UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

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

PRIVACIDAD EN ALGORITMOS ICR

TESIS

QUE PARA OPTAR POR EL GRADO DE:

MAESTRO EN CIENCIAS MATEMÁTICAS

PRESENTA:

ALEJANDRO ANTONIO ESTRADA FRANCO

DIRECTORA:

DRA. SANDRA PALAU CALDERÓN

I.I.M.A.S.

CIUDAD DE MÉXICO, NOVIEMBRE 2025

Índice general

Introducción

Dentro del universo de los datos cada vez toma mayor importancia la palabra *privacidad* la cuál se concibe como la necesidad que el usuario de distintos sistemas y aplicaciones tiene de que sus datos personales no puedan ser conocidos o vistos sin su consentimiento. La realidad es que para hacer uso de la mayoría de las aplicaciones que se conectan a internet el usuario debe otorgar acceso a datos sensibles, estos pueden incluir su ubicación, su nombre, su ocupación, lo que busca en los navegadores, sus compras a través de la red, etc. Grandes empresas tecnológicas desde sus centros de investigación han empujado iniciativas para lograr métodos que a la vez les permitan usar datos de los usuarios para distintos fines, pero garantizando a estos privacidad.

Estos métodos son aquellos enmarcados dentro del concepto de *privacidad diferencial*. Un área crítica en la que datos son usados es el *aprendizaje máquina* (machine learning), aquí se cuenta con una amplia gama de modelos estadísticos que se usan principalmente para predecir y/o clasificar. De manera muy general podemos decir que en la mayoría de los modelos de machine learning se busca calcular los parámetros del modelo a través de un conjunto de entrenamiento minimizando una función que mide el error de las predicciones. La familia de algoritmos de *descenso del gradiente* proporciona un gran surtido de métodos para minimizar la función de error o pérdida.

Los datos atraviesan este proceso de optimización, y en determinados contextos se han desarrollado técnicas que son capaces de vulnerar la privacidad de los datos, aún después de que estos han sido procesados por estos algoritmos.

La tesis se desarrolla en torno al tema de privacidad diferencial dentro de algoritmos iterativos los cuales generalizan cierto tipo de algoritmos *descenso del gradiente*. En particular se estudia el **teorema 22** de [Fel+18] el cual proporciona garantías de privacidad para los algoritmos citados.

En el capítulo 1 se introducen los conceptos básicos usados en el enfoque matemático sobre la privacidad en bases de datos, así como una breve reseña histórica del abordaje de este problema.

En el capítulo 2 se abordan la definición de privacidad diferencial así como las herramientas matemáticas usadas en su estudio.

En el capítulo 3 se da un repaso de la divergencia de Renyi se introducen sus motivaciones así como resultados que son usados en capítulos posteriores.

El capítulo 4 trata de la distancia Wasserstein, en particular para este trabajo se usa la distancia Wasserstein infinito, se revisan los conceptos básicos y resultados útiles.

Dentro del capítulo 5 se exponen los teoremas principales, aquellos que proporcionan garantías de privacidad para algoritmos iterativos.

Finalmente en los apéndices A y B se repasan resultados y conceptos de medida, probabilidad y conjuntos convexos que son importantes dentro del resto de los capítulos.

Capítulo 1

Privacidad

1.1. Introducción

El concepto matemático de privacidad surge en medio del tratamiento computacional de datos. Por ejemplo:

1. **Análisis estadístico de datos:** Cuando empresas o instituciones gubernamentales publican estadísticas agregadas (promedios, conteos, histogramas), la aplicación de la privacidad permite hacerlo sin riesgo de re-identificación.

2. **Aprendizaje estadístico (Machine Learning) usando datos sensibles:** Al usar datos personales para entrenar modelos de *Machine Learning* (historiales de salud, imágenes médicas, textos escritos por usuarios, etc.) se pone en peligro la privacidad de los usuarios.

3. **Recolección de datos del usuario:** En aplicaciones móviles y/o en navegadores web se recolectan datos del usuario para mejorar productos o personalizar recomendaciones. En este caso también la privacidad es vulnerable.

La privacidad computacional se ha formulado conceptualmente de distintas maneras según los problemas que se intenten resolver. Ésta surgió como respuesta a las, cada vez mayores, preocupaciones sobre la exposición y uso de datos personales en una época caracterizada por la explotación masiva de información y los límites de las técnicas tradicionales de anonimización. En los años noventa Latanya Sweeney propone la técnica del k-anonimato, sin embargo, investigaciones posteriores demostraron que, incluso, tras eliminar identificadores directos (como nombres); la combinación de ciertos datos (como edad, código postal y sexo) podía permitir re-identificar personas con sorprendente facilidad. Fue así como surgió la necesidad de un nuevo paradigma de privacidad que no dependiera de ocultar o transformar datos, sino de garantías matemáticas robustas.

La definición formal (matemática) de privacidad diferencial fue propuesta por Cynthia Dwork, Frank McSherry, Kobbi Nissim y Adam Smith, en 2006, en el artículo *Calibrating Noise to Sensitivity in Private Data Analysis* publicado en *Proceedings of the Third Theory of Cryptography Conference (TCC 2006)*. Su propuesta en lugar de buscar que ciertos datos pudieran anonimizarse por completo, se propuso lograr medir el cambio en la salida de un algoritmo cuando se agrega o elimina un elemento

del conjunto de datos. Así, se podía asegurar que ningún resultado revelara información sensible sobre una persona en particular, sin importar cuánto conocimiento previo tuviera un atacante.

En 2014 Google implementó privacidad diferencial en su sistema *RAPPOR (Randomized Aggregatable Privacy-Preserving Ordinal Response)* este es una aplicación que recopila estadísticas de uso del navegador Chrome de los usuarios, la implementación de privacidad diferencial tenía como objetivo conseguir información agregada (matemáticamente hablando estadísticos) sin revelar datos individuales de los usuarios, para ello *RAPPOR* implementa una forma de privacidad diferencial local donde los datos son aleatorizados en el dispositivo personal antes de ser enviados a los servidores. En 2016 Apple implementó privacidad diferencial en el sistema operativo iOS 10 con el objetivo de recolectar información de los usuarios de forma útil y segura a la vez para implementar mejoras en las sugerencias de texto, identificar errores ortográficos, etc. de igual manera fue usada la técnica de privacidad diferencial local. Como estos ejemplos hay otros y se han ido multiplicando.

La privacidad diferencial se ha convertido en el estándar de referencia en la protección de datos, influyendo actualmente incluso en el desarrollo de políticas públicas.

1.2. Conceptos básicos

Llamamos base de datos a un conjunto de n-tuplas $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, donde para cada $i = 1, \ldots, n$, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})$. Pensamos en una tabla del siguiente estilo

	Columna 1	Columna 2	...	Columna k
$\mathbf{x}_1$	x_{11}	x_{12}	...	x_{1k}
$\mathbf{x}_2$	x_{21}	x_{22}	...	x_{2k}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\mathbf{x}_n$	x_{n1}	x_{n2}	...	x_{nk}

Cuadro 1.1: Base de datos.

Cada x_{ij} es algún tipo de dato computacional (flotante, entero, texto, booleano, etc.). Una base de datos puede ser incluso un conjunto de tablas, es decir cada $\mathbf{x}_i$ puede ser pensado como una matriz con entradas en algún tipo de dato computacional.

Una base de datos estadística es aquella en la cual no se puede consultar a los datos particulares, solo se pueden hacer consultas agregadas es decir; únicamente se tiene acceso a información estadística sobre los datos (media, varianza, conteos, moda, etc.), en este contexto se denomina curador de los datos al sujeto que puede observar los datos particulares.

Una base de datos estadística no es protección suficiente a la privacidad de la información individual de la base, por ejemplo, usando el promedio en subconjuntos diferenciados por un dato es posible obtener información. En la tabla del cuadro 1

supongamos que la columna 1 es numérica, usando el promedio para $j = 1, \ldots, s$ y para $j = 1, \ldots, s+1$ se puede obtener

$$x_{(s+1)j} = (s+1)\left(\frac{1}{s+1}\sum_{j=1}^{s+1} x_{j1}\right) - s\left(\frac{1}{s}\sum_{j=1}^{s} x_{j1}\right);$$

esta en realidad es una consulta muy sencilla, por ejemplo en SQL:

```
SELECT
(s+1)*(SELECT
       AVG(Columna 1)
       FROM tabla
       WHERE idx < s+2)
-
s*(SELECT
   AVG(Columna 1)
   FROM tabla
   WHERE idx < s+1).
```

La idea de privacidad diferencial es agregar ruido al resultado de las consultas para que no sea posible distinguir de que tabla procede pero sin que pierda utilidad para propósitos estadísticos.

Capítulo 2

Privacidad Diferencial

2.1. Generalidades

Pensemos en una base de datos como un conjunto X no vacío cuyos elementos son n-tuplas. Una consulta es una función f que toma subconjuntos de la base X y les asigna algún objeto s. Por ejemplo, pensemos en la siguiente base de datos

$$X = \{(x_1, x_2, x_3) | x_1 \in \{0.25, 0.5, 0.75, 1\}, x_2 \in \{A, B, C, D\}, x_3 = \{0, 1\}\}.$$

Sea $A \subseteq X$. La función $f : \{0,1\}^X \to \{\text{VERDADERO}, \text{FALSO}\}$ dada por:

$$f(A) = \begin{cases} \text{VERDADERO} & si \quad x_1 > 0.25 \ y \ x_2 = B, \\ \text{FALSO} & en \ otro \ caso. \end{cases}$$

es una consulta.

Para fines de privacidad se busca aleatorizar la salida de las consultas a la base de datos introduciendo cierto nivel de ruido que permita que los resultados proporcionen información útil pero sin que se pueda determinar el dato preciso al hacer comparaciones entre resultados sobre distintos subconjuntos de la base. Para esto introducimos las siguientes definiciones. Denotamos como $\{0,1\}^X$ al conjunto potencia de X.

Definición 2.1. *Sea X un conjunto no vacío. Consideremos una distancia $d : \{0,1\}^X \times \{0,1\}^X \to \mathbb{R}$. Dos conjuntos D, D' se dicen p-adyacentes o p-vecinos con respecto a d si $d(D, D') \leq p$. Para $p = 1$ diremos simplemente conjuntos vecinos o adyacentes.*

Definición 2.2. *Sea $(\Omega, \mathcal{F}, \mathbb{P})$ un espacio de probabilidad, $E \neq \emptyset$ y $(G, \mathcal{G})$ un espacio medible. Un algoritmo aleatorizado de E en G es una función $\mathcal{A} : \{0,1\}^E \times \Omega \to G$ tal que para todo $A \in \{0,1\}^E$, $\mathcal{A}(A, \cdot)$ es una función medible. Denotada como $\mathcal{A}(A)$.*

Para cada $A \subseteq E$ tenemos una distribución $\mathbb{P}_{\#A}(\cdot) = \mathbb{P}[\mathcal{A}(A) \in \cdot]$. Deseamos que al observar alguna salida s del algoritmo no sea posible determinar de que distribución fue muestreado, esto resultará posible si es difícil distinguir entre distribuciones generadas por conjuntos p-vecinos (para algún p). Entonces necesitamos un número que nos permita medir qué tanto "se parecen" dos distribuciones.

Pero algo más, ya que las técnicas para obtener datos de individuos en una base estadística se fundamentan en aplicar distintas consultas a subconjuntos vecinos de la base, requerimos que las consultas aleatorizadas se mantengan indistinguibles al aplicarse individualmente así como al combinarse. Llamamos posprocesamiento a esta propiedad. Demos formalidad matemática a esto en la siguiente definición.

Definición 2.3. *Sea $(\Omega, \mathcal{F}, \mathbb{P})$ un espacio de probabilidad, $(G, \mathcal{G})$, $(H, \mathcal{H})$ espacios medibles, $f : G \to H$ una función medible y $\mathcal{P}(G)$ el conjunto de las medidas de probabilidad sobre G, a una función*

$$D_{PM}(\cdot||\cdot) : \mathcal{P}(G)^2 \to \mathbb{R}^+$$

le llamamos medida de privacidad si cumple

$$D_{PM}(\mu_f||\nu_f) \leq D_{PM}(\mu||\nu)$$

para todo $\mu, \nu \in \mathcal{P}(G)$ y $f : G \to H$ medible.

Aquí nótese que $\mu_f(\cdot) = \mu(f^{-1}(\cdot))$ es una medida en $(H, \mathcal{H})$, Sin embargo también la podemos pensar como una medida en $\{f^{-1}(B) : B \in \mathcal{H}\} \subseteq \mathcal{P}(G)$, por lo cuál la expresión $D_{PM}(\mu_f||\nu_f)$ esta bien definida, en términos de evaluar elementos en $\mathcal{P}(G)$.

A la propiedad $D_{PM}(\mu_f||\nu_f) \leq D_{PM}(\mu||\nu)$ le llamamos posprocesamiento. Entonces decimos que $D_{PM}(\cdot||\cdot)$ es una medida de privacidad si cumple la propiedad de posprocesamiento.

Como ejemplo de medida de privacidad tenemos la medida clásica de privacidad diferencial:

$$D_\infty(\mu||\nu) = \sup_{B \in \mathcal{H}} \log \frac{\mu(B)}{\nu(B)}.$$

Ya que contamos con una manera de medir la privacidad podemos determinar que tanto una consulta o una combinación de consultas aleatorizadas mantienen la privacidad en una base estadística, es decir, establecer una manera de calibrar algoritmos en términos de privacidad.

Definición 2.4. *Sea $(\Omega, \mathcal{F}, \mathbb{P})$ un espacio de probabilidad, $E \neq \emptyset$, $(G, \mathcal{G})$ un espacio medible, d una distancia en $\{0,1\}^E$, D_{PM} una medida de privacidad en $\mathcal{P}(G)$. Un algoritmo aleatorizado $\mathcal{A}$ de E en G es $(\varepsilon_{IN}, \varepsilon_{OUT})$-privado si para todo par $D, D' \subseteq E$ tales que $d(D, D') \leq \varepsilon_{IN}$ se tiene $D_{PM}(\mathcal{A}(D)||\mathcal{A}(D')) \leq \varepsilon_{OUT}$.*

Equivalentemente un algoritmo aleatorizado $\mathcal{A}$ es $(\varepsilon_{IN}, \varepsilon_{OUT})$-privado si

$$\sup_{d(D,D') \leq \varepsilon_{IN}} D_{PM}(\mathcal{A}(D)||\mathcal{A}(D')) \leq \varepsilon_{OUT}.$$

En general con las bases estadísticas se busca hacer indistinguible el resultado de dos consultas aplicadas a conjuntos adyacentes (usando la distancia de diferencia simétrica), en estos casos $\varepsilon_{IN} = 1$, y decimos que $\mathcal{A}$ es ε-privado si cumple con la definición, donde $\varepsilon_{OUT} = \varepsilon$. Si dos conjuntos D, D' son adyacentes escribimos $D \sim D'$, ahora para $\varepsilon_{IN} = 1$, y $D_{PM} = D_\infty$ un algoritmo aleatorizado $\mathcal{A}$ es ε-privado si

$$\sup_{D \sim D'} \sup_{S \in \mathcal{G}} \log \frac{\mathbb{P}[\mathcal{A}(D) \in S]}{\mathbb{P}[\mathcal{A}(D') \in S]} \leq \varepsilon,$$

que es la definición clásica de privacidad.

