado se da en el contexto de un problema de aprendizaje estadístico donde se trabaja con una muestra aleatoria de características con sus respectivas

etiquetas denotada por $S_n = (Z_1, \ldots, Z_n) \sim \mu^{\otimes n}$, un algoritmo de aprendizaje W que toma valores en una clase de hipótesis $\mathcal{W}$ y una función de pérdida l.

Lema 5.1 (Xu y Raginsky 2017, Teorema 1)**.** *Si $l(w, Z)$ es σ-sub-Gaussiana bajo μ para toda $w \in \mathcal{W}$, entonces*

$$|gen(\mu, P_{W\,|\,S_n})| \leq \sqrt{\frac{2\sigma^2}{n} I(S_n; W)}. \tag{5.2}$$

Demostración. Sea $\tilde{S}_n = (\tilde{Z}_1, \ldots, \tilde{Z}_n) \sim \mu^{\otimes n}$ una muestra aleatoria independiente de W. Basándose en (Jacod y Protter 2004, ejercicio 23.7) se sigue que

$$L_\mu(W) = \mathbb{E}(l(W, \tilde{Z}_i)\,|\,W), \;\; \forall i \in [n]. \tag{5.3}$$

Así, para todo $i \in [n]$, $\mathbb{E}(L_\mu(W)) = \mathbb{E}(l(W, \tilde{Z}_i))$. Luego,

$$\mathbb{E}(L_\mu(W)) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(l(W, \tilde{Z}_i)). \tag{5.4}$$

Además, con base en (4.2), es directo que

$$\mathbb{E}(L_{S_n}(W)) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(l(W, Z_i)). \tag{5.5}$$

Sea $f : \mathcal{W} \times \mathcal{Z}^n \to \mathbb{R}$ definida por $f(\mathbf{w}, \mathbf{s}) = \frac{1}{n} \sum\limits_{i=1}^{n} l(\mathbf{w}, \mathbf{z}_i)$ con $\mathbf{s} = (z_1, \ldots, z_n)$. De tal manera que la diferencia entre (5.4) y (5.5) puede representarse como

$$\mathbb{E}([L_\mu(W) - L_{S_n}(W)] = \mathbb{E}[f(W, \tilde{S}_n)] - \mathbb{E}[f(W, S_n)].$$

Luego,

$$\text{gen}(\mu, P_{W\,|\,S_n}) = \mathbb{E}(f(W, \tilde{S}_n)) - \mathbb{E}(f(W, S_n)). \tag{5.6}$$

Por otro lado, debido a que $\tilde{Z}_1, \ldots, \tilde{Z}_n$ son independientes, se obtiene

$$\begin{aligned} \log \mathbb{E}(e^{\lambda[f(\mathbf{w}, \tilde{S}_n) - \mathbb{E}(f(\mathbf{w}, \tilde{S}_n))]}) &= \log \mathbb{E}\left[\exp\left(\frac{\lambda}{n} \sum_{i=1}^{n} \Big(l(\mathbf{w}, \tilde{Z}_i) - \mathbb{E}[l(w, \tilde{Z}_i)]\Big)\right)\right] \\ &= \sum_{i=1}^{n} \log \mathbb{E}\left(e^{\frac{\lambda}{n}[l(\mathbf{w}, \tilde{Z}_i) - \mathbb{E}(l(\mathbf{w}, \tilde{Z}_i)]}\right). \end{aligned}$$

Así, si $l(\mathbf{w}, Z)$ es σ-sub-Gaussiana bajo μ para todo $\mathbf{w} \in \mathcal{W}$, entonces

$$\log \mathbb{E}\left(e^{\frac{\lambda}{n}[l(\mathbf{w}, \tilde{Z}_i) - \mathbb{E}(l(\mathbf{w}, \tilde{Z}_i)]}\right) \leq \lambda^2 \sigma^2 / 2n^2, \;\; \forall i \in \{1, \ldots n\}.$$

Por lo que

$$\log \mathbb{E}(e^{\lambda[f(\mathbf{w},\tilde{S}_n)-\mathbb{E}(f(\mathbf{w},\tilde{S}_n))]}) \leq \frac{\lambda^2\sigma^2}{2n}, \ \ \forall \mathbf{w} \ \in \mathcal{W}. \tag{5.7}$$

Es decir, $f(\mathbf{w}, \tilde{S}_n)$ es $(\sigma/\sqrt{n})$-sub-Gaussiana. Defínase $\tau = \frac{\sigma}{\sqrt{n}}$ y

$$\phi(\mathbf{w}, \mathbf{s}) = f(\mathbf{w}, \mathbf{s}) - \mathbb{E}(f(\mathbf{w}, \tilde{S}_n)).$$

La representación variacional de la divergencia de Kullback-Leibler (Polyanskiy y Wu 2019, Teorema 3.5) garantiza que para cualesquiera medidas de probabilidad P y Q en un espacio $\mathcal{X}$, si $\mathcal{C}$ denota el conjunto de funciones $f : \mathcal{X} \to \mathbb{R}$ tal que $\mathbb{E}_Q[\exp(f(X)] < \infty$, entonces para toda $f \in \mathcal{C}$ se tiene que $\mathbb{E}_P[f(X)]$ existe y

$$D\left(P \,||\, Q\right) = \sup_{f \in \mathcal{C}} \mathbb{E}_P\left[f(X)\right] - \log \mathbb{E}_Q\left[e^{f(X)}\right].$$

Luego, con base en lo anterior, es posible asegurar, para todo $\lambda \in \mathbb{R}$,

$$D(P_{WS_n} \,||\, P_W \otimes P_{S_n}) \geq \mathbb{E}[\lambda\phi(W, S_n)] - \log \mathbb{E}[e^{\lambda\phi(W,\tilde{S}_n)}]. \tag{5.8}$$

Así, al aplicar el teorema de Fubini (Jacod y Protter 2004, Teorema 10.3) y el hecho de que $f(w, \tilde{S}_n)$ es τ-sub-Gaussiana, se tiene

$$\mathbb{E}[e^{\lambda\phi(W,\tilde{S}_n)}] = \int \mathbb{E}\left(e^{\lambda[f(w,\tilde{S}_n)-\mathbb{E}(f(w,\tilde{S}_n))]}\right) P_W(dw) \leq e^{\frac{\lambda^2\tau^2}{2}} \tag{5.9}$$

Luego, al juntar (5.8) y (5.9), se obtiene

$$I(W; S_n) \geq \lambda E[\phi(W, S_n)] - \frac{\lambda^2\tau^2}{2}. \tag{5.10}$$

De tal suerte que la función $\lambda \mapsto \frac{\lambda^2\tau^2}{2} - \lambda\mathbb{E}(\phi(W, S_n)) + I(W; S_n)$ define una parábola no-negativa. Se sigue que el discriminante de ésta debe ser no-positivo, es decir,

$$[\mathbb{E}(\phi(W, S_n))]^2 - 2\tau^2 I(W; S_n) \leq 0. \tag{5.11}$$

Nótese que de (5.6) se deduce inmediatamente que $\mathbb{E}[\phi(W, S_n)] = \text{gen}(\mu, P_{W\,|\,S_n})$. Así, al reacomodar (5.11) concluye la prueba. □

Como se menciona en Bu, Zou y Veeravalli (2020), es altamente relevante notar que $I(W; S_n)$ depende de todos los elementos principales de un problema de aprendizaje máquina; i.e., la clase de hipótesis $\mathcal{W}$, la distribución μ de cada elemento de la muestra y el kernel de transición $P_{W\,|\,S_n}$. Sin embargo, Bu, Zou y Veeravalli (2020) señalan algunas limitantes de esta cota. En nuestro caso, la más relevante se deriva

de que si W es una función determinista de S_n, por ejemplo en ERM, entonces $I(W; S_n) = \infty$.

La alternativa que se propone en Bu, Zou y Veeravalli (2020) se centra en utilizar una cota basada en la información mutua entre cada instancia de la muestra y el algoritmo de aprendizaje (*Individual Sample Mutual Information*) conforme se presenta en el siguiente lema. Además, de acuerdo con la Proposición 2 en Bu, Zou y Veeravalli (2020), ésta es una cota más estricta que la del Lema 5.1.

Lema 5.2 (Bu, Zou y Veeravalli 2020, Proposición 1.1)**.** *Si $l(w, Z)$ es σ-sub-Gaussiana bajo μ para todo $w \in \mathcal{W}$, entonces*

$$|gen(\mu, P_{W \mid S_n})| \leq \frac{1}{n} \sum_{i=1}^{n} \sqrt{2\sigma^2 I(W; Z_i)}. \tag{5.12}$$

Demostración. De las ecuaciones (5.4) y (5.5) se sigue

$$|\text{gen}(\mu, P_{W \mid S_n})| = \frac{1}{n} \left| \sum_{i=1}^{n} [\mathbb{E}(l(W, \tilde{Z}_i)) - \mathbb{E}(l(W, Z_i))] \right|.$$

Así, aplicar la desigualdad del triángulo y el Lema 5.1 reemplazando S_n por Z_i resulta en

$$|\text{gen}(\mu, P_{W \mid S_n})| \leq \frac{1}{n} \sum_{i=1}^{n} \sqrt{2\sigma^2 I(W; Z_i)}.$$

□

5.4. Acotando el error de generalización del *SGLD*

En esta sección se presenta el resultado principal de Wang y col. (2021): una cota para el error de generalización esperado en un problema de aprendizaje estadístico que usa como algoritmo de aprendizaje el *SGLD*. A pesar de que el Lema 5.2 nos brinda una cota que es función de la información mutua entre el *SGLD* y la muestra, ésta no resulta muy útil debido a la dificultad de calcular dicha cantidad en la práctica (para darse una idea de dicho problema véase, por ejemplo, Zbili y Rama 2021; Goebel y col. 2005). A fin de lidiar con esto, el Teorema 1 en Wang y col. (2021) resulta de continuar acotando superiormente con ayuda de algunas desigualdades de procesamiento de la información y el Lema 3.2.