2.2. Privacidad diferencial clásica

Definición 2.5 (Privacidad diferencial clásica)**.** *Un algoritmo aleatorizado $\mathcal{A}$ es ε-diferencial privado, si para todo par de conjuntos adyacentes D, $D' \in \{0,1\}^E$ y para todo $S \subseteq G$ se cumple:*

$$\log \frac{\mathbb{P}[\mathcal{A}(D) \in S]}{\mathbb{P}[\mathcal{A}(D') \in S]} \leq \varepsilon.$$

Para revisar ejemplos de algoritmos ε-privados es necesario introducir un par de conceptos.

Para una consulta $f : \{0,1\}^E \to \mathbb{R}^n$ la sensibilidad global de f está definida como

$$\triangle_{GS} f = \sup_{D \sim D'} ||f(D) - f(D')||_1.$$

Este número es importante en términos de algoritmos aleatorizados ε-privados, pues los algoritmos más usados consisten en sumar ruido a la consulta por ejemplo si este ruido es laplaciano, o gaussiano, la sensibilidad representa el desplazamiento de la media de la distribución de un algoritmo respecto de otro. Veamos el algoritmo laplaciano.

Sea $f = (f_1, \ldots, f_n)$ una consulta como la antes mencionada, definimos el mecanismo de Laplace como $\mathcal{A} : \{0,1\}^E \times \Omega \to \mathbb{R}^n$ tal que

$$\mathcal{A}(D, \omega) = f(D) + (X_1, \ldots, X_n)(\omega),$$

donde las X_i son independientes, idénticamente distribuidas con $X_i \sim Lap\left(0, \frac{n\triangle_{GS} f}{\varepsilon}\right)$ y $\varepsilon > 0$. Tenemos entonces que $\mathcal{A}(D)_i \sim Lap\left(f(D)_i, \frac{n\triangle_{GS} f}{\varepsilon}\right)$. Entonces, la función de densidad de $\mathcal{A}(D)$ está dada por

$$\rho_{\mathcal{A}(D)}(r) = \prod_{i=1}^{n} \frac{\varepsilon}{2n\triangle_{GS} f} \exp\left\{-\frac{\varepsilon |r_i - f(D)_i|}{n\triangle_{GS} f}\right\}, \quad r \in \mathbb{R}^n.$$

Tomamos ahora $\rho_{\mathcal{A}(D')}$ y hacemos el cociente

$$\frac{\rho_{\mathcal{A}(D)(r)}}{\rho_{\mathcal{A}(D')(r)}} = \prod_{i=1}^{n} \exp\left\{\frac{\varepsilon}{n\triangle_{GS} f}\left(|r_i - f(D')_i| - |r_i - f(D)_i|\right)\right\},$$

tomando en cuenta $|f(D)_i - f(D')_i| \geq |f(D')_i - r_i| - |f(D)_i - r_i|$, escribimos

$$\frac{\rho_{\mathcal{A}(D)(r)}}{\rho_{\mathcal{A}(D')(r)}} \leq \prod_{i=1}^{n} \exp\left\{\frac{\varepsilon}{n\triangle_{GS} f}\left(|f(D')_i - f(D)_i|\right)\right\}.$$

Luego $|f(D') - f(D)| \leq \triangle_{GS} f$, por tanto

$$\frac{\rho_{\mathcal{A}(D)(r)}}{\rho_{\mathcal{A}(D')(r)}} \leq \prod_{i=1}^{n} \exp\left\{\frac{\varepsilon}{n}\right\} = \exp\{\varepsilon\}.$$

Integrando, tenemos para todo $S \in \mathcal{B}(\mathbb{R}^n)$

$$\int_S \rho_{\mathcal{A}(D)}(r)dr \leq \exp\{\varepsilon\} \int_S \rho_{\mathcal{A}(D')}(r)dr.$$

Lo cual resulta en que

$$\sup_{S\in\mathcal{B}(\mathbb{R}^n)} \log \frac{\mathbb{P}[\mathcal{A}(D)\in S]}{\mathbb{P}[\mathcal{A}(D')\in S]} \leq \varepsilon.$$

Ejemplo 2.1. *Consideremos el conjunto*

$$D = \{88.68, 70.65, 38.45, 58.42, 29.32, 11.11, 60.37, 19.91, 52.04, 96.39\},$$

y $D' = D - \{96.39\}$, f *la consulta que obtiene el promedio,* $\mathcal{A}$ *el algoritmo aleatorizado que suma un ruido laplaciano* $X \sim Lap(0,1)$, *en la gráfica tenemos las distribuciones* $\mathcal{A}(D)$ *y* $\mathcal{A}(D')$.

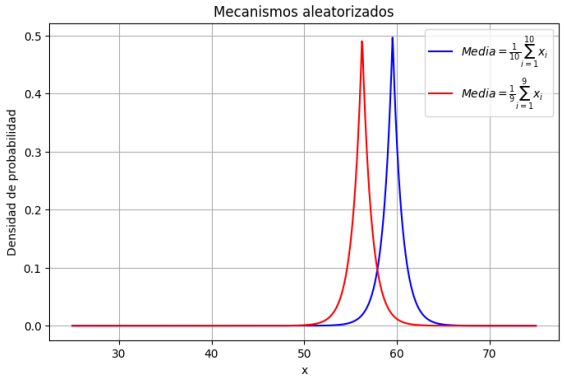

Figura 2.1: Distribuciones de los algoritmos $\mathcal{A}(D)$, $\mathcal{A}(D')$.

Otro algoritmo aleatorizado muy usado es el gaussiano, la idea es la misma que para el laplaciano, $\mathcal{A}(D) = f(D) + N(0, \mathbb{I}\sigma^2)$.

Capítulo 3

Divergencia de Renyi

3.1. Introducción

Para comprender la funcionalidad de la herramienta presentada en este capítulo, daremos un breve repaso a conceptos básicos de teoría de la información pensando en lectores poco familiarizados con el tema.

Supongamos que se tiene un sistema de mensajes codificados en matrices (a_{ij}) de 5×2, donde cada elemento de cada matriz es un cuadro, los cuadros toman colores como valores, pueden ser negros, amarillos, azules o rojos; los mensajes se envían en dígitos dibujados en la matriz. Tomemos por ejemplo un mensaje de dos dígitos:

Dígito 1: $a_{i,j} = b$ si y sólo si $j = 1$,

Dígito 2: $a_{i,j} = n$ si $j = 2$ o $i \in \{1, 3, 5\}$,

donde n quiere decir que el cuadro es negro, a que el cuadro es amarillo y b que el cuadro es azul y v cuadro verde.

Supongamos que este código se transmite a través de un canal de información que sólo transmite dos valores; por ejemplo 0 y 1, una manera de hacer llegar los datos correctos es usando preguntas, se inicia la comunicación y por cada cuadro de la matriz tendremos dos dígitos de ceros y unos, el primero corresponde a la pregunta: ¿Este cuadro es negro o amarillo? si la respuesta es afirmativa se pregunta ¿El cuadro es negro? si la respuesta no es afirmativa se pregunta ¿El cuadro es azul? No se necesitan más preguntas. Para cada pregunta el valor 1 es afirmativo y el 0 negativo. Así que para comunicar que el cuadro es negro se envía 11, para comunicar que es amarillo 10, para azul 01, para rojo 00. Para cada color necesitamos dos preguntas.

Podemos interpretar que el costo de transmitir un mensaje con este sistema es de dos preguntas con respuesta binaria por cuadro, en el lenguaje de teoría de la información, cada cuadro contiene 2 *bits* de información.

Consideremos el caso cuando el código está configurado de tal manera que en 16 cuadros transmitidos 8 veces tenemos negro, 4 veces amarillo, 2 azul y 2 rojo. Con esta formulación elegiremos otro sistema para decodificar la información, esta vez se impone un orden en la sucesión de las preguntas.

1.- ¿El color es negro?

2.- ¿Es amarillo?

3.- ¿Es azul?

El orden esta basado en la frecuencia de aparición de cada color en este sistema. Así para el color negro necesitamos una pregunta, para el amarillo 2, y para los colores azul y rojo 3. En promedio el número de preguntas en este sistema es el siguiente:

$$\frac{1}{2} * 1 + \frac{1}{4} * 2 + \frac{1}{8} * 3 + \frac{1}{8} * 3 = 1.75$$

Mientras que con el sistema anterior el promedio de preguntas es 2, en este caso diremos que el cuadro en el segundo sistema contiene $1.75\, bits$. Nótese que el número de preguntas binarias (con dos posibles respuesta únicamente) asociado a cada color está relacionado con la frecuencia relativa del mismo

$$\text{frecuencia relativa} = 2^{(\#\text{ preguntas})}.$$

En términos de probabilidades, supongamos que $X : \Omega \to E$ es una variable aleatoria donde E es a lo más numerable, con función de masa de probabilidad $p(x) = \mathbb{P}[X = x]$. La cantidad de información en $bits$ de X se define como:

$$H(X) = -\sum_{x \in E} p(x) \log_2 p(x).$$

Esta definición fue propuesta en 1948 por *Claude Shannon* [Sha48] este número se denomina también entropía de Shannon, o simplemente entropía.

Consideremos nuevamente una variable aleatoria X como arriba, esta vez pensándola sobreyectiva para hacer claro el argumento, el número de preguntas binarias determinado por la distribución de X para llegar al elemento x es $\log_2(\mathbb{P}[X = x])$ consideremos ahora otra variable aleatoria Y con el mismo dominio e imagen, también sobreyectiva, la distribución de Y impone un sistema de preguntas sobre E, es decir para llegar a $y \in Im(Y)$ necesitamos $\log_2(\mathbb{P}[Y = y])$ preguntas. Consideremos ahora una función Z tal que a cada elemento $x \in E$ le asigna su número de $bits$ con respecto a la distribución de Y. El valor esperado de preguntas para cada $x \in E$ con respecto a la distribución de X está dado por:

$$H(X; Y) := -\sum_{x \in E} p(x) \log_2 q(x).$$

Donde p es la masa de probabilidad de X y q la de Y. Llamamos a $H(X; Y)$ la entropía cruzada X a Y. Que podemos interpretar como la cantidad de información en bits en un sistema en el cual los elementos de E tienen una frecuencia relativa dada por la distribución de X con un sistema de preguntas diseñado con la distribución de Y. Tenemos, también la entropía relativa de X a Y:

$$D_{KL}(X||Y) = H(X; Y) - H(X).$$

Que se interpreta cómo la pérdida de información al establecer un sistema de preguntas dado por una distribución distinta a la real de los datos. La notación es en

honor a *S. Kullback* y *R. A. Leibler* quienes propusieron esta definición en 1951 [S K51].
Redefinamos la entropía, la entropía cruzada y la entropía relativa en términos de distribuciones de probabilidad y cambiando el logaritmo base 2 (debido a la naturaleza binaria de las preguntas) por logaritmo natural; tenemos para distribuciones de probabilidad P, Q:

- $H(P) := -\sum_x P(x) \log P(x)$.
- $H(P;Q) := -\sum_x P(x) \log Q(x)$.
- $D_{KL}(P||Q) := H(P;Q) - H(P) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$.

En términos de distribuciones se hace evidente que para que se pueda calcular la entropía relativa necesitamos que $Q(x) = 0$ siempre que $P(x) = 0$, es decir $Q \ll P$.
Una generalización de la entropía relativa se da por la siguiente expresión:

$$D_\alpha(P||Q) = \frac{1}{\alpha - 1} \log \left[\sum_x \left(\frac{P(x)}{Q(x)} \right)^\alpha Q(x) \right].$$

Para $\alpha \in (0,1) \cup (1,\infty)$, y distribuciones P, Q, tales que $Q \ll P$. Tenemos que $\lim_{\alpha \to 1} D_\alpha(P||Q) = D_{KL}(P||Q)$([Erv10] *teorema 5*). La interpretación de esta especie de nueva medida de entropía es en el mismo sentido que en el de la entropía relativa; sólo que aquí el parámetro α nos permite calibrar la sensibilidad de la medida; conforme α es mayor, es mas sensible a diferencias entre las distribuciones. Esta medida de entropía recibe el nombre de divergencia de Renyi, al fijar dos distribuciones esta es una función creciente con respecto a α.

Ejemplo 3.1. *Consideremos dos distribuciones:*
$p = (0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1)$,
$q = (0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.11, 0.09)$.

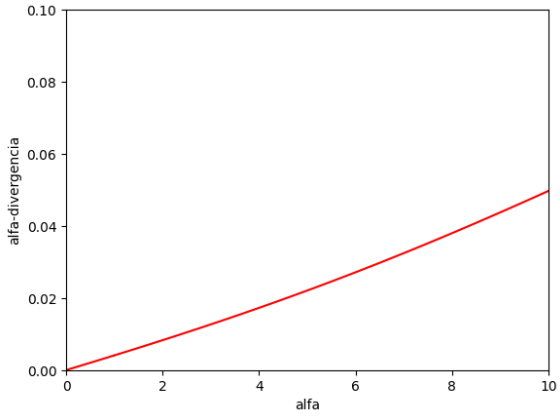

Figura 3.1: Valor de la divergencia de Renyi $D_\alpha(p||q)$ en función del parámetro α

$D_\alpha(P||Q)$ recibe su nombre en honor a *Alfred Renyi* quien la propuso por primera vez en 1961 [Ren61]. Veamos ahora una definición más general de esta herramienta en términos de teoría de la medida.

Definición 3.1 (Divergencia de Renyi)**.** *Considérese un espacio medible* $(\Omega, \mathcal{F})$*,* $\mathfrak{m}, \nu, \mu$ *medidas sobre* $(\Omega, \mathcal{F})$ *tales que* $\mu \ll \nu \ll \mathfrak{m}$*, sea* $\alpha \in (0,1) \cup (1, \infty)$*. Se define la* α*-divergencia de Renyi entre* μ *y* ν *como:*

$$D_\alpha(\mu||\nu) = \frac{1}{\alpha - 1} \log \int \left(\frac{d\mu}{d\mathfrak{m}}\right)^\alpha \left(\frac{d\nu}{d\mathfrak{m}}\right)^{1-\alpha} d\mathfrak{m}.$$

Observación 3.1. *Dentro de la integral estamos haciendo referencia a la derivada de Radón-Nikodym.*

Observación 3.2. *Ya que*

$$\frac{(d\mu/d\mathfrak{m})}{(d\nu/d\mathfrak{m})} = \frac{d\mu}{d\nu},$$

Podemos escribir usando cambio de variable:

$$D_\alpha(\mu||\nu) = \frac{1}{\alpha - 1} \log \int \left(\frac{d\mu}{d\nu}\right)^\alpha d\nu.$$

Una importante observación en el contexto de privacidad diferencial es que

$$\lim_{\alpha \to \infty} D_\alpha(\mu||\nu) = \sup_{D \in \mathcal{F}} \log \frac{\mu(D)}{\nu(D)}.$$

Tenemos pues que $\lim_{\alpha\to\infty} D_\alpha(\mu||\nu) = D_\infty(\mu||\nu)$, así este límite de la divergencia de Renyi coincide con la medida clásica de privacidad diferencial. En adelante consideraremos $\alpha > 1$ que es el parámetro usado en privacidad diferencial.