Primeramente, el siguiente resultado nos permite acotar los sumandos que aparecen en el Lema 5.2. Desafortunadamente, obtener una expresión para la cota de cada uno de éstos vuelve a depender de calcular la información mutua entre dos

variables. Sin embargo, como se verá más adelante, resulta sencillo volver a acotar estas nuevas cantidades con ayuda de la distancia de Wasserstein. Adicionalmente, el Lema 5.3 implica que, al usar el *SGLD*, la información de las instancias que incorpora el algoritmo en cada paso decae conforme pasan las iteraciones.

Lema 5.3 (Wang y col. 2021, Lema 2)**.** *En el contexto del SGLD (véase la Sección 4.3.2 para recordar los detalles de este algoritmo), supóngase que el espacio parametral $\mathcal{W}$ es compacto con diámetro D y $\|\nabla_{\boldsymbol{w}} \hat{l}(\boldsymbol{w}, \boldsymbol{z})\| \leq K$ para todo $\boldsymbol{w}, \boldsymbol{z}$. Si el muestreo de los mini-batches se hace con reemplazo y el punto Z_i fue usado por última vez en la t-ésima iteración, entonces*

$$I(W_T; Z_i) \leq I(W_t; Z_i) \cdot \prod_{t'=t+1}^{T} q_{t'}, \tag{5.13}$$

donde

$$q_{t'} := 1 - 2\bar{\Phi}\left((D + 2\eta_{t'}K)\sqrt{\frac{\beta_{t'}}{8\eta_{t'}}}\right) \in (0, 1) \tag{5.14}$$

y $\bar{\Phi}(\cdot)$ es la función de supervivencia de una variable aleatoria normal estándar.

Demostración. Para la t-ésima iteración se puede escribir la recursión en (4.13) como

$$U_t = W_{t-1} - \eta_t \nabla_{\mathbf{w}} \hat{l}(W_{t-1}, B_{b_t}) \tag{5.15}$$

$$V_t = U_t + \sqrt{\frac{2\eta_t}{\beta_t}} \cdot N \tag{5.16}$$

$$W_t = \text{Proj}_{\mathcal{W}}(V_t) \tag{5.17}$$

Sea Z_i una instancia de la muestra y sea t la última iteración donde se utilizó Z_i. Dado que los *mini-batches* son disjuntos, entonces

$$Z_i - U_t - V_t - W_t - \cdots - W_{T-1} - U_T - V_T - W_T. \tag{5.18}$$

Si $\mathcal{U}_T$ es el soporte de U_T, debido a la desigualdad del triángulo se tiene

$$\begin{aligned}
\text{diam}(\mathcal{U}_T) &= \sup_{\mathbf{w}, \mathbf{w}^* \in \mathcal{W}} \|\mathbf{w} - \eta_T \nabla \hat{l}(\mathbf{w}, B_{b_T}) - [\mathbf{w}^* - \eta_T \nabla \hat{l}(\mathbf{w}^*, B_{b_T})]\| \\
&\leq \sup_{\mathbf{w}, \mathbf{w}^* \in \mathcal{W}} \left(\|\mathbf{w} - \mathbf{w}^*\| + \eta_T \|\nabla\, \hat{l}(\mathbf{w}, B_{b_T})\| + \eta_T \|\nabla\, \hat{l}(\mathbf{w}^*, B_{b_T})\|\right)
\end{aligned}$$

Dado que $\text{diam}(\mathcal{W}) = D$ y $\|\nabla_{\mathbf{w}}\, \hat{l}(\mathbf{w}, \mathbf{z})\| \leq K$ para todo $\mathbf{w}, \mathbf{z}$, lo anterior implica

$$\text{diam}(\mathcal{U}_T) \leq D + 2\eta_T K. \tag{5.19}$$

Luego, debido a la desigualdad de procesamiento de la información para la información mutua (Proposición 2.9.3)

$$I(W_{t+1}; Z_i) \leq I(V_{t+1}; Z_i). \tag{5.20}$$

Posteriormente, apelando a la Proposición 2.12

$$I(V_{t+1}; Z_i) \leq \eta_{\mathrm{KL}}(P_{V_{t+1}\,|\,U_{t+1}}) \cdot I(U_{t+1}; Z_i). \tag{5.21}$$

Dado que (5.19) es válida para todo $T > t$, del Ejemplo 2.11 se deduce

$$\eta_{\mathrm{KL}}(P_{V_{t+1}\,|\,U_{t+1}}) \cdot I(U_{t+1}; Z_i) \leq \left(1 - 2\bar{\Phi}\left(\frac{(D + 2\eta_{t+1}K)}{2\sqrt{\frac{2\eta_{t+1}}{\beta_{t+1}}}}\right)\right) \cdot I(U_{t+1}, Z_i). \tag{5.22}$$

Juntando (5.20)-(5.22) e invocando, de nuevo, la desigualdad de procesamiento de la información para la información mutua, se concluye que

$$I(W_{t+1}; Z_i) \leq I(W_t, Z_i) \cdot q_{t+1}.$$

De la misma forma, si k es un entero positivo tal que $t + k + 1 \leq T$, se asegura que

$$I(W_{t+k+1}, Z_i) \leq q_{t+k+1} I(W_{t+k}, Z_i).$$

Luego, la proposición del enunciado se sigue por inducción sobre k. □

Como ya se ha mencionado, para obtener una cota útil en un contexto práctico, resulta de vital importancia acotar la información mutua entre (W_t, Z_i), donde t es la última iteración en la cual se utilizó Z_i para entrenar el algoritmo. Concretamente, esta nueva cota a encontrar depende del objeto de la siguiente definición

Definición 5.2. Se define la varianza total de un vector aleatorio $X \in \mathbb{R}^d$ como

$$\mathrm{V}(X) := \mathbb{E}\left(\|X - \mathbb{E}(X)\|_2^2\right). \tag{5.23}$$

Es fácil notar que si $\mathrm{Var}(X)$ es la matriz de varianza y covarianza de X, entonces $\mathrm{V}(X) = \mathrm{tr}(\mathrm{Var}(X))$.

Con base en lo anterior, nos encontramos listos para enunciar y probar el resultado principal de Wang y col. (2021). Este teorema aplica los Lemas 5.2 y 5.3 para acotar el error de generalización esperado. Además, continúa acotando superiormente para deducir una desigualdad que pueda ser aplicada sin necesidad de recurrir a aproximaciones de la información mutua.

Teorema 5.4 (Wang y col. 2021, Teorema 1)**.** *Supóngase que el espacio parametral $\mathcal{W}$ es compacto con diámetro D y $\|\nabla_{\boldsymbol{w}} \hat{l}(\boldsymbol{w}, \boldsymbol{z})\| \leq K$ para todo $\boldsymbol{w}, \boldsymbol{z}$ y que la función de pérdida $l(\boldsymbol{w}, Z)$ es σ-sub-Gaussiana bajo $Z \sim \mu$ para todo $\boldsymbol{w} \in \mathcal{W}$. Entonces el error de generalización del algoritmo SGLD al tiempo T está acotado por*

$$\frac{\sqrt{2M}\sigma}{n} \sum_{i=1}^{m} \sqrt{\sum_{t \in \mathcal{T}_i} \beta_t \eta_t \cdot V\left(\nabla_{\boldsymbol{w}} \hat{\ell}(W_{t-1}, B_i)\right) \cdot \prod_{\substack{t'=t+1 \\ t' \notin \mathcal{T}_i}}^{T} q_{t'}}, \tag{5.24}$$

donde el conjunto $\mathcal{T}_i$ contiene los índices de las iteraciones en las cuales el mini-batch B_i es usado; $q_{t'}$ está definida como en (5.14); y

$$\nabla_{\boldsymbol{w}} \hat{\ell}(W_{t-1}, B_i) \coloneqq \frac{1}{M} \sum_{Z \in B_i} \nabla_{\boldsymbol{w}} \hat{l}(W_{t-1}, Z). \tag{5.25}$$

Demostración. Primero, se trabajará con el caso más sencillo: cuando el tamaño del *mini-batch* es uno; i.e. $|B_i| = 1$ para todo $i \in [m]$. Así, el Lema 5.3 y la desigualdad de procesamiento de la información implican

$$I(W_T; Z_i) \leq \prod_{t'=t+1}^{T} q_{t'} \cdot I(W_t; Z_i) \leq \prod_{t'=t+1}^{T} q_{t'} \cdot I(V_t; Z_i). \tag{5.26}$$

A continuación, se busca acotar superiormente $I(V_t; Z_i)$. Reescribiendo V_t y U_t como en (5.16) y (5.15) respectivamente, resulta en

$$I(V_t; Z_i) = I\left(W_{t-1} - \eta_t \nabla_{\mathbf{w}} \hat{l}(W_{t-1}, B_{b_t}) + \sqrt{\frac{2\eta_t}{\beta_t}} N; Z_i\right). \tag{5.27}$$

Sea $\phi : (x, y) \mapsto x + y$. En virtud de la desigualdad de procesamiento de la información de la información mutua (Proposición 2.9.3),

$$\begin{aligned} I(V_t; Z_i) &= I\left(\phi\left(W_{t-1}, -\eta_t \nabla_{\mathbf{w}} \hat{l}(W_{t-1}, B_{b_t}) + \sqrt{\frac{2\eta_t}{\beta_t}} N\right); Z_i\right) \\ &\leq I\left(W_{t-1}, -\eta_t \nabla_{\mathbf{w}} \hat{l}(W_{t-1}, B_{b_t}) + \sqrt{\frac{2\eta_t}{\beta_t}} N; Z_i\right). \end{aligned} \tag{5.28}$$

Dado que Z_i se usa en la t-ésima iteración y $|B_{b_t}| = 1$, entonces $B_{b_t} = \{Z_i\}$. Más aún, al aplicar la identidad de Kolmogorov (Proposición 2.9.2) y, de nuevo, la desigualdad de procesamiento de la información, de la expresión en (5.28) se deduce

$$I(V_t; Z_i) \leq I(W_{t-1}, Z_i) + I\left(-\nabla_{\mathbf{w}} \hat{l}(W_{t-1}, Z_i) + \sqrt{\frac{2}{\beta_t \eta_t}} N; Z_i \,\middle|\, W_{t-1}\right). \tag{5.29}$$