3.2. Propiedades

Lema 3.1 (Preprocesamiento)**.** *[Erv10] Sea* $(\Omega, \mathcal{F})$ *espacio de medida,* μ*,* ν *medidas sobre este espacio;* $\mathcal{G}$ *una sub-sigma álgebra de* $\mathcal{F}$*. Si se denotan* $\mu|_\mathcal{G}$*,* $\nu|_\mathcal{G}$ *las correspondientes restricciones, entonces se cumple:*

$$D_\alpha(\mu|_\mathcal{G} \,||\, \nu|_\mathcal{G}) \leq D_\alpha(\mu||\nu).$$

Demostración.

$$D_\alpha(\mu|_\mathcal{G} \,||\, \nu|_\mathcal{G}) = \frac{1}{\alpha - 1} \log \int \left(\frac{d\mu|_\mathcal{G}}{d\nu|_\mathcal{G}}\right)^\alpha d\nu.$$

Por el lema A.9 del apéndice continuamos con

$$D_\alpha(\mu|_\mathcal{G} \,||\, \nu|_\mathcal{G}) = \frac{1}{\alpha - 1} \log \int \left(E\left[\frac{d\mu}{d\nu} |\mathcal{G}\right]\right)^\alpha d\nu.$$

Usando la desigualdad de Jensen

$$D_\alpha(\mu|_\mathcal{G} \,||\, \nu|_\mathcal{G}) \leq \frac{1}{\alpha - 1} \log \int E\left[\left(\frac{d\mu}{d\nu}\right)^\alpha |\mathcal{G}\right] d\nu.$$

Finalmente por definición de esperanza condicional,

$$D_\alpha\left(\mu|_{\mathcal{G}} \,||\, \nu|_{\mathcal{G}}\right) = \frac{1}{\alpha-1}\log\int\left(\frac{d\mu}{d\nu}\right)^\alpha d\nu$$
$$= D_\alpha\left(\mu||\nu\right).$$

□

Observación 3.3. *El preprocesamiento nos dice que la pérdida de información de una distribución ν a otra μ, cuando estas están sobre subfamilias de la σ-álgebra original de conjuntos medibles, no rebasa la pérdida en el conjunto de datos completo.*

La siguiente proposición presenta dos propiedades importantes de la divergencia de *Renyi*. En la primera; si tenemos dos conjuntos de datos S_1, y S_2, donde actuando sobre el primero tenemos las distribuciones μ y ν, y sobre el segundo μ' y ν' en el conjunto de datos $S_1 \otimes S_2$ la pérdida de información de $\nu \times \nu'$ a $\mu \times \mu'$ es igual a sumar las perdidas de información de ν a μ y de ν' a μ' en S_1 y en S_2 respectivamente. Con respecto a la segunda propiedad llamada *Posprocesamiento* tenemos que la pérdida de información no se incrementa al darle un procesamiento a los datos a través de una función determinista f.

Proposición 3.1. *Lo siguiente se cumple para todo $\alpha \in (1,\infty)$, y distribuciones μ, μ', ν, ν':*

- ***Aditividad:*** $D_\alpha(\mu \times \mu' || \nu \times \nu') = D_\alpha(\mu||\nu) + D_\alpha(\mu'||\nu')$.
- ***Posprocesamiento:*** *Para cualquier función determinista f,*

  $$D_\alpha\left(f(\mu)||f(\nu)\right) \leq D_\alpha(\mu||\nu),$$

  donde $f(\mu)$ es la distribución de $f(X)$ con $X \sim \mu$.

Demostración. ▪ **Aditividad:**

$$\begin{aligned}
D_\alpha\left(\mu \times \mu' || \nu \times \nu'\right) &= \frac{1}{\alpha-1}\log\int\left(\frac{d(\mu\times\mu')/d(\mathfrak{m}\times\mathfrak{m})}{d(\nu\times\nu')/d(\mathfrak{m}\times\mathfrak{m})}\right)^\alpha \frac{d(\nu\times\nu')}{d(\mathfrak{m}\times\mathfrak{m})} d(\mathfrak{m}\times\mathfrak{m}) \\
&= \frac{1}{\alpha-1}\log\int\left(\frac{d(\mu\times\mu')}{d(\nu\times\nu')}\right)^\alpha \frac{d(\nu\times\nu')}{d(\mathfrak{m}\times\mathfrak{m})} d(\mathfrak{m}\times\mathfrak{m}) \\
&= \frac{1}{\alpha-1}\log\int\left(\frac{d\mu}{d\nu}\right)^\alpha\left(\frac{d\mu'}{d\nu'}\right)^\alpha \frac{d\nu}{d\mathfrak{m}}\frac{d\nu'}{d\mathfrak{m}} d(\mathfrak{m}\times\mathfrak{m}) \\
&= \frac{1}{\alpha-1}\log\left[\int\left(\frac{d\mu}{d\nu}\right)^\alpha \frac{d\nu}{d\mathfrak{m}} d\mathfrak{m} \int\left(\frac{d\mu'}{d\nu'}\right)^\alpha \frac{d\nu'}{d\mathfrak{m}} d\mathfrak{m}\right] \\
&= \frac{1}{\alpha-1}\log\int\left(\frac{d\mu}{d\nu}\right)^\alpha \frac{d\nu}{d\mathfrak{m}} d\mathfrak{m} + \frac{1}{\alpha-1}\log\int\left(\frac{d\mu'}{d\nu'}\right)^\alpha \frac{d\nu'}{d\mathfrak{m}} d\mathfrak{m} \\
&= D_\alpha(\mu||\nu) + D_\alpha(\mu'||\nu').
\end{aligned}$$

- **Posprocesamiento:**
  Aquí usaremos la propiedad **preprocesamiento**; que dice lo siguiente:
  Si P, Q son distribuciones sobre el espacio medible $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ con $\mathcal{G} \leq \mathcal{F}$. Entonces:
  $$D_\alpha(P|_\mathcal{G} \,||\, Q|_\mathcal{G}) \leq D_\alpha(P||Q).$$
  Tenemos para $B \in \mathcal{B}(\mathcal{Z})$, $f(\mu)[B] = \mu[f^{-1}B]$ donde μ es la distribución de X. Podemos poner entonces:
  $$D_\alpha(f(\mu)||f(\nu)) = D_\alpha(\mu|_{\sigma(f)}||\nu|_{\sigma(f)}) \leq D_\alpha(\mu||\nu).$$

□

La aditividad es válida para cualquier número de factores en las entradas de la divergencia de Renyi. Es decir para todo número natural N, para distribuciones $\mu_1, \mu_2, \ldots, \mu_N$, $\nu_1, \ldots, \nu_N$ sobre conjuntos de datos $S_1, \ldots, S_N$ consideramos las distribuciones $\bigotimes_{n=1}^N \mu_i$, $\bigotimes_{n=1}^N \nu_i$ sobre el conjunto de datos $\bigotimes_{n=1}^N S_i$. Bajo estas condiciones tenemos:

$$D_\alpha\left(\bigotimes_{n=1}^N \mu_i \,\middle|\middle|\, \bigotimes_{n=1}^N \nu_i\right) = \sum_{n=1}^N D_\alpha\left(\mu_i||\nu_i\right).$$

La prueba de esto es por inducción, el paso inductivo es una calca de la prueba que se ha hecho para la propiedad de aditividad en el caso de dos distribuciones.

Observación 3.4. *Dado que la divergencia de Renyi cumple con la propiedad de posprocesamiento, tenemos que es una medida de privacidad.*

Capítulo 4

Distancia Wasserstein

La distancia de Wasserstein surge como una herramienta importante en el análisis de la privacidad diferencial de Renyi. A diferencia de otras funciones que comparan distribuciones, la distancia de Wasserstein captura no solo las diferencias en masa entre distribuciones, sino también el "costo" de transportar esa masa, lo que resulta útil al analizar algoritmos estocásticos en términos de su estabilidad frente a perturbaciones. En el contexto de la privacidad de Renyi, esta distancia se utiliza para acotar la divergencia entre salidas inducidas por bases de datos vecinas, ofreciendo una forma de medir la fuga de información. Además, su simetría y su compatibilidad con técnicas de optimización convexa la convierten en una herramienta muy útil para analizar mecanismos privados con mejores garantías de privacidad.

4.1. Acoplamientos y transporte

Definición 4.1. *Consideremos* $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$, $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$ *espacios de probabilidad* $(\mathcal{Z}_1, \mathcal{E}_1)$, $(\mathcal{Z}_2, \mathcal{E}_2)$ *espacios medibles y* $X_1 : \Omega_1 \to \mathcal{Z}_1$, $X_2 : \Omega_2 \to \mathcal{Z}_2$ *objetos aleatorios. Por un acoplamiento entre* X_1 *y* X_2 *entendemos un objeto aleatorio* $(\hat{X}_1, \hat{X}_2)$ *sobre el espacio* $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{\mathbb{P}})$ *que toma valores en el espacio medible* $(\mathcal{Z}_1 \times \mathcal{Z}_2, \mathcal{E}_1 \otimes \mathcal{E}_2)$*, y que además cumple:*

$$\hat{X}_1 \overset{d}{=} X_1 \qquad y \qquad \hat{X}_2 \overset{d}{=} X_2.$$

Pensemos en las distribuciones $\mu(\cdot) = \mathbb{P}_1(X_1 \in \cdot)$, $\nu(\cdot) = \mathbb{P}_2(X_2 \in \cdot)$ y $\gamma(\cdot) = \hat{\mathbb{P}}((\hat{X}_1, \hat{X}_2) \in \cdot)$, bajo las condiciones de la definición de arriba decimos que γ es un acoplamiento de μ y ν y escribimos $\gamma \in \Gamma(\mu, \nu)$. Donde $\Gamma(\mu, \nu)$ es el conjunto de todos los acoplamientos entre μ y ν, el cuál no es vacío porque siempre esta la medida producto $\mu \times \nu$.

Los acoplamientos también son llamados ***planes de transferencia***. En este contexto γ se piensa como una medida de la "transferencia" de conjuntos en $\mathcal{Z}_1$ hacia $\mathcal{Z}_2$. Estamos pensando aquí que se esta "transportando", o también transformando el conjunto $\mathcal{Z}_1$ en el conjunto $\mathcal{Z}_2$, y γ mide como se reparte $\mathcal{Z}_1$ en $\mathcal{Z}_2$, es decir $\gamma(A \times \mathcal{Z}_2) = \mu(A)$ dice cuanto de A llega a $\mathcal{Z}_2$, $\gamma(\mathcal{Z}_1 \times B) = \nu(B)$ dice cuanto de $\mathcal{Z}_1$ se ha puesto en B, y $\gamma(A \times B)$ cuanto de A se ha puesto en B.

Ejemplo 4.1. *Supongamos que se tienen los conjuntos* $U = \{x_1, x_2, x_3, x_4\}$ *y* $V = \{y_1, y_2\}$ *con distribuciones* $p_{_U} = (0.75, 0.10, 0.10, 0.05)$, $p_{_V} = (0.75, 0.25)$ *un acopla-*

miento es la medida producto $p_{_U} \times p_{_V}$. Este ***plan de transferencia*** *queda esquematizado en la siguiente tabla:*

U/V	y_1	y_2	V
x_1	0.5625	0.1875	0.75
x_2	0.075	0.025	0.10
x_3	0.075	0.025	0.10
x_4	0.0375	0.0125	0.05
U	0.75	0.25	1

Cuadro 4.1: En este caso el plan de transferencia queda determinado por la cantidad de x_i que se pondrá en y_j

.

La cantidad de x_1 que se esta transfiriendo a V se reparte sobre los elementos de V, en este caso y_1, y_2, según el plan de transferencia dado por $p_U \times p_V$, es decir la cantidad de x_1 transferida a y_1 es

$$(p_{_U} \times p_{_V})(\{x_1\} \times \{y_1\}) = p_{_U}(x_1)p_{_V}(y_1) = 0.5625,$$

mientras que la cantidad de x_1 transferida a y_2 es

$$(p_{_U} \times p_{_V})(\{x_1\} \times \{y_2\}) = p_{_U}(x_1)p_{_V}(y_2) = 0.1875.$$

Para cualquier plan de transferencia p la cantidad de x_1 que se transfiere a V es la misma: $p(x_1 \times V) = p_{|_U}(x_1)$=0.75.

La noción de ***transporte*** se construye a partir del concepto de ***plan de transferencia***, consideremos el mismo conjunto del ejemplo 3.1, ahora introduciremos una función $c : U \times V \to \mathbb{R}^+$ que se puede pensar como el costo de transportar la medida; por ejemplo $c(x_1, y_2)$ es el costo de transportar la medida de x_1 a y_2. De esta manera podemos obtener la media del costo ponderada por la distribución (plan de transferencia) $p_{_U} \times p_{_V}$:

$$\sum_{i,j} c(x_i, y_j)(p_{_U} \times p_{_V})(x_i, y_j).$$

Al par de medidas $p_{_U}$, $p_{_V}$ se le asocia el número:

$$\inf_{\gamma \in \Gamma(p_{_U}, p_{_V})} \sum_{i,j} c(x_i, y_j)\gamma(x_i, y_j).$$

A este número lo pensamos como la manera óptima de transportar la medida $p_{_U}$ en la medida $p_{_V}$ con respecto a la función de costo c.

Más en general si $\mathcal{X}$ y $\mathcal{Y}$ son espacios métricos completos y separables, μ y ν distribuciones de Borel sobre $\mathcal{X}$ y $\mathcal{Y}$ respectivamente, $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ una función continua. El costo total de transportar μ a ν asociado a $\gamma \in \Gamma(\mu, \nu)$ con respecto a c se define como:

$$C(\gamma) = \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)d\gamma(x, y).$$

Nosotros nos colocaremos en una situación mas específica. Sea $\mathcal{Z}$ un espacio de Banach, tomemos la función de costo como la distancia inducida por la métrica; si μ y ν son dos distribuciones de Borel sobre $\mathcal{Z}$, el costo total asociado a $\gamma \in \Gamma(\mu, \nu)$ se calcula:

$$C(\gamma) = \int_{\mathcal{Z}^2} ||x - y|| d\gamma(x, y).$$

Partiendo de estos conceptos podemos obtener una distancia sobre un conjunto de medidas.

4.2. Espacio de Wasserstein

Consideremos $\mathcal{P}(\mathcal{Z})$ el conjunto de las distribuciones de Borel sobre $\mathcal{Z}$. Definimos el p-espacio de *Wasserstein* en $\mathcal{Z}$ como:

$$\mathcal{W}_p(\mathcal{Z}) = \left\{ \mu \in \mathcal{P}(\mathcal{Z}) : \int_{\mathcal{Z}} ||x||^p d\mu(x) < \infty \right\}.$$

En la siguiente definición tenemos una distancia sobre este conjunto la cuál lo convierte en un espacio métrico [Vil03].

Definición 4.2. *Sea $\mathcal{Z}$ espacio de Banach, $p \geq 1$. Para μ, $\nu \in \mathcal{P}(\mathcal{Z})$ se define:*

$$W_p(\mu, \nu) := \left[\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathcal{Z}^2} ||x - y||^p d\gamma(x, y) \right]^{\frac{1}{p}}.$$

En este trabajo estaremos usando un caso límite de esta distancia. La llamamos distancia ∞-Wasserstein.