En busca de simplificar la notación, defínase $L \coloneqq -\nabla_{\mathbf{w}}\hat{l}(W_{t-1}, Z_i)$. Nótese que, usando la definición de la información mutua condicional (2.46) y la Proposición 2.6.1, para $\mathbf{w} \in \mathcal{W}$,

$$I\left(L + \sqrt{\frac{2}{\eta_t\beta_t}}N; Z_i \,\middle|\, W_{t-1} = \mathbf{w}\right)$$
$$= \int D\left(P_{L+\sqrt{\frac{2}{\eta_t\beta_t}}N \,|\, Z_i, W_{t-1}}(\cdot \,|\, \mathbf{z}, \mathbf{w}) \,\middle\|\, P_{L+\sqrt{\frac{2}{\eta_t\beta_t}}N \,|\, W_{t-1}}(\cdot \,|\, \mathbf{w})\right) P_{Z_i \,|\, W_{t-1}}(d\mathbf{z} \,|\, \mathbf{w}). \tag{5.30}$$

Sean $\mathbf{w} \in \mathcal{W}$, $\mathbf{z} \in \mathcal{Z}$ y $Z_{\mathbf{w}} \sim P_{Z_i \,|\, W}(\cdot \,|\, \mathbf{w})$. Es claro que la variable $L^*_{\mathbf{z},\mathbf{w}} \coloneqq -\nabla_{\mathbf{w}}\hat{l}(\mathbf{w}, \mathbf{z}) + \sqrt{\frac{2}{\eta_t\beta_t}}N$ se distribuye de acuerdo a la ley $P_{L+\sqrt{\frac{2}{\eta_t\beta_t}}N \,|\, Z_i, W_{t-1}}(\cdot \,|\, \mathbf{z}, \mathbf{w})$ y la variable $L^*_{\mathbf{w}} \coloneqq -\nabla_{\mathbf{w}}\hat{l}(\mathbf{w}; Z_{\mathbf{w}}) + \sqrt{\frac{2}{\eta_t\beta_t}}N$ tiene distribución $P_{L+\sqrt{\frac{2}{\eta_t\beta_t}}N \,|\, W_{t-1}}(\cdot \,|\, \mathbf{w})$. Así, al aplicar el Lema 3.2 se obtiene

$$D\left(P_{L+\sqrt{\frac{2}{\eta_t\beta_t}}N \,|\, Z_i, W_{t-1}}(\cdot \,|\, \mathbf{z}, \mathbf{w}) \,\middle\|\, P_{L+\sqrt{\frac{2}{\eta_t\beta_t}}N \,|\, W_{t-1}}(\cdot \,|\, \mathbf{w})\right) = D\left(P_{L^*_{\mathbf{z},\mathbf{w}}} \,\middle\|\, P_{L^*_{\mathbf{w}}}\right) \tag{5.31}$$
$$\leq \frac{\eta_t\beta_t}{4} W_2^2\left(P_{\nabla_{\mathbf{w}}\hat{l}(\mathbf{w},\mathbf{z})}, P_{\nabla_{\mathbf{w}}\hat{l}(\mathbf{w},Z_{\mathbf{w}})}\right). \tag{5.32}$$

Es claro que $\nabla_{\mathbf{w}}\hat{l}(\mathbf{w}, \mathbf{z})$ es, de hecho, una variable determinista, por lo que la definición de la distancia de Wasserstein implica

$$\frac{\eta_t\beta_t}{4} W_2^2\left(P_{\nabla_{\mathbf{w}}\hat{l}(\mathbf{w},\mathbf{z})}, P_{\nabla_{\mathbf{w}}\hat{l}(\mathbf{w},Z_{\mathbf{w}})}\right) = \frac{\eta_t\beta_t}{4}\mathbb{E}\left(\left\|\nabla_{\mathbf{w}}\hat{l}(\mathbf{w}, \mathbf{z}) - \nabla_{\mathbf{w}}\hat{l}(\mathbf{w}, Z_{\mathbf{w}})\right\|_2^2\right). \tag{5.33}$$

Nótese que $\|a - b\|_2^2 \leq 2\|a\|_2^2 + 2\|b\|_2^2$. Luego, tomando $\mathbf{e} \coloneqq \mathbb{E}\left(\nabla_{\mathbf{w}}\hat{l}(W_{t-1}, Z_i)\right)$, se sigue que

$$\frac{\eta_t\beta_t}{4}\mathbb{E}\left(\|\nabla_{\mathbf{w}}\hat{l}(\mathbf{w}, \mathbf{z}) - \nabla_{\mathbf{w}}\hat{l}(\mathbf{w}, Z_i)\|_2^2\right)$$
$$\leq \frac{\eta_t\beta_t}{4}\left[2\left\|\nabla_{\mathbf{w}}\hat{l}(\mathbf{w}, \mathbf{z}) - \mathbf{e}\right\|_2^2 + 2\mathbb{E}\left(\left\|\nabla_{\mathbf{w}}\hat{l}(\mathbf{w}, Z_{\mathbf{w}}) - \mathbf{e}\right\|_2^2\right)\right]. \tag{5.34}$$

Juntando (5.31)-(5.34) e integrando respecto a $\mathbf{z}$ bajo la medida de probabilidad $P_{Z_i \,|\, W_{t-1}}(\cdot \,|\, \mathbf{w})$ resulta en

$$\int D\left(P_{L+\sqrt{\frac{2}{\eta_t\beta_t}}N \,|\, Z_i, W_{t-1}}(\cdot \,|\, \mathbf{z}, \mathbf{w}) \,\middle\|\, P_{L+\sqrt{\frac{2}{\eta_t\beta_t}}N \,|\, W_{t-1}}(\cdot \,|\, \mathbf{w})\right) P_{Z_i \,|\, W_{t-1}}(d\mathbf{z} \,|\, \mathbf{w})$$
$$\leq \eta_t\beta_t\mathbb{E}\left(\left\|\nabla_{\mathbf{w}}\hat{l}(\mathbf{w}, Z_{\mathbf{w}}) - \mathbf{e}\right\|_2^2\right). \tag{5.35}$$

Es decir, con base en (5.30),

$$I\left(L+\sqrt{\frac{2}{\eta_t\beta_t}N};Z_i\,\middle|\,W_{t-1}=\mathbf{w}\right)\leq\eta_t\beta_t\mathbb{E}\left(\left\|\nabla_{\mathbf{w}}\hat{l}(\mathbf{w},Z_{\mathbf{w}})-\mathbf{e}\right\|_2^2\right). \tag{5.36}$$

Al integrar la desigualdad anterior con respecto a $\mathbf{w}$ bajo la medida $P_{W_{t-1}}$ se asegura que

$$I\left(L+\sqrt{\frac{2}{\eta_t\beta_t}N};Z_i\,\middle|\,W_{t-1}\right)\leq\eta_t\beta_t\mathbb{E}\left(\left\|\nabla_{\mathbf{w}}\hat{l}(W_{t-1},Z_i)-\mathbf{e}\right\|_2^2\right). \tag{5.37}$$

Así, combinando (5.29) y (5.37) se obtiene

$$I(V_t;Z_i)\leq I(W_{t-1};Z_i)+\eta_t\beta_t\mathrm{V}(\nabla_{\mathbf{w}}\hat{l}(W_{t-1},Z_i)). \tag{5.38}$$

Usando lo anterior en (5.26), resulta en

$$I(W_T;Z_i)\leq\prod_{t'=t+1}^{T}q_{t'}\cdot\left(I(W_{t-1};Z_i)+\eta_t\beta_t\mathrm{V}(\nabla_{\mathbf{w}}\hat{l}(W_{t-1},Z_i))\right). \tag{5.39}$$

Nótese que estos mismos argumentos implican que

$$I(W_{t-1};Z_i)\leq\prod_{t'=t^*+1}^{t-1}q_{t'}\cdot\left(I(W_{t^*-1};Z_i)+\eta_{t^*}\beta_{t^*}\mathrm{V}(\nabla_{\mathbf{w}}\hat{l}(W_{t^*-1},Z_i))\right), \tag{5.40}$$

donde t^* es la última iteración donde se utilizó Z_i hasta el momento $t-1$. Al sustituir (5.40) en (5.39) y aplicar este procedimiento de manera recursiva se sigue

$$I(W_T;Z_i)\leq\sum_{t\in\mathcal{T}_i}\beta_t\eta_t\prod_{\substack{t'=t+1\\t'\notin\mathcal{T}_i}}^{T}q_{t'}\cdot\mathrm{V}\left(\nabla_{\mathbf{w}}\hat{l}(W_{t-1},Z_i)\right). \tag{5.41}$$

Aplicando el Lema 5.2 se concluye que

$$\mathrm{gen}(\mu,P_{W_T\,|\,S_n})\leq\frac{\sqrt{2}\sigma}{n}\sum_{i=1}^{n}\sqrt{\sum_{t\in\mathcal{T}_i}\beta_t\eta_t\prod_{\substack{t'=t+1\\t'\notin\mathcal{T}_i}}^{T}q_{t'}\cdot\mathrm{V}\left(\nabla_{\mathbf{w}}\hat{l}(W_{t-1},Z_i)\right)}. \tag{5.42}$$

Ahora, se considerará el caso donde el tamaño del *mini-batch* es mayor a uno. Tómese una nueva función de pérdida

$$\ell(\mathbf{w},B_i)\coloneqq\frac{1}{M}\sum_{Z\in B_i}l(\mathbf{w},Z)$$

y una nueva función de pérdida sustituta $\hat{\ell}(\mathbf{w}, B_i) := \frac{1}{M} \sum_{Z \in B_i} \hat{l}(\mathbf{w}, Z)$; de tal manera que

$$\nabla_{\mathbf{w}} \hat{\ell}(\mathbf{w}, B_i) = \frac{1}{M} \sum_{Z \in B_i} \nabla_{\mathbf{w}} \hat{l}(\mathbf{w}, Z). \tag{5.43}$$