Definición 4.3. *Sea $\mathcal{Z}$ espacio de Banach. Para μ, $\nu \in \mathcal{P}(\mathcal{Z})$ se define:*

$$W_\infty(\mu, \nu) := \inf_{\gamma \in \Gamma(\mu,\nu)} \operatorname{ess\,sup}_{(X,Y)\sim\gamma} ||X - Y||.$$

En el siguiente lema se prueba la conexión entre p-distancia y la distancia infinito.

Lema 4.1. *Sean $\mu, \nu \in \mathcal{P}(\mathcal{Z})$. Se cumple:*

$$\lim_{p\to\infty} W_p(\mu, \nu) = W_\infty(\mu, \nu).$$

Demostración. Para $\mu = \nu$ es inmediato ya que para todo p se tiene

$$W_p(\mu, \nu) = 0 = W_\infty(\mu, \nu).$$

Consideremos entonces $\mu \neq \nu$, sea $\gamma \in \Gamma(\mu, \nu)$, $\varepsilon > 0$, y $E_\varepsilon = \{(x, y) \in \mathcal{Z}^2 : ||x - y|| \geq \operatorname{ess\,sup}_{(X,Y)\sim\gamma} ||X - Y|| - \varepsilon\}$; para $(x, y) \in E_\varepsilon$ tenemos:

$$||x - y||^p \geq \left(\operatorname*{ess\,sup}_{(X,Y)\sim\gamma} ||X - Y|| - \varepsilon \right)^p.$$

Integrando respecto a γ:

$$\int_{\mathcal{Z}} ||x-y||^p d\gamma(x,y) \geq \int_{E_\epsilon} ||x-y||^p d\gamma(x,y) \geq \int_{E_\varepsilon} \left(\operatorname*{ess\,sup}_{(X,Y)\sim\gamma} ||X-Y|| - \varepsilon \right)^p d\gamma(x,y).$$

Obteniendo el ínfimo sobre los acoplamientos, y luego raíz p-ésima:

$$\begin{aligned} W_p(\mu,\nu) &\geq \left[\operatorname*{ínf}_{\gamma\in\Gamma(\mu,\nu)} \int_{E_\varepsilon} \left(\operatorname*{ess\,sup}_{(X,Y)\sim\gamma} ||X-Y|| - \varepsilon \right)^p d\gamma(x,y) \right]^{\frac{1}{p}} \\ &= \left[\operatorname*{ínf}_{\gamma\in\Gamma(\mu,\nu)} \left(\operatorname*{ess\,sup}_{(X,Y)\sim\gamma} ||X-Y|| - \varepsilon \right)^p \gamma(E_\varepsilon) \right]^{\frac{1}{p}} \\ &\geq \left[\operatorname*{ínf}_{\gamma\in\Gamma(\mu,\nu)} \left(\operatorname*{ess\,sup}_{(X,Y)\sim\gamma} ||X-Y|| - \varepsilon \right)^p \operatorname*{ínf}_{\gamma\in\Gamma(\mu,\nu)} \gamma(E_\varepsilon) \right]^{\frac{1}{p}} \\ &= \left[\operatorname*{ínf}_{\gamma\in\Gamma(\mu,\nu)} \left(\operatorname*{ess\,sup}_{(X,Y)\sim\gamma} ||X-Y|| - \varepsilon \right)^p \right]^{\frac{1}{p}} \left[\operatorname*{ínf}_{\gamma\in\Gamma(\mu,\nu)} \gamma(E_\varepsilon) \right]^{\frac{1}{p}} \\ &= \operatorname*{ínf}_{\gamma\in\Gamma(\mu,\nu)} \left(\operatorname*{ess\,sup}_{(X,Y)\sim\gamma} ||X-Y|| - \varepsilon \right) \operatorname*{ínf}_{\gamma\in\Gamma(\mu,\nu)} \gamma(E_\varepsilon)^{\frac{1}{p}}. \end{aligned}$$

Tomamos límite inferior y considerando que $\gamma(E_\varepsilon) \in (0,\infty)$ y tomando ε lo suficientemente pequeño para que $\left(\operatorname{ess\,sup}_{(X,Y)\sim\gamma} ||X-Y|| - \varepsilon\right) > 0$:

$$\begin{aligned} \operatorname*{lím\,ínf}_{p\uparrow\infty} W_p(\mu,\nu) &\geq \operatorname*{lím\,ínf}_{p\uparrow\infty} \operatorname*{ínf}_{\gamma\in\Gamma(\mu,\nu)} \left(\operatorname*{ess\,sup}_{(X,Y)\sim\gamma} ||X-Y|| - \varepsilon \right) \gamma(E_\varepsilon)^{\frac{1}{p}} \\ &= \operatorname*{ínf}_{\gamma\in\Gamma(\mu,\nu)} \operatorname*{ess\,sup}_{(X,Y)\sim\gamma} ||X-Y|| - \varepsilon. \end{aligned}$$

Llevando ε a 0 este desaparece de la desigualdad:

$$\operatorname*{lím\,ínf}_{p\uparrow\infty} W_p(\mu,\nu) \geq \operatorname*{ínf}_{\gamma\sim\Gamma(\mu,\nu)} \operatorname*{ess\,sup}_{(X,Y)\sim\gamma} ||X-Y|| = W_\infty(\mu,\nu).$$

Por otro lado; para $\gamma \in \Gamma(\mu,\nu)$, si $(X,Y) \sim \gamma$ tenemos que γ-casi seguramente $||X-Y|| \leq \operatorname{ess\,sup}_\gamma ||X-Y||$, luego para $p > q$:

$$||X-Y||^{p-q} ||X-Y||^q \leq \left(\operatorname*{ess\,sup}_{(X,Y)\sim\gamma} ||X-Y|| \right)^{p-q} ||X-Y||^q.$$

Integrando con respecto a la medida γ:

$$\begin{aligned} \int_{\mathcal{Z}} ||x-y||^{p-q} ||x-y||^q d\gamma(x,y) &\leq \int_{\mathcal{Z}} \left(\operatorname*{ess\,sup}_{(X,Y)\sim\gamma} ||X-Y|| \right)^{p-q} ||x-y||^q d\gamma(x,y) \\ &= \left(\operatorname*{ess\,sup}_{(X,Y)\sim\gamma} ||X-Y|| \right)^{p-q} \int_{\mathcal{Z}} ||x-y||^q d\gamma(x,y). \end{aligned}$$

Luego tomando raíz p-ésima:

$$\left(\int_{\mathcal{Z}} ||x-y||^{p-q}||x-y||^q d\gamma(x,y)\right)^{\frac{1}{p}}$$
$$\leq \left(\operatorname*{ess\,sup}_{(X,Y)}||X-Y||\right)^{\frac{p-q}{p}} \left(\int_{\mathcal{Z}} ||x-y||^q d\gamma(x,y)\right)^{\frac{1}{p}}.$$

Tomamos ínfimo sobre los acoplamientos de μ y ν:

$$\operatorname*{ínf}_{\gamma\in\Gamma(\mu,\nu)} \left(\int_{\mathcal{Z}} ||x-y||^{p-q}||x-y||^q d\gamma(x,y)\right)^{\frac{1}{p}}$$
$$\leq \operatorname*{ínf}_{\gamma\in\Gamma(\mu,\nu)} \left(\operatorname*{ess\,sup}_{(X,Y)\sim\gamma}||X-Y||\right)^{\frac{p-q}{p}} \left(\int_{\mathcal{Z}} ||x-y||^q d\gamma(x,y)\right)^{\frac{1}{p}}$$
$$\leq \operatorname*{ínf}_{\gamma\in\Gamma(\mu,\nu)} \left(\operatorname*{ess\,sup}_{(X,Y)\sim\gamma}||X-Y||\right)^{\frac{p-q}{p}} \operatorname*{ínf}_{\gamma\in\Gamma(\mu,\nu)} \left(\int_{\mathcal{Z}} ||x-y||^q d\gamma(x,y)\right)^{\frac{1}{p}}.$$

Podemos escribir:

$$W_p(\mu,\nu) \leq \operatorname*{ínf}_{\gamma\in\Gamma(\mu,\nu)} \left(\operatorname*{ess\,sup}_{(X,Y)\sim\gamma}||x-y||\right)^{\frac{p-q}{p}} W_q(\mu,\nu)^{\frac{q}{p}}$$
$$\leq \operatorname*{ínf}_{\gamma\in\Gamma(\mu,\nu)} \operatorname*{ess\,sup}_{(X,Y)\sim\gamma}||X-Y|| \operatorname*{ínf}_{\gamma\in\Gamma(\mu,\nu)} \left(\operatorname*{ess\,sup}_{(X,Y)\sim\gamma}||X-Y||\right)^{-\frac{q}{p}} W_q(\mu,\nu)^{\frac{q}{p}}.$$

Tomando límite superior sobre p:

$$\limsup_{p\uparrow\infty} W_p(\mu,\nu) \leq W_\infty(\mu,\nu).$$

Concluimos:

$$\lim_{p\to\infty} W_p(\mu,\nu) = W_\infty(\mu,\nu).$$

□

Es de ayuda recordar que en notación de conjuntos podemos escribir:

$$W_\infty(\mu,\nu) := \text{ínf}\left\{||x-y||_{L^\infty(\gamma)} : \gamma\in\Gamma(\mu,\nu)\right\}.$$

Veamos ahora que $W_\infty(\mu,\nu)$ es una distancia en $\mathcal{P}_\infty(\mathcal{Z})$, que es el conjunto de todas las medidas de soporte compacto sobre $\mathcal{Z}$:

1. Supongamos $\mu \neq \nu$, entonces existe A subconjunto medible de $\mathcal{Z}$ tal que $\mu(A) \neq \nu(A)$, sin pérdida de generalidad suponemos $\mu(A) - \nu(A) > 0$. Ya que

$||x-y||_{L^\infty(\mathcal{Z}\times A,\gamma)} \leq ||x-y||_{L^\infty(\mathcal{Z}\times\mathcal{Z},\gamma)}$; donde para $||x-y||_{L^\infty(\mathcal{Z}\times A,\gamma)}$ tenemos:

$$\begin{aligned}\gamma\left[(x,y)\in\mathcal{Z}\times A : ||x-y||>0\right] &= \gamma\left[\mathcal{Z}\times A - \{(y,y) : y\in A\}\right]\\ &= \gamma\left[\mathcal{Z}\times A\right] - \gamma\left[\{(y,y) : y\in A\}\right]\\ &\geq \mu(A) - \gamma\left[A\times\mathcal{Z}\right]\\ &= \mu(A)-\nu(A) > 0.\end{aligned}$$

Por tanto $||x-y||_{L^\infty(\mathcal{Z}\times A,\gamma)} > 0$ para todo $\gamma\in\Gamma(\nu,\mu)$, así para todo $\gamma\in\Gamma(\nu,\mu)$ $W_\infty(\nu,\mu) > 0$

2. Para todo $p\geq 1$ tenemos que $W_p(\mu,\mu)=0$, de donde $W_\infty(\mu,\mu)=0$.

3. Para revisar la simetría tenemos:

$$W_\infty(\nu,\mu) = \lim_{p\to\infty} W_p(\nu,\mu) = \lim_{p\to\infty} W_p(\mu,\nu) = W_\infty(\mu,\nu).$$

4. Desigualdad triangular: Para todo $\mu,\nu,\zeta\in\mathcal{P}(\mathcal{Z})$ se tiene:

$$W_\infty(\nu,\mu) = \lim_{p\to\infty} W_p(\nu,\mu) \leq \lim_{p\to\infty}\left(W_p(\nu,\zeta)+W_p(\zeta,\mu)\right) = W_\infty(\nu,\zeta)+W_\infty(\zeta,\mu).$$

Cerramos esta sección con el siguiente lema, el cuál es uno de los soportes importantes de los resultados principales de este trabajo.

Lema 4.2. *Los siguientes enunciados son equivalentes para cualesquiera distribuciones μ,ν sobre $\mathcal{Z}$.*

1) $W_\infty(\mu,\nu)\leq s$.

2) Existe un vector aleatorio (U,V) tal que $U\sim\mu$, $V\sim\nu$ y $\mathbb{P}\left[||U-V||\leq s\right]=1$.

3) Existe un vector aleatorio (U,V) tal que $U\sim\mu$, $U+W\sim\nu$ y $\mathbb{P}\left[||W||\leq s\right]=1$.

Demostración. 1) $\Rightarrow$ 2):
Tenemos que:

$$\inf_{\gamma\in\Gamma(\mu,\nu)} \operatorname*{ess\,sup}_{(\mathrm{X},\mathrm{Y})\sim\gamma} ||X-Y|| \leq s$$

implica que para todo $\varepsilon>0$ existe una medida $\gamma\in\Gamma(\mu,\nu)$ tal que $\operatorname{ess\,sup}||U-V||\leq s+\varepsilon$ con $(U,V)\sim\gamma$, donde $U\sim\mu$, y $V\sim\nu$, tenemos que $\text{ínf}\{x\in\mathbb{R} : \mathbb{P}\{\omega\in\Omega : ||U(\omega)-V(\omega)||>x\}=0\}\leq s+\varepsilon$ de donde tenemos que existe $x\in\mathbb{R}$, $x\leq s+2\varepsilon$ tal que $\mathbb{P}\{\omega\in\Omega : ||U(\omega)-V(\omega)||>x\}=0$. Así cuando $\mathbb{P}$ es una medida de probabilidad tenemos $\mathbb{P}[||U-V||\leq s+2\varepsilon]=1$. para todo $\varepsilon>0$. Finalmente tomando el límite $\varepsilon\to 0$ se concluye $\mathbb{P}[||U-V||\leq s]=1$.

2)$\Rightarrow$3):
Definimos $W=V-U$, entonces $V=W+U$, y por tanto $W+U\sim\nu$ y además $\mathbb{P}[||W||\leq s]$.

3)$\Rightarrow$1)
Definimos $V=U+W$; ya que $\mathbb{P}[||U-V||\leq s]$, tenemos que $\mathbb{P}[||U-V||>s]=0$ por tanto $\text{ínf}\{x\in\mathbb{R} : \mathbb{P}[||U-V||>x]=0\}\leq s$, es decir $\operatorname{ess\,sup}||U-V||\leq s$, y así tenemos $W_\infty(\mu,\nu)\leq s$. □

Capítulo 5

Garantías de Privacidad en ICR

5.1. Algoritmos iterativos

En aprendizaje automático (*Machine learning*) es común encontrar algoritmos iterativos, es decir que actualizan progresivamente sus parámetros en una sucesión de iteraciones. En cada iteración el algoritmo usa un dato, o un cierto subconjunto de datos del conjunto de entrenamiento para lograr una dirección de mejora y así ir ajustando los parámetros del modelo.

En este contexto se ha demostrado que si no se publican los resultados parciales del algoritmo iterativo, la privacidad mejora [Fel+18]. Más en específico se esta pensando en algoritmos aleatorizados construidos usando funciones contractivas como por ejemplo el *descenso ruidoso del gradiente*.

Ejemplo 5.1. *El algoritmo "descenso ruidoso del gradiente" hace más eficiente la exploración del espacio de parámetros, evitando que la optimización se estanque en mínimos locales o zonas planas, además de prevenir o regular el sobre ajuste. La actualización de parámetros de este algoritmo es la siguiente:*

$$M(w) = w - \eta \nabla L(w) + \zeta.$$

Donde $\eta \in \mathbb{R}$ es el radio de aprendizaje, ζ el ruido es una variable aleatoria y ∇ es el gradiente.

Llamamos ruido a una variable aleatoria $X \sim \zeta$ o a su distribución, y denominamos sucesión de ruidos a una sucesión de variables aleatorias $\{X_n\}_{n\in I}$ o de distribuciones $\{\zeta_n\}_{n\in I}$ mutuamente independientes.

Definición 5.1 (Contracción)**.** *Para un espacio de Banach $(\mathcal{Z}, ||\cdot||)$ una función $\phi : \mathcal{Z} \to \mathcal{Z}$ se dice que es una contracción si para todo $x, y \in \mathcal{Z}$*

$$||\phi(x) - \phi(y)|| \leq ||x - y||.$$

Una definición general de los algoritmos en consideración es la siguiente:

Definición 5.2. *[Iteración de contracciones con ruido (ICR)] Dado un estado aleatorio inicial $X_0 \in \mathcal{Z}$ una sucesión de contracciones $\{\phi_t : \mathcal{Z} \to \mathcal{Z}\}_{t\leq T}$, y una sucesión de ruidos $\{\zeta_t\}_{t\leq T}$ definimos la iteración de contracciones con ruido (ICR) con la siguiente regla de actualización:*

$$X_t = \phi_t(X_{t-1}) + Z_t, \quad t \in \{1, \ldots, T\}.$$

Donde $Z_t \sim \zeta_t$ para cada $t \in \{1, \ldots, T\}$. Después de T pasos obtenemos la variable aleatoria X_T la cual será denotada como $ICR_T\left(X_0, \{\phi_t\}_{t\leq T}, \{\zeta_t\}_{t\leq T}\right)$.

5.2. Privacidad en algoritmos iterativos

Retomamos la divergencia de Renyi (*definición 3.1*) en la *observación 3.4* resaltamos que esta divergencia es una medida de privacidad, por tanto hay una versión de algoritmo aleatorizado privado (*definición 2.2*) para esta medida, sólo cabe destacar lo siguiente; para cada α tenemos una divergencia de Renyi distinta. Se plantea una definición de privacidad diferencial adaptada para incorporar a α como parámetro.

Definición 5.3. *[Privacidad diferencial de Renyi (RDP)] Sea $\alpha \in [1, \infty]$, $\varepsilon > 0$, $\mathcal{A}$ un algoritmo aleatorizado es (α, ε)-RDP si*

$$\sup_{D \sim D'} D_\alpha(\mathcal{A}(D)||\mathcal{A}(D')) \leq \varepsilon.$$

Para estudiar la privacidad (medida con D_α) en algoritmos *ICR* introducimos las siguientes definiciones. La idea es captar la distribución de salida en cada iteración, más en específico la distribución *pushforward* inducida por un mapeo contractivo o contracción.

Definición 5.4 (Divergencia de Renyi deslizada)**.** *Sean μ y ν distribuciones definidas en un espacio de Banach $(\mathcal{Z}, ||\cdot||)$. Para parámetros $z \geq 0$ y $\alpha \geq 1$, la divergencia z-deslizada de Renyi entre μ y ν se define como:*

$$D_\alpha^{(z)}(\mu||\nu) = \inf_{\mu' \in \overline{B}_{W_\infty}(\mu, z)} D_\alpha(\mu'||\nu).$$

Donde $\overline{B}_{W_\infty}(\mu, z)$ es la bola cerrada con centro en μ y radio z con respecto a la distancia W_∞.

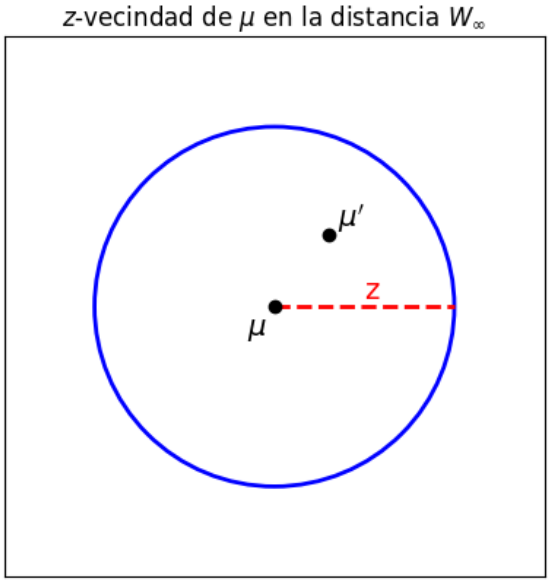

Figura 5.1: Vecindad de radio z en el espacio métrico $(\mathcal{P}_\infty(\mathcal{Z}), W_\infty)$.

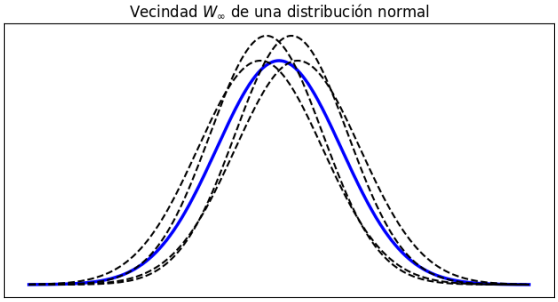

Figura 5.2: Así se ven las densidades dentro de una vecindad de una distribución $N(0,1)$, las distribuciones en la vecindad no necesariamente son normales, si no que corresponden a distribuciones que están cerca en términos de la distancia en el espacio donde las variables aleatorias toman valores.

Proposición 5.1. *La divergencia deslizada de Renyi satisface las siguientes condiciones para cualquier par de medidas μ, ν y $\alpha \in (1, \infty)$.*

- ***Monotonía:*** *Para $z \in (0, z')$, $D_\alpha^{(z')}(\mu||\nu) \leq D_\alpha^{(z)}(\mu||\nu)$.*
- ***Desplazamiento:*** *Para todo $x \in \mathcal{Z}$, $D_\alpha^{(||x||)}(\mu||\nu) \leq D_\alpha(\mu * \mathbf{x}||\nu)$ donde $\mathbf{x}$ denota la distribución de la variable aleatoria constante x, y $*$ la operación convolución.*

Demostración. Para monotonía tenemos; ya que $z < z'$, $\overline{B}_{W_\infty}(\mu, z) \subseteq \overline{B}_{W_\infty}(\mu, z')$ por tanto:

$$\inf_{\mu' \in \overline{B}_{W_\infty}(\mu, z)} D_\alpha(\mu'||\nu) \leq \inf_{\mu' \in \overline{B}_{W_\infty}(\mu, z')} D_\alpha(\mu'||\nu).$$

Luego para el desplazamiento notemos que

$$\begin{aligned} W_\infty(\mu, \mu * \mathbf{x}) &= \inf_{\gamma \in \Gamma(\mu, \mu * \mathbf{x})} \operatorname*{ess\,sup}_{(X,Y)\sim\gamma} ||X - Y|| \\ &\leq \operatorname*{ess\,sup}_{(X,Y)\sim(X, X+x)} ||X - Y|| \\ &= \operatorname*{ess\,sup}_{(X,Y)\sim(X, X+x)} ||X - (X + x)|| \\ &= \operatorname*{ess\,sup}_{(X,Y)\sim(X, X+x)} ||x|| \\ &= ||x||. \end{aligned}$$

Por tanto $\mu * \mathbf{x} \in \overline{B}_{W_\infty}(\mu, ||x||)$, así

$$\inf_{\mu' \in \overline{B}_{W_\infty}(\mu, ||x||)} D_\alpha(\mu'||\nu) \leq D_\alpha(\mu * \mathbf{x}||\nu).$$

□

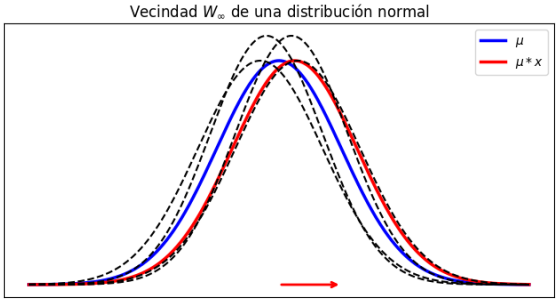

Figura 5.3: $\mu * \mathbf{x}$ está dentro de la vecindad de radio $||x||$ alrededor de μ.

Definición 5.5. *Para una distribución ζ sobre un espacio de Banach $(\mathcal{Z}, ||\cdot||)$ definimos la magnitud de ruido de radio a como:*

$$R_\alpha(\zeta, a) = \sup_{x:||x||\leq a} D_\alpha(\zeta * \mathbf{x}||\zeta).$$

La magnitud de ruido de radio a nos permitirá poner una cota a la privacidad de un algoritmo *ICR* en función de los ruidos involucrados en su construcción.

Lema 5.1 (Reducción de ruido)**.** *Sean $z \geq 0$, μ, ν y ζ distribuciones sobre el espacio de Banach $(\mathcal{Z}, ||\cdot||)$. Entonces para todo $a \geq 0$:*

$$D_\alpha^{(z)}(\mu * \zeta||\nu * \zeta) \leq D_\alpha^{(z+a)}(\mu||\nu) + R_\alpha(\zeta, a).$$

Demostración. Primero asumamos $z = 0$. Ya que $W_\infty(\mathcal{Z})$ es completo existe μ' tal que $W_\infty(\mu', \mu) \leq a$ y $D_\alpha(\mu'||\nu) = D_\alpha^{(a)}(\mu||\nu)$. Sea (U, W) la variable aleatoria dada por el *lema 4.2 inciso 3)*. De ahí tenemos $\mathbb{P}[||W|| \leq a] = 1$, $U \sim \mu$, y $U + W \sim \mu'$. Sea también $V \sim \nu$, y finalmente, $Y \sim \zeta$ variable aleatoria independiente de las anteriores. Podemos escribir:

$$\begin{aligned} D_\alpha(\mu * \zeta||\nu * \zeta) &= D_\alpha(U + Y||V + Y) \\ &= D_\alpha(U + W - W + Y||V + Y). \end{aligned}$$

Usando la propiedad de posprocesamiento con la función determinista $(x, y) \xrightarrow{f} x+y$ tenemos

$$D_\alpha(U + W - W + Y||V + Y) \leq D_\alpha((U + W, -W + Y)||(V, Y)).$$

Por tanto:

$$D_\alpha(\mu * \zeta||\nu * \zeta) \leq D_\alpha((U + W, -W + Y)||(V, Y)). \tag{5.1}$$

Ahora usamos la independencia para obtener:

$$\begin{aligned}
&\exp\left[(\alpha-1)D_\alpha\left((U+W,-W+Y)||(V,Y)\right)\right]\\
&=\int\int\left(\frac{f_{(U+W,-W+Y)}(v,y)}{f_{(V,Y)}(v,y)}\right)^\alpha f_{(V,Y)}(v,y)dvdy\\
&=\int\int\left(\frac{f_{-W+Y|U+W}(y|v)f_{U+W}(v)}{f_V(v)f_Y(y)}\right)^\alpha f_V(v)f_Y(y)dvdy\\
&=\int\left(\frac{f_{U+W}(v)}{f_V(v)}\right)^\alpha f_V(v)\left(\int\left(\frac{f_{-W+Y|U+W}(y|v)}{f_Y(y)}\right)^\alpha f_Y(y)dy\right)dv\\
&\leq\int\left(\frac{f_{U+W}(v)}{f_V(v)}\right)^\alpha f_V(v)\left(\operatorname*{ess\,sup}_{v'\sim\nu}\int\left(\frac{f_{-W+Y|U+W}(y|v)}{f_Y(y)}\right)^\alpha f_Y(y)dy\right)dv\\
&=\int\left(\frac{f_{U+W}(v)}{f_V(v)}\right)^\alpha f_V(v)dv\ \operatorname*{ess\,sup}_{v'\sim\nu}\int\left(\frac{f_{-W+Y|U+W}(y|v)}{f_Y(y)}\right)^\alpha f_Y(y)dy. \qquad (5.2)
\end{aligned}$$

Luego consideremos lo siguiente:

$$\begin{aligned}
f_{-W+Y|U+W}(y|v)&=\frac{f_{U+W,-W+Y}(v,y)}{f_{U+W}(v)}\\
&=\frac{\int f_{U+W,-W+Y,W}(v,y,w)dw}{f_{U+W}(v)}\\
&=\int\frac{f_{-W+Y|U+W,W}(y|v,w)f_{U+W,W}(v,w)}{f_{U+W}(v)}dw\\
&=\int f_{-W+Y|U+W,W}(y|v,w)f_{W|U+W}(w|v)dw.
\end{aligned}$$

De aquí escribimos:

$$\begin{aligned}
&\int\left(\frac{f_{-W+Y|U+W}(y|v)}{f_Y(y)}\right)^\alpha f_Y dy\\
&=\int\left(\int\frac{f_{-W+Y|U+W,W}(y|v,w)}{f_Y(y)}f_{W|U+W}(w|v)dw\right)^\alpha f_Y(y)dy.
\end{aligned}$$

Ya que $\alpha>1$ usamos la desigualdad de Jenssen con respecto a $f_{W|U+W}(w|v)dw$:

$$\leq\int\left[\int\left(\frac{f_{-W+Y|U+W,W}(y|v,w)}{f_Y(y)}\right)^\alpha f_{W|U+W}(w|v)dw\right]f_Y(y)dy.$$

Luego usando el teorema de Tonelli:

$$\begin{aligned}
&=\int\left[\int\left(\frac{f_{-W+Y|U+W,W}(y|v,w)}{f_Y(y)}\right)^\alpha f_Y(y)dy\right]f_{W|U+W}(w|v)dw\\
&\leq\operatorname*{ess\,sup}_{w\sim\mathbb{P}_W}\int\left(\frac{f_{-W+Y|U+W,W}(y|v,w)}{f_Y(y)}\right)^\alpha f_Y(y)dy.
\end{aligned}$$

Ahora tenemos:

$$\begin{aligned}
&\operatorname*{ess\,sup}_{v'\sim\nu}\int\left(\frac{f_{-W+Y|U+W}(y|v)}{f_Y(y)}\right)^\alpha f_Y(y)dy\\
&\leq\operatorname*{ess\,sup}_{v'\sim\nu}\operatorname*{ess\,sup}_{w\sim\mathbb{P}_W}\int\left(\frac{f_{-W+Y|U+W,W}(y|v,w)}{f_Y(y)}\right)^\alpha f_Y(y)dy.
\end{aligned}$$

Regresando a (5.2) hemos llegado a que:

$$\exp\left[(\alpha-1)D_\alpha\left((U+W,-W+Y)||(V,Y)\right)\right] \leq$$
$$\int\left(\frac{f_{U+W}(v)}{f_V(v)}\right)^\alpha f_V(v)dv \operatorname*{ess\,sup}_{v'\sim\nu}\operatorname*{ess\,sup}_{w\sim\mathbb{P}_W}\int\left(\frac{f_{-W+Y|U+W,W}(y|v,w)}{f_Y(y)}\right)^\alpha f_Y(y)dy.$$

Hay que detenerse aquí un momento para notar lo siguiente

$$\frac{f_{-W+Y,U+W,W}(y,v,w)}{f_{U+W,W}(v,w)} = \frac{f_{Y,U+W,W}(y+w,v,w)}{f_{U+W,W}(v,w)};$$

ya que Y es independiente al resto de las variables

$$\begin{aligned}\frac{f_{Y,U+W,W}(y+w,v,w)}{f_{U+W,W}(v,w)} &= f_Y(y+w)\frac{f_{U+W,W}(v,w)}{f_{U+W,W}(v,w)}\\ &= f_Y(y+w)\\ &= f_{Y-W}(y).\end{aligned}$$

Es decir:

$$f_{-W+Y|U+W,W}(y|v,w) = f_{Y-W}(y).$$

Así que:

$$\begin{aligned}&\exp\left[(\alpha-1)D_\alpha\left((U+W,-W+Y)||(V,Y)\right)\right]\\ &\leq\int\left(\frac{f_{U+W}(v)}{f_V(v)}\right)^\alpha f_V(v)dv\ \operatorname*{ess\,sup}_{v'\sim\nu}\ \operatorname*{ess\,sup}_{w\sim\mathbb{P}_W}\int\left(\frac{f_{-W+Y}(y)}{f_Y(y)}\right)^\alpha f_Y(y)dy\\ &=\int\left(\frac{f_{U+W}(v)}{f_V(v)}\right)^\alpha f_V(v)dv\quad\operatorname*{ess\,sup}_{w\sim\mathbb{P}_W}\int\left(\frac{f_{-W+Y}(y)}{f_Y(y)}\right)^\alpha f_Y(y)dy\\ &\leq\int\left(\frac{f_{U+W}(v)}{f_V(v)}\right)^\alpha f_V(v)dv\quad\sup_{x:||x||\leq a}\int\left(\frac{f_{x+Y}(y)}{f_Y(y)}\right)^\alpha f_Y(y)dy.\end{aligned}$$

Regresamos a la divergencia de Renyi reescribiendo de la siguiente manera

$$\int\left(\frac{f_{U+W}(v)}{f_V(v)}\right)^\alpha f_V(v)dv\quad\sup_{x:||x||\leq a}\int\left(\frac{f_{x+Y}(y)}{f_Y(y)}\right)^\alpha f_Y(y)dy$$

$$= \exp\left[(\alpha-1)D_\alpha(\mu'||\nu)\right]\sup_{x:||x||\leq a}\exp\left[(\alpha-1)D_\alpha(\zeta*\mathbf{x}||\zeta)\right]$$

$$\begin{aligned}&= \exp\left[(\alpha-1)D_\alpha(\mu'||\nu)\right]\exp\left[(\alpha-1)\sup_{x:||x||\leq a}D_\alpha(\zeta*\mathbf{x}||\zeta)\right]\\ &= \exp\left[(\alpha-1)D_\alpha(\mu'||\nu)\right]\exp\left[(\alpha-1)R_\alpha(\zeta,a)\right].\end{aligned}$$

Retomando desde la desigualdad *(5.1)* tenemos:

$$D_\alpha(\mu*\zeta||\nu*\zeta) \leq D_\alpha^{(a)}(\mu||\nu) + R_\alpha(\zeta,a).$$

Ya esta para $z=0$. Ahora definimos para $z>0$:

$$h_z(x) = \begin{cases} x & si\ ||x||\leq z,\\ \frac{x}{||x||}z & si\ z<||x||.\end{cases}$$

Notemos que $||h_z(x)|| \leq z$ para todo x, y si $||x|| \leq z + a$, tenemos

$$||x - h_z(x)|| \leq ||x|| - ||h_z(x)|| \leq z + a + ||h_z(x)|| \leq a.$$

Consideremos ahora μ' tal que $D_\alpha(\mu'||\nu) = D_\alpha^{(z+a)}(\mu||\nu)$; consideremos la variable aleatoria (U, W) tal que $||W|| \leq a$ con probabilidad 1, $U \sim \mu$ y $U + W \sim \mu'$. Establezcamos $W_1 = h_z(W)$, y $W_2 = W - W_1$. En base a las observaciones anteriores tenemos $||W_1|| \leq z$ y $||W_2|| \leq a$ con probabilidad 1. Luego entonces tenemos:

$$\begin{aligned} W_\infty(\mu * \zeta, \mu * \mathbb{P}_{W_1} * \zeta) &= \inf_{\gamma \in \Gamma(\mu*\zeta, \mu*\mathbb{P}_{W_1}*\zeta)} \operatorname*{ess\,sup}_{(x,y)\sim\gamma} ||x - y|| \\ &\leq \operatorname*{ess\,sup}_{(x,y)\sim(U+Y, U+W_1+Y)} ||x - y|| \\ &= \operatorname*{ess\,sup}_{(U+Y, U+W_1+Y)} ||U + Y - (U + W_1 + Y)|| \\ &= \operatorname*{ess\,sup}_{(U+Y, U+W_1+Y)} ||W_1|| \\ &\leq z. \end{aligned}$$

Así que $\mu * \mathbb{P}_{W_1} * \zeta \in \overline{B}_{||\cdot||}(\mu * \zeta, z)$, por tanto:

$$\begin{aligned} D_\alpha^{(z)}(U + Y||V + Y) &\leq D_\alpha(U + W_1 + Y||V + Y) \\ &\leq D_\alpha^{(a)}(U + W_1||V) + R_\alpha(\zeta, a) \quad \textit{por el caso } z = 0. \end{aligned}$$

Luego tenemos también, en analogía a lo anterior $W_\infty(\mu * \mathbb{P}_{W_1}, \mu * \mathbb{P}_{W_1} * \mathbb{P}_{W_2}) \leq a$, implica:

$$\begin{aligned} D_\alpha^{(a)}(U + W_1||V) &\leq D_\alpha(U + W_1 + W_2||V) \\ &= D_\alpha(U + W||V) \\ &= D_\alpha^{(z+a)}(U||V). \end{aligned}$$

Concluimos:

$$D_\alpha^{(z)}(\mu * \zeta||\nu * \zeta) \leq D_\alpha^{(z+a)}(\mu||\nu) + R_\alpha(\zeta, a).$$

□

Lema 5.2 (Contracción reduce $D_\alpha^{(z)}$)**.** *Sea $z \geq 0$, y ϕ, ϕ' contracciones en $(\mathcal{Z}, ||\cdot||)$ tales que $\sup_x ||\phi(x) - \phi'(x)|| \leq s$. Entonces para variables aleatorias X y X' sobre $\mathcal{Z}$ se cumple:*

$$D_\alpha^{(z+s)}(\phi(X)||\phi'(X')) \leq D_\alpha^{(z)}(X||X').$$

Demostración. Usamos el *lema 4.2* Para establecer una variable aleatoria Y tal que:

$$D_\alpha^{(z)}(X||X') = D_\alpha(Y||X')$$

con $\mathbb{P}[||Y - X|| \leq z] = 1$, es decir, sea Y tal que el ínfimo se alcanza, como se tiene $W_\infty(\mathbb{P}_Y, \mathbb{P}_X) \leq z$ por el ya citado lema, tenemos que con probabilidad 1 se cumple la desigualdad.
Luego observamos que, para esta variable aleatoria Y con probabilidad 1 tenemos, por ser ϕ contracción y por hipótesis:

$$\begin{aligned} ||\phi(X) - \phi'(Y)|| &= ||\phi(X) - \phi(Y) + \phi(Y) - \phi'(Y)|| \\ &\leq ||\phi(X) - \phi(Y)|| + ||\phi(Y) - \phi'(Y)|| \\ &\leq ||X - Y|| + s \\ &\leq z + s. \end{aligned}$$

Es decir; $\mathbb{P}\left[||\phi(X) - \phi'(Y)|| \leq z + s\right] = 1$, nuevamente usando el *lema 4.2* tenemos que $W_\infty(\mathbb{P}_{\phi(X)}, \mathbb{P}_{\phi'(Y)}) \leq z+s$, por tanto: $D_\alpha^{(z+s)}(\phi(X)||\phi'(X')) \leq D_\alpha(\phi'(Y)||\phi'(X'))$ por definición de divergencia de Renyi deslizada. Ahora usando la propiedad de $post - processing$ para la divergencia de Renyi tenemos:

$$D_\alpha^{(z+s)}(\phi(X)||\phi'(X')) \leq D_\alpha(\phi'(Y)||\phi'(X')) \leq D_\alpha(Y||X') = D_\alpha^{(z)}(X||X').$$

□

El siguiente teorema da garantías de privacidad para algoritmos ICR calibrando los ruidos que se suman en cada iteración.

Teorema 5.3. *Sean X_T, $X_T^{'}$ salidas de $ICR_T(X_0, \{\phi_t\}, \{\zeta_t\})$ y $ICR_T(X_0, \{\phi_t'\}, \{\zeta_t\})$ respectivamente. Sea $s_t \geq \sup_x ||\phi_t(x) - \phi_t^{'}(x)||$. Sea $a_1, \ldots, a_T$ una sucesión de números reales positivos y sea $z_t = \sum_{i\leq t} s_i - \sum_{i\leq t} a_i$. Si $z_t \geq 0$ para todo t, entonces:*

$$D_\alpha^{(z_T)}(X_T||X_T^{'}) \leq \sum_{t=1}^{T} R_\alpha(\zeta_t, a_t).$$

En particular si $z_T = 0$, entonces:

$$D_\alpha(X_T||X_T^{'}) \leq \sum_{t=1}^{T} R_\alpha(\zeta_t, a_t).$$

Demostración. Procedemos por inducción sobre T. Caso $T = 1$: Sea $Z_1 \sim \zeta$, tenemos:

$$X_1 = \phi_1(X_0) + Z_1 \qquad \text{y} \qquad X_1' = \phi_1'(X_0) + Z_1.$$

Para $T = 1$ tenemos $a_1 \geq 0$, $s_1 = \sup_x ||\phi_1(x) - \phi_1'(x)||$, y $z_1 = s_1 - a_1$ bajo la suposición $z_1 \geq 0$. Usando el lema de reducción de ruido tenemos:

$$\begin{aligned} D_\alpha^{(z_1)}(X_1||X_1') &= D_\alpha^{(z_1)}(\phi_1(X_0) + Z_1||\phi_1'(X_0) + Z_1) \\ &\leq D_\alpha^{(z_1+a_1)}(\phi_1(X_0)||\phi_1'(X_0)) + R_\alpha(\zeta_1, a_1). \end{aligned}$$

Luego por el lema *Contracción reduce* $D_\alpha^{(z)}$ tenemos:

$$\begin{aligned} D_\alpha^{(z_1+a_1)}(\phi_1(X_0)||\phi_1'(X_0)) + R_\alpha(\zeta_1, a_1) &= D_\alpha^{(s)}(\phi_1(X_0)||\phi_1'(X_0)) + R_\alpha(\zeta_1, a_1) \\ &\leq D_\alpha^{(0)}(X_0||X_0) + R_\alpha(\zeta_1, a_1) \\ &= R_\alpha(\zeta_1, a_1). \end{aligned}$$

Concluimos el caso $T = 1$:

$$D_\alpha^{(z_1)}(X_1||X_1') \leq R_\alpha(\zeta_1, a_1).$$

Supongamos ahora que se cumple para $T = k$ por demostrar que es válido para $k + 1$: Usando el lema de reducción de ruido:

$$\begin{aligned} D_\alpha^{(z_{k+1})}(X_{k+1}||X_{k+1}') &= D_\alpha^{(z_k+s_{k+1}-a_{k+1})}(\phi_{k+1}(X_k) + Z_{k+1}||\phi_{k+1}'(X_k') + Z_{k+1}) \\ &\leq D_\alpha^{(z_k+s_{k+1})}(\phi_{k+1}(X_k)||\phi_{k+1}'(X_k')) + R_\alpha(\zeta_{k+1}, a_{k+1}). \end{aligned}$$

Por *Contracción reduce* $D_\alpha^{(z)}$

$$D_\alpha^{(z_{k+1})}(X_{k+1}||X'_{k+1}) \leq D_\alpha^{(z_k)}(X_k||X'_k) + R_\alpha(\zeta_{k+1}, a_{k+1}).$$

Por hipótesis de inducción:

$$D_\alpha^{(z_{k+1})}(X_{k+1}||X'_{k+1}) \leq \sum_{i\leq k} R_\alpha(\zeta_i, a_i) + R_\alpha(\zeta_{k+1}, a_{k+1}).$$

□

Este teorema es central en las pruebas de resultados sobre la conservación de la privacidad en algoritmos de descenso del gradiente. Veamos un ejemplo tomado de [Fel+18], *Proyected noisy stochastic gradient descent* (*PNSGD*).

Algoritmo: Proyected Noisy Stochastic Gradient Descent (PNSGD) en $\mathbb{R}^n$

Entrada: Conjunto de datos $S = \{x_1, \ldots, x_n\}$, función $f : K \times X \to \mathbb{R}$ convexa en el primer argumento, tasa de aprendizaje η, punto inicial $w_0 \in K \subseteq \mathbb{R}^n$, parámetro de ruido σ.

Procedimiento: Para cada iteración $t \in \{0, \ldots, n-1\}$ hacer:

$$v_{t+1} \leftarrow w_t - \eta\left(\nabla_w f(w_t, x_{t+1}) + Z_t\right), \quad Z_t \sim \mathcal{N}(0, \sigma^2 I_d)$$
$$w_{t+1} \leftarrow \Pi_K(v_{t+1}) = \arg\min_{\theta\in K} \|\theta - v_{t+1}\|_2^2.$$

Salida: El punto final w_n. En este momento destacaremos que

$$R_\alpha(N(0, \sigma^2\mathbb{I}), a) = \frac{\alpha a^2}{2\sigma^2}, \tag{5.3}$$

esto se deduce fácilmente del siguiente resultado que se puede encontrar en [LV87] pág. 45.

$$D_\alpha(N(\mu_0\Sigma)||N(\mu_1, \Sigma)) = \frac{1}{\alpha}(\mu_1 - \mu_0)^T\Sigma^{-1}(\mu_0 - \mu_1).$$

El siguiente resultado es una aplicación del *teorema 5.3*.

Teorema 5.4. *Sea $\mathcal{K} \subseteq \mathbb{R}^n$ un conjunto convexo, $\{f(\cdot, x)\}_{x\in\mathcal{X}}$ una familia de funciones L-Lipschitz, con gradiente β-Lipschitz sobre $\mathcal{K}$. Entonces para todo $\eta < 2/\beta$, $\sigma > 0$, $\alpha > 1$, $t \in \{1, \ldots, n\}$, $w_0 \in \mathcal{K}$, $S \in \mathcal{X}^n$, $PNSGD(S, w_0, \eta, \sigma)$ satisface*

$$\left(\alpha, \frac{\alpha\varepsilon}{n+1-t}\right) - RDP$$

para su t-ésima entrada, donde $\varepsilon = \frac{2L^2}{\sigma^2}$.

Demostración. Consideremos $S = (x_1, \ldots, x_n)$ y $S' = (x_1, \ldots, x_{t-1}, x'_t, x_{t+1}, \ldots, x_n)$ De los teoremas B.2 y B.1 en el apéndice deducimos que $PNSGD(S, w_0, \eta, \sigma)$ es una algoritmo ICR (*definición 5.2*) bajo la hipótesis de que ∇f es β-Lipschitz y

$\eta \leq 2/\beta$, ya que la proyección es una contracción y también gradiente. El algoritmo ICR sobre el conjunto S se conforma por las contracciones

$$\psi_i(w) = \Pi_{\mathcal{K}}(w) - \eta \nabla f(\Pi_{\mathcal{K}}(w), x_i),$$

y los ruidos $Z_i \sim N(0, (\eta\sigma)^2\mathbb{I})$, para S', las contracciones quedan como $\psi_i' = \psi_i$, excepto para t, aquí f esta evaluada en x_t'; ya que para todo $x \in \mathcal{X}$, y $w \in \mathcal{K}$ f es L-Lipschitz tenemos:

$$\begin{aligned} &\sup_w ||\psi_t(w) - \psi_t'(w)|| \\ &= \sup_w ||\eta \nabla f(\Pi_{\mathcal{K}}(w), x_t) - \eta \nabla f(\Pi_{\mathcal{K}}(w), x_t')|| \leq 2\eta L. \end{aligned}$$

Aplicamos el *teorema 5.3* con $a_1, \ldots, a_{t-1} = 0$ y $a_t \ldots, a_n = \frac{2\eta L}{n-t+1}$, donde definimos $s_t = 2\eta L$ y $s_i = 0$ para $i \neq t$. De esta manera $z_i \geq 0$ para $i \leq n$, con $z_n = 0$. Por este mismo teorema y por la expresión 5.3 se obtiene

$$D_\alpha(X_n || X_n') \leq \frac{\alpha}{2\eta^2\sigma^2} \sum_{i=1}^{n} a_t^2 \leq \frac{2\alpha L^2}{\sigma^2(n-t+1)}.$$

□

Una aplicación sencilla, pero ilustrativa es la regresión lineal. Considérese un conjunto de datos $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ donde $\mathbf{x}_i \in \mathbb{R}^d$ y $y_i \in \mathbb{R}$, como espacio de pesos $\mathcal{K} = \{\mathbf{w} \in \mathbb{R}^d \,|\, ||\mathbf{w}||_2 \leq R\}$, función de aproximación $h_{\mathbf{w}}(\mathbf{x}_i) = \langle \mathbf{w}, \mathbf{x}_i \rangle$ donde $\mathbf{w} \in \mathcal{K}$ es el vector de pesos, la función de pérdida es el error cuadrático medio $f(\mathbf{w}, \mathbf{x}_i) = \frac{1}{2}(y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle)^2$.

El gradiente de la pérdida es $\nabla f(\mathbf{w}, \mathbf{x}_i) = (\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i)\mathbf{x}_i$. Nótese que, cuando $||\mathbf{x}_i||_2 \leq 1$ y $|y_i| \leq 1$, entonces $f(\mathbf{w}, \mathbf{x}_i)$ es L-Lipschitz con $L = R + 1$, el gradiente es 1-Lipschitz.

Así para todo $\eta < 2$, $\sigma > 0$, $\alpha > 1$, $t = \{1, \ldots, n\}$, $\mathbf{w}_0 \in \mathcal{K}$, S y $\zeta_i \sim N(0, \sigma^2\mathbb{I})$ tenemos, por el *teorema 5.4*, que el algoritmo $ICR\left(\mathbf{w}_0, \{f(\cdot, \mathbf{x}_i)\}_{i=1}^n, \{\zeta\}\right)$ es $\left(\alpha, \frac{\alpha\varepsilon}{n+1-t}\right) - RDP$ para la t-ésima iteración, donde $\varepsilon = \frac{2(R+1)^2}{\sigma^2}$ para todo i.

Capítulo 6

Conclusiones

Se dispone de un método garantizado para calibrar la privacidad en algoritmos de descenso del gradiente aplicados a funciones de pérdida ampliamente utilizadas, bajo ciertas modificaciones como la proyección sobre conjuntos convexos. Este enfoque es factible siempre que se pueda acotar una región alrededor del mínimo global. En tales condiciones, al añadir ruido en cada iteración y posponer la publicación de los resultados hasta el final, se asegura la garantía de privacidad.

La relevancia de este teorema radica en que permite resguardar la información del conjunto de entrenamiento en algoritmos de aprendizaje automático. Además, la teoría alcanza un nivel suficiente de generalidad mediante el uso de la herramienta matemática conocida como magnitud de ruido de radio a, lo que ha posibilitado parametrizar distribuciones arbitrarias para el ruido agregado en cada iteración de los algoritmos. Asimismo, la distancia infinito de Wasserstein permite calibrar el radio alrededor de la distribución inicial en el cual la garantía de privacidad queda asegurada.

Apéndice A

Probabilidad y Medida

Usaremos las definiciones comunes de σ-álgebra, espacio medible, medida y medida de probabilidad. Un espacio de *Banach* es un espacio vectorial normado y completo en la métrica definida por su norma. Si μ y ν son dos medidas sobre el espacio medible $(\Omega, \mathcal{F})$ decimos que ν es absolutamente continua con respecto a μ cuando $\nu(A) = 0$ siempre que $\mu(A) = 0$ para $A \in \mathcal{F}$.

A continuación se exponen un conjunto de definiciones y resultados clásicos en la teoría de la medida y de probabilidad. Comenzamos esta sección recordando conceptos elementales, esto será útil también para indicar la notación que usaremos a lo largo de este trabajo.

Si X es un conjunto no vacío escribimos $\sigma(X)$ para denotar a la σ-álgebra generada por dicho conjunto; es decir la σ-álgebra más pequeña que contiene a X.

La *definición A.1*, así como el *teorema A.1*, y el *Corolario A.1.1* son tomados del libro de *Gravinsky* [Gra09] las demostraciones se pueden encontrar ahí mismo, el tema que se aborda es el teorema de clases monótonas.

Definición A.1. *Consideremos un conjunto $X \neq \emptyset$.*

- *Un conjunto $\mathcal{C} \subseteq \{0,1\}^X$, $\mathcal{C} \neq \emptyset$ se llama π-sistema si es cerrado bajo intersecciones finitas.*
- *Un conjunto $\mathcal{L} \subseteq \{0,1\}^X$, $\mathcal{L} \neq \emptyset$ se llama un sistema de Dynkin cuando:*

  1.- $X \in \mathcal{L}$.

  2.- Si $E, F \in \mathcal{L}$ y $F \subset E$, entonces $E - F \in \mathcal{L}$.

  3.- Si $\{E_n\}_{n\in\mathbb{N}} \subset \mathcal{L}$ es una sucesión creciente, entonces $\bigcup_{n\in\mathbb{N}} E_n \in \mathcal{L}$.

El siguiente teorema es llamado de *clases monótonas* o teorema de *Dynkin*.

Teorema A.1. *[Clases Monótonas] Sean $X \neq \emptyset$, $\mathcal{C} \in \{0,1\}^X$ un π-sistema, y $\mathcal{L}$ un sistema de Dynkin tal que $\mathcal{C} \subseteq \mathcal{L}$, entonces $\sigma(\mathcal{C}) \subseteq \mathcal{L}$.*

Luego el siguiente corolario es una observación que indica en qué sentido este teorema es útil.

Corolario A.1.1. *Si $\mathcal{C} \subseteq \{0,1\}^X$ es un π-sistema, entonces:*

$$\sigma(\mathcal{C}) = \bigcap \{\mathcal{D} \text{ sistema de Dynkin en } \{0,1\}^X : \mathcal{C} \subseteq \mathcal{D}\}.$$

Dado que haremos uso frecuente de la derivada de *Radón-Nikodym* recordaremos el teorema que introduce el concepto, su demostración se puede encontrar en [Roy68].

Teorema A.2 (Radón-Nikodym)**.** *Sea $(\Omega, \mathcal{F}, \mu)$ un espacio de medida σ-finito, ν una medida definida sobre $\mathcal{F}$ tal que $\nu \ll \mu$. Entonces existe una función medible no negativa f única en casi todas partes respecto a μ, tal que para todo $A \in \mathcal{F}$ se cumple:*

$$\nu(A) = \int_A f d\mu.$$

El significado de la expresión *casi todas partes respecto a μ* es el habitual en el contexto de análisis real, en este caso para todo $g : \Omega \to \mathbb{R}^+ \cup \{0\}$, $\mu\{\omega \in \Omega : f(\omega) \neq g(\omega)\} = 0$. A la función medible f que proporciona el teorema se le llama derivada de *Radón-Nikodym*, usamos también la siguiente notación:

$$\frac{d\nu}{d\mu} := f.$$

Los siguientes lemas y proposiciones son propiedades de la derivada de *Radón-Nikodym* que se usarán en desarrollos posteriores.

Lema A.3. *Sea $(\Omega, \mathcal{F})$ un espacio medible, μ, ν, $\mathfrak{m}$ medidas sobre este espacio tales que $\mu \ll \nu \ll \mathfrak{m}$. Se cumple:*

- *Para cualquier variable aleatoria $X : \Omega \to \mathbb{R}$:*

$$\int X d\nu = \int X \frac{d\nu}{d\mathfrak{m}} d\mathfrak{m}.$$

- *Para $\frac{d\nu}{d\mu} \neq 0$ casi en todas partes $[\mathfrak{m}]$, entonces $\frac{(d\mu/d\mathfrak{m})}{(d\nu/d\mathfrak{m})} = \frac{d\mu}{d\nu}$.*

Demostración. Para todo $A \in \mathcal{F}$ tenemos:

$$\int \frac{d\nu}{d\mathfrak{m}} \mathbf{1}_A d\mathfrak{m} = \nu(A) = \int \mathbf{1}_A d\nu.$$

Siguiendo el razonamiento canónico; si f es una función simple con valores $f(\omega) = \sum_{i=1}^n a_i \mathbf{1}_i(\omega)$ integrando:

$$\begin{aligned} \int \frac{d\nu}{d\mathfrak{m}} f d\mathfrak{m} &= \int \frac{d\nu}{d\mathfrak{m}} \left(\sum_{i=1}^n a_i \mathbf{1}_{A_i} \right) d\mathfrak{m} \\ &= \sum_{i=1}^n a_i \int \frac{d\nu}{d\mathfrak{m}} \mathbf{1}_{A_i} d\mathfrak{m} \end{aligned}$$

$$= \sum_{i=1}^{n} a_i \nu(A_i) = \int f d\nu.$$

Luego toda función medible positiva g puede ser aproximada por alguna sucesión de funciones simples, por ejemplo $\{g_n\}$. Usando la propiedad (lema) de convergencia monótona tenemos:

$$\begin{aligned}
\int \frac{d\nu}{d\mu} g d\mathfrak{m} &= \int \frac{d\nu}{d\mu} \lim_{n\uparrow\infty} g_n d\mathfrak{m} \\
&= \lim_{n\uparrow\infty} \int \frac{d\nu}{d\mu} g_n d\mathfrak{m} \\
&= \lim_{n\uparrow\infty} \int g_n d\nu \\
&= \int \lim_{n\uparrow\infty} g_n d\nu \\
&= \int g d\nu.
\end{aligned}$$

Finalmente si h es cualquier función medible escribimos $h = h^+ - h^-$, y:

$$\begin{aligned}
\int \frac{d\nu}{d\mathfrak{m}} h d\mathfrak{m} &= \int \frac{d\nu}{d\mathfrak{m}} \left(h^+ - h^-\right) d\mathfrak{m} \\
&= \int \frac{d\nu}{d\mathfrak{m}} h^+ d\mathfrak{m} \;\; - \;\; \int \frac{d\nu}{d\mathfrak{m}} h^- d\mathfrak{m} \\
&= \int h^+ d\nu \;\; - \;\; \int h^- d\nu \\
&= \int \left(h^+ - h^-\right) d\nu \\
&= \int h d\nu.
\end{aligned}$$

Para el segundo punto se tiene para todo $A \in \mathcal{F}$:

$$\int_A \frac{d\mu}{d\mathfrak{m}} d\mathfrak{m} = \mu(A) = \int_A \frac{d\mu}{d\nu} d\nu = \int_A \frac{d\mu}{d\nu} \frac{d\nu}{d\mathfrak{m}} d\mathfrak{m}.$$

De donde c.s.-$\mathfrak{m}$ se tiene:

$$\frac{d\mu}{d\mathfrak{m}} = \frac{d\mu}{d\nu} \frac{d\nu}{d\mathfrak{m}}.$$

Y se concluye el segundo punto si $d\nu/d\mathfrak{m} \neq 0$ casi seguramente $\mathfrak{m}$. □

Si $\mathcal{F}_1$ y $\mathcal{F}_2$ son dos σ-álgebras, denotamos:

$$\mathcal{F}_1 \otimes \mathcal{F}_2 := \sigma\{A \times B : A \in \mathcal{F}_1 \ y \ B \in \mathcal{F}_2\}.$$

Además $\mu \times \nu$ denota la única medida ([Roy68] pág. 265) sobre $\mathcal{F}_1 \otimes \mathcal{F}_2$ tal que $(\mu \times \nu)(A \times B) = \mu(A)\nu(B)$, para todo $A \in \mathcal{F}_1$, $B \in \mathcal{F}_2$.

Definición A.2. *Sean $(\Omega_i, \mathcal{F}_i, \mu_i)$, $i \in \{1, 2\}$, espacios de medida; consideremos el espacio $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2, \mu_1 \times \mu_2)$, sea $E \in \mathcal{F}_1 \otimes \mathcal{F}_2$, dado $x \in \Omega_1$ definimos $E_x := \{y \in \Omega_2 | (x, y) \in \Omega_1 \times \Omega_2\}$, dado $y \in \Omega_2$ definimos $E_y := \{x \in \Omega_1 | (x, y) \in \Omega_1 \times \Omega_2\}$.*

Las demostraciones del *lema A.4*, la *proposición A.1*, *teorema A.5* y *teorema A.6* están en el libro de Royden [Roy68] en el capítulo 12 sección 4. Estos resultados se aplican de manera análoga con E_y.

Lema A.4. *Consideremos todo el contexto de la definición A.2 justo arriba. Sea $E \in \mathcal{F}_1 \otimes \mathcal{F}_2$ tal que $(\mu_1 \times \mu_2)(E) = 0$. Entonces para casi todo $x \in \Omega_1$ se tiene $\mu_2(E_x) = 0$.*

Proposición A.1. *En el mismo contexto sea $E \in \mathcal{F}_1 \otimes \mathcal{F}_2$ tal que $(\mu_1 \times \mu_2)(E) < \infty$. Entonces para casi todo $x \in \Omega_1$ el conjunto E_x es $\mathcal{F}_2$-medible. Además la función definida como*

$$g(x) = \mu_2(E_x).$$

es una función medible definida para casi todo $x \in \Omega_1$; y

$$\int_{\Omega_1} g \, d\mu_1 = (\mu_1 \times \mu_2)(E).$$

Los teoremas de Fubini y Tonelli, son muy usados en el contexto de σ-álgebras y medidas producto. Un espacio de medida $(\Omega, \mathcal{F}, \mu)$ es completo sí y solo si $A \subseteq N$ y $\mu(N) = 0$ implica $A \in \mathcal{F}$.

Teorema A.5. *[Fubini] Sean $(\Omega_1, \mathcal{F}_1, \mu)$, $(\Omega_2, \mathcal{F}_2, \nu)$ dos espacios de medida completos y f una función integrable en $\Omega_1 \times \Omega_2$. Entonces:*

i.- Para casi todo u la función f_u definida por $f_u(v) = f(u, v)$ es una función integrable en Ω_2.

i'.- Para casi todo v la función f_v definida por $f_v(u) = f(u, v)$ es una función integrable en Ω_1.

ii.- $\int_{\Omega_2} f(u, v) d\nu(v)$ es una función integrable en Ω_1.

ii'.- $\int_{\Omega_1} f(u, v) d\mu(u)$ es una función integrable en Ω_2.

iii.- $\int_{\Omega_1} \left[\int_{\Omega_2} f d\nu \right] d\mu = \int_{\Omega_1 \times \Omega_2} f d(\nu \times \mu) = \int_{\Omega_2} \left[\int_{\Omega_1} f d\mu \right] d\nu$.

Teorema A.6. *[Tonelli] Sean $(\Omega_1, \mathcal{F}_1, \mu)$, $(\Omega_2, \mathcal{F}_2, \nu)$ dos espacios de medida completos y f una función medible no negativa en $\Omega_1 \times \Omega_2$. Entonces:*

i.- Para casi todo u la función f_u definida por $f_u(v) = f(u,v)$ es una función integrable en Ω_2.

i'.- Para casi todo v la función f_v definida por $f_v(u) = f(u,v)$ es una función integrable en Ω_1.

ii.- $\int_{\Omega_2} f(u,v)d\nu(v)$ es una función integrable en Ω_1.

ii'.- $\int_{\Omega_1} f(u,v)d\mu(u)$ es una función integrable en Ω_2.

iii.- $\int_{\Omega_1}\left[\int_{\Omega_2} f d\nu\right]d\mu = \int_{\Omega_1\times\Omega_2} f\, d(\nu\times\mu) = \int_{\Omega_2}\left[\int_{\Omega_1} f d\mu\right]d\nu$.

A continuación se exponen algunos resultados sobre medida que usaremos más adelante.

Lema A.7. *Sean $(\Omega_i, \mathcal{F}_i)$, $i \in \{1,2\}$, espacios medibles con medidas μ, ν finitas, sobre $(\Omega_1, \mathcal{F}_1)$, y medidas μ', ν' finitas sobre $(\Omega_2, \mathcal{F}_2)$ tales que $\mu \ll \nu$ y $\mu' \ll \nu'$ entonces $\mu\times\mu' \ll \nu\times\nu'$.*

Demostración. Sea $E \in \sigma(\Omega_1\times\Omega_2)$ supongamos $(\nu\times\nu)'(E) = 0$, entonces por el *lema A.4* tenemos $\nu'(E_\omega) = 0$ c.s.-ν en la variable ω. Por hipótesis $\mu' \ll \nu'$, por tanto $\mu'(E_\omega) = 0$ c.s.-ν, también $\mu \ll \nu$, de donde $\mu\{\omega \in \Omega_1 : \mu'(E_\omega) \neq 0\} = 0$ por tanto tenemos $\mu'(E_\omega) = 0$ c.s.-μ finalmente por la *Proposición A.1*:

$$(\mu\times\mu)'(E) = \int \mu'(E_\omega)d\mu = 0.$$

□

Lema A.8. *Continuando*

$$d(\mu\times\mu')/d(\nu\times\nu') = (d\mu/d\nu)\,(d\mu'/d\nu').$$

Demostración. Sea $A\times B \subseteq \Omega_1\times\Omega_2$ rectángulo medible, usando el *lema A.3* y el teorema de Fubini tenemos:

$$\begin{aligned}\int_{A\times B}\frac{d(\mu\times\mu')}{d(\nu\times\nu')}d(\nu\times\nu') &= (\mu\times\mu')(A\times B)\\ &= \mu(A)\mu'(B)\\ &= \int_A \frac{d\mu}{d\nu}d\nu\cdot\int_B\frac{d\mu'}{d\nu'}d\nu'\\ &= \int_{A\times B}\frac{d\mu}{d\nu}\frac{d\mu'}{d\nu'}d(\nu\times\nu').\end{aligned}$$

Ahora consideremos el conjunto:

$$\mathfrak{D} = \left\{ D \text{ medible de } \{0,1\}^{\Omega_1\times\Omega_2} \;\middle|\; \int_D \frac{d(\mu\times\mu')}{d(\nu\times\nu')}d(\nu\times\nu') = \int_D \frac{d\mu}{d\nu}\frac{d\nu'}{d\mu'}d(\nu\times\nu') \right\}.$$

Veamos que el conjunto es de Dynkin.

a) $\Omega_1 \times \Omega_2 \in \mathfrak{D}$ porque es un rectángulo medible.

b) Consideremos $E, F \in \mathfrak{D}$, $F \subseteq E$:

$$\begin{aligned}\int_{E-F} \frac{d\mu}{d\nu}\frac{d\mu'}{d\nu'} d(\nu\times\nu') &= \int \frac{d\mu}{d\nu}\frac{d\mu'}{d\nu'}\left(\mathbf{1}_E - \mathbf{1}_F\right) d(\nu\times\nu') \\ &= \int_E \frac{d\mu}{d\nu}\frac{d\mu'}{d\nu'} d(\nu\times\nu') - \int_F \frac{d\mu}{d\nu}\frac{d\mu'}{d\nu'} d(\nu\times\nu') \\ &= \int_E \frac{d(\mu\times\mu')}{d(\nu\times\nu')} d(\nu\times\nu') - \int_F \frac{d(\mu\times\mu')}{d(\nu\times\nu')} d(\nu\times\nu') \\ &= \int_{E-F} \frac{d(\mu\times\mu')}{d(\nu\times\nu')} d(\nu\times\nu').\end{aligned}$$

c) Sea $\{E_n\}_{n\in\mathbb{N}} \subseteq \mathfrak{D}$; consideremos $\{F_j\}_{j\in\mathbb{N}}$ donde $F_j = E_n - \bigcup_{n=1}^{j-1} E_n$, tenemos:

$$\begin{aligned}\int_{\bigcup_{j=1}^{\infty} E_n} \frac{d(\mu\times\mu')}{d(\nu\times\nu')} d(\nu\times\nu') &= \int_{\bigcup_{j=1}^{\infty} F_n} \frac{d(\mu\times\mu')}{d(\nu\times\nu')} d(\nu\times\nu') \\ &= \sum_{j=1}^{\infty} \int_{F_j} \frac{d(\mu\times\mu')}{d(\nu\times\nu')} d(\nu\times\nu') \\ &= \sum_{j=1}^{\infty} \int_{F_j} \frac{d\mu}{d\nu}\frac{d\mu'}{d\nu'} d(\nu\times\nu') \\ &= \int_{\bigcup_{j=1}^{\infty} F_j} \frac{d\mu}{d\nu}\frac{d\mu'}{d\nu'} d(\nu\times\nu').\end{aligned}$$

Por a), b) y c); $\mathfrak{D}$ es de Dynkin y al aplicar el teorema de clases monótonas se concluye que para todo conjunto medible E se tiene:

$$\int_E \frac{d(\mu\times\mu')}{d(\nu\times\nu')} d(\nu\times\nu') = \int_E \frac{d\mu}{d\nu}\frac{d\mu'}{d\nu'} d(\nu\times\nu').$$

Tenemos entonces $d(\mu\times\mu')/d(\nu\times\nu') = (d\mu/d\nu)\,(d\mu'/d\nu')$. *c.s.* $-\ \nu\times\nu'$. □

Lema A.9. *Sea $(\Omega, \mathcal{F})$ un espacio medible; μ, ν medidas tales que $\mu \ll \nu$, $\mathcal{G}$ sub-sigma álgebra de $\mathcal{F}$. Se cumple:*

$$E\left[\frac{d\mu}{d\nu}\,|\mathcal{G}\right] = \frac{d\mu|_{\mathcal{G}}}{d\nu|_{\mathcal{G}}} \qquad \nu - c.s.$$

Demostración. Comenzamos por revisar que para $X : \Omega \to \mathbb{R}$ que es $\mathcal{G}$-medible se cumple para todo $B \in \mathcal{G}$:

$$\int_B X d\nu|_G = \int_B X d\nu.$$

Esto es sencillo, observando que para todo A, $B \in \mathcal{G}$

$$\int_B \mathbf{1}_A d\nu = \nu\left(A \cap B\right) = \nu|_G\left(A \cap B\right) = \int_B \mathbf{1}_A d\nu|_G.$$

Por la linealidad de la integral esto mismo se cumple para variables aleatorias simples, en consecuencia para variables aleatorias positivas, y luego para toda variable aleatoria.
Si $\mu \ll \nu$, también $\mu|_{\mathcal{G}} \ll \nu|_{\mathcal{G}}$, si ν es sigma finita, también lo es $\nu|_{\mathcal{G}}$, por tanto podemos escribir para todo $B \in \mathcal{G}$:

$$\int_B \frac{d\mu}{d\nu} d\nu = \mu(B) = \mu|_{\mathcal{G}}(B) = \int_B \frac{d\mu|_{\mathcal{G}}}{d\nu|_{\mathcal{G}}} d\nu|_{\mathcal{G}} = \int_B \frac{d\mu|_{\mathcal{G}}}{d\nu|_{\mathcal{G}}} d\nu.$$

Por tanto $E\left[\frac{d\mu}{d\nu}\,|\mathcal{G}\right] = \frac{d\mu|_{\mathcal{G}}}{d\nu|_{\mathcal{G}}}$ c.s.-ν. □

Apéndice B

Funciones Lipschitz

Las funciones Lipschitz y el análisis convexo juegan un papel fundamental en el estudio de la privacidad diferencial, ya que proporcionan herramientas matemáticas precisas para controlar la sensibilidad de los algoritmos a cambios en los datos individuales. En particular, una función Lipschitz acotada garantiza que pequeñas variaciones en la entrada no produzcan grandes fluctuaciones en la salida, lo cual es esencial para limitar la cantidad de información que puede filtrarse sobre un individuo específico. Por otro lado, el análisis convexo permite formular y resolver problemas de optimización que surgen en mecanismos de privacidad, como el diseño de ruido óptimo o la caracterización de garantías de privacidad en algoritmos iterativos. La combinación de ambas herramientas permite desarrollar mecanismos diferenciales robustos, eficientes y con garantías matemáticas verificables.

Definición B.1. *Consideremos* $(\mathcal{Z}_1, ||\cdot||_1)$, $(\mathcal{Z}_2, ||\cdot||_2)$ *espacios de Banach,* $f : \mathcal{Z}_1 \to \mathcal{Z}_2$ *decimos que* f *es* K*-Lipschitz, si existe* $K > 0$ *tal que:*

$$||f(x) - f(y)||_2 \leq K||x - y||_1 \qquad \textit{para todo } x, y \in \mathcal{Z}_1.$$

La forma más general de esta definición es en espacios métricos, en este trabajo se usará para espacios de Banach.

Recordamos también que un campo escalar diferenciable $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ tiene asociado el campo vectorial $\nabla f : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ dado por:

$$\nabla f(\mathbf{x}) = \left(\frac{\partial f(\mathbf{x})}{\partial x_1}, \ldots, \frac{\partial f(\mathbf{x})}{\partial x_n} \right)$$

llamado gradiente de f.

Teorema B.1. *Sea* $\mathcal{K}$ *un conjunto convexo en* $\mathbb{R}^n$*. Definimos el operador proyección como:*

$$\Pi_{\mathcal{K}}(x) = \min_{y \in \mathcal{K}} d(x, y).$$

Entonces $\Pi_{\mathcal{K}}$ *es una contracción.*

Demostración. Consideremos $x, z \in \mathbb{R}^n$ entonces:

$$||\Pi_{\mathcal{K}}(x) - \Pi_{\mathcal{K}}(z)|| = || \min_{y \in \mathcal{K}} d(x, y) - \min_{y \in \mathcal{K}} d(z, y)||.$$

En primer lugar notemos que para todo $w \in \mathcal{K}$ se tiene $\langle z - \Pi_{\mathcal{K}}(z), w - \Pi_{\mathcal{K}}(z)\rangle \leq 0$, cuando $z \in \mathcal{K}$ es evidente; para el caso $z \in \mathcal{K}^c$ tenemos que $z - \Pi_{\mathcal{K}}(z)$ es un vector normal a $\mathcal{K}$, por ser $\mathcal{K}$ convexo; el ángulo entre $z - \Pi_{\mathcal{K}}(z)$ y $w - \Pi_{\mathcal{K}}(z)$ está entre $\pi/2$ y π por tanto se tiene la desigualdad con respecto al producto interior canónico en $\mathbb{R}^n$ que esta en función del coseno del ángulo. Para una demostración analítica de este hecho, que involucra otros conceptos de convexidad consultar *Lema 3.1.4* en [Nes04].Por las mismas razones tenemos $\langle x - \Pi_{\mathcal{K}}(x), w' - \Pi_{\mathcal{K}}(x)\rangle \leq 0$ para todo $w' \in \mathcal{K}$. Sumando ambas expresiones tenemos

$$\langle z - \Pi_{\mathcal{K}}(z), w - \Pi_{\mathcal{K}}(z)\rangle + \langle x - \Pi_{\mathcal{K}}(x), w' - \Pi_{\mathcal{K}}(x)\rangle \leq 0;$$

sustituyendo $w = \Pi_{\mathcal{K}}(x)$ y $w' = \Pi_{\mathcal{K}}(z)$ se sigue

$$\begin{aligned} 0 \geq & \langle z - \Pi_{\mathcal{K}}(z), \Pi_{\mathcal{K}}(x) - \Pi_{\mathcal{K}}(z)\rangle + \langle x - \Pi_{\mathcal{K}}(x), \Pi_{\mathcal{K}}(z) - \Pi_{\mathcal{K}}(x)\rangle \\ = & \langle z - \Pi_{\mathcal{K}}(z), \Pi_{\mathcal{K}}(x) - \Pi_{\mathcal{K}}(z)\rangle - \langle x - \Pi_{\mathcal{K}}(x), \Pi_{\mathcal{K}}(x) - \Pi_{\mathcal{K}}(z)\rangle \\ = & \langle z - x - \Pi_{\mathcal{K}}(z) + \Pi_{\mathcal{K}}(x), \Pi_{\mathcal{K}}(x) - \Pi_{\mathcal{K}}(z)\rangle \\ = & \langle z - x, \Pi_{\mathcal{K}}(x) - \Pi_{\mathcal{K}}(z)\rangle + \langle \Pi_{\mathcal{K}}(x) - \Pi_{\mathcal{K}}(z), \Pi_{\mathcal{K}}(x) - \Pi_{\mathcal{K}}(z)\rangle . \end{aligned}$$

De donde

$$\begin{aligned} ||\Pi_{\mathcal{K}}(x) - \Pi_{\mathcal{K}}(z)||^2 &= \langle \Pi_{\mathcal{K}}(x) - \Pi_{\mathcal{K}}(z), \