Durante la demostración del Lema 5.1 se mostró que si $l(\mathbf{w}, Z)$ es σ-sub-Gaussiana bajo $Z \sim \mu$ para todo $\mathbf{w} \in \mathcal{H}$, entonces $\ell(\mathbf{w}, Z)$ es $(\sigma/\sqrt{M})$-subgaussiana. Posteriormente, se toma cada *mini-batch* como una observación de la muestra para un problema de aprendizaje estadístico con función de périda ℓ, algoritmo de aprendizaje *SGLD* y función de pérdida sustituta $\hat{\ell}$. En virtud de (5.42), se deduce

$$\operatorname{gen}(\mu, P_{W_T \mid S_n}) \leq \frac{\sqrt{2}\sigma}{m\sqrt{M}} \sum_{i=1}^{m} \sqrt{\sum_{t \in \mathcal{T}_i} \beta_t \eta_t \prod_{\substack{t'=t+1 \\ t' \notin \mathcal{T}_i}}^{T} q_{t'} \cdot \mathrm{V}\left(\nabla_{\mathbf{w}} \hat{\ell}(W_{t-1}, B_i)\right)}. \tag{5.44}$$

Nótese que $n = m \cdot M$, por lo que

$$\operatorname{gen}(\mu, P_{W_T \mid S_n}) \leq \frac{\sqrt{2M}\sigma}{n} \sum_{i=1}^{m} \sqrt{\sum_{t \in \mathcal{T}_i} \beta_t \eta_t \prod_{\substack{t'=t+1 \\ t' \notin \mathcal{T}_i}}^{T} q_{t'} \cdot \mathrm{V}\left(\nabla_{\mathbf{w}} \hat{\ell}(W_{t-1}, B_i)\right)}. \tag{5.45}$$

□

Discusión

Conforme se ejemplifica bien en este capítulo, la teoría de la información brinda herramientas para analizar procedimientos de naturaleza estadística. Con ayuda de la información mutua, Wang y col. (2021) pudieron deducir una cota para el error de generalización esperado que depende del algoritmo que se usa para entrenar el modelo y de la distribución μ. Según se menciona en el mismo artículo, este resultado es novedoso debido a que las cotas más conocidas dependen solamente de la clase de hipótesis con la que se trabaja.

De acuerdo con Wang y col. (2021), la cota obtenida tiene una correlación más alta con el error de generalización que una cota similar deducida en Negrea y col. (2019). Sin embargo, es importante notar que en Wang y col. (2021) no se verifica el ajuste de ésta por lo que sólo es útil para describir el comportamiento del error de generalización pero no es una opción viable para garantizar que este error no excederá algún número **razonable**.

Recapitulando, el Teorema 1 en Xu y Raginsky (2017) (Lema 5.1) sirve como una base para acotar el error de generalización esperado mediante la información

mutua. En este resultado se acota $\text{gen}(\mu, P_{W \mid S_n})$ mediante la información mutua entre la muestra y el algoritmo de aprendizaje; sin embargo, este acercamiento tiene sus limitantes. Debido a esto, en Bu, Zou y Veeravalli (2020) se mejora esta cota a través del Lema 5.2. Posteriormente, Wang y col. (2021) se apoyan de este lema y algunas propiedades de los canales de ruido Gaussiano para obtener una cota de generalización para el *SGLD*.

Es importante notar que, con la ayuda de los capítulos de teoría preliminar, se logró enunciar y probar estos resultados limitándonos casi completamente a proposiciones enunciadas en este texto. Principalmente, este capítulo es el desenlace de una historia que narra cómo deducir algunas aportaciones de Wang y col. (2021) utilizando el conocimiento adquirido en una maestría con enfoque en probabilidad y estadística.

Capítulo 6

Conclusiones

Entre estas páginas se presentó la teoría necesaria para estudiar, en toda generalidad, problemas de aprendizaje estadístico supervisado y se provee de una bibliografía suficiente para ahondar y dominar estos conceptos. Con ayuda de tales cimientos, y mucho esfuerzo, se pueden estudiar minuciosamente las propiedades de algoritmos de aprendizaje automático modernos y sofisticados. Sin embargo, al trabajar con estas metodologías altamente complejas es necesario recurrir a otras ramas de las matemáticas. Particularmente, en este trabajo se muestra cómo Wang y col. (2021) se apoyan de la teoría de la información y el transporte óptimo para describir la capacidad de generalización del descenso por gradiente de la dinámica de Langevin (*SGLD*).

Con base en lo anterior, es sencillo notar que la teoría de la información brinda herramientas clave para analizar problemas estadísticos y aporta una nueva perspectiva desde la cual estudiar preguntas aún sin resolver; por ejemplo, la impresionante capacidad de generalización de las redes neuronales. Con esto en mente, este texto aporta una introducción con fuertes bases probabilísticas a esta disciplina. En consecuencia, la teoría aquí desarrollada puede resultar más clara, para aquellos estudiantes que hayan tomado un curso avanzado de probabilidad, que la teoría presente en textos introductorios donde sólo se trabaja con variables aleatorias discretas o absolutamente continuas.

Finalmente, se logró presentar, de manera autocontenida, un resultado contemporáneo de un artículo de investigación (Wang y col. 2021) relevante para entender mejor el comportamiento de un algoritmo de aprendizaje de máquina. Más aún, en el capítulo correspondiente se discuten un poco los resultados del artículo. Asimismo, se menciona que la cota para el error de generalización esperado obtenida en éste no es justa. De tal manera que aún queda camino por recorrer para obte-

ner cotas de esta índole que brinden resultados numéricos útiles para la toma de decisiones y que no solamente describan a grandes rasgos un fenómeno.

Referencias

Apostol, T. (2014). *Calculus II*. Editorial Reverté.

Bu, Y., S. Zou y V. Veeravalli (2020). «Tightening Mutual Information-Based Bounds on Generalization Error». En: *IEEE Journal on Selected Areas of Information Theory* 1.1, págs. 121-130.

Çinlar, E. (2011). *Probability and Stochastics*. Graduate Texts in Mathematics. Springer.

Cohen, J., J.H.B. Kempermann y G. Zbaganu (1998). *Comparisons of Stochastic Matrices with Applications in Information Theory, Statistics, Economics and Population*. Springer.

Duchi, J. (2019). *Lecture Notes for Statistics 311/Electrical Engineering 377*. Stanford University.

Dudley, R. (1999). *Uniform Central Limit Theorems*. Cambridge University Press.

Gelfand, S.B. y S.K. Mitter (1991). «Recursive stochastic algorithms for global optimization». En: *SIAM Journal on Control and Optimization* 29.5, págs. 999-1018.

Goebel, B., Z. Dawy, J. Hagenauer y J.C. Mueller (2005). «An approximation to the distribution of finite sample size mutual information estimates». En: *IEEE International Conference on Communications, 2005. ICC 2005. 2005.* Vol. 2, 1102-1106 Vol. 2.

Hazan, E. (2016). «Introduction to online convex optimization». En: *Foundations and Trends® in Optimization* 2.3-4, págs. 157-325.

Jacod, J. y P. Protter (2004). *Probability Essentials*. Springer.

Kallenberg, O. (2002). *Foundations of Modern Probability*. 2.ª ed. Springer.

Li, C., C. Chen, D. Carlson y L. Carin (2016). «Preconditioned stochastic gradient Langevin dynamics for deep neural networks». En: *Thirtieth AAAI Conference on Artificial Intelligence*.

Negrea, J., M. Haghifam, G.K Dziugate, A. Khisti y D.M. Roy (2019). «Information theoretic generalization bounds for SGLD via data-dependent estimates». En: *Advances in Neural Information Processing Systems*, págs. 11015-11025.

Nesterov, Y. (2004). *Introductory Lectures on Convex Optimization*. Applied Optimization. Springer.

Niculescu, C. y L. Persson (2018). *Convex Functions and Their Applications: A contemporary approach*. 2.ª ed. Springer.

Polyanskiy, Y. e Y. Wu (2016). «Dissipation of Information in Channels With Input Constraints». En: *IEEE Transactions on Information Theory,* 62.1.

Polyanskiy, Y. e Y. Wu (2017). «Strong data-processing inequalities for channels and Bayesian networks». En: *Convexity and Concentration*. Springer, págs. 211-249.
— (2019). *Lecture Notes on Information Theory*. Lecture Notes for 6.441 (MIT), ECE 563 (UIUC), STAT 364 (Yale).
Raginsky, M. (2016). «Strong Data Processing Inequalities and Φ-Sobolev Inequalities for Discrete Channels». En: *IEEE Transactions on Information Theory,* 62.6.
Ranginsky, M. e I. Sason (2012). «Concentration of Measure Inequalities in Information Theory». En: *arXiv preprint arXiv:2102.02976v1*.
Rockafellar, J. (1997). *Convex Analysis*. Princeton University Press.
Seber, G. y A. Lee (2003). *Linear Regression Analysis*. 2.ª ed. John Wiley & Sons.
Shalev-Schwarz, S. y S. Ben-David (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press.
Spivak, M. (2014). *Calculus. Cuarta Edición*. Editorial Reverté.
Villani, C. (2009). *Optimal Transport: Old and New*. Vol. 338. A Series of Comprehensive Studies in Mathematics. Springer-Verlag.
Wang, H., Y. Huang, R. Gao y F. Calmon (2021). «Learning While Dissipating Information: Understanding the Generalization Capability of SGLD». En: *arXiv preprint arXiv:2102.02976v1*.
Welling, M. e Y. W. Teh (2011). «Bayesian learning via stochastich gradient Langevin dynamics». En: *Proceedings of the 28th international conference on machine learning (ICML)*, págs. 681-688.
Xu, A. y M. Raginsky (2017). «Information-theoretic analysis of generalization capability of learning algorithms». En: *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, págs. 2524-2533.
Zbili, M. y S. Rama (2021). «A quick and easy way to estimate entropy and mutual information for neuroscience». En: *Frontiers in Neuroinformatics* 15, pág. 25.
Zhang, C., S. Bengio, M. Hardt, B. Recht y O. Vinyals (2021). «Understanding Deep Learning (Still) Requires Rethinking Generalization». En: *Communications of the ACM* 64.3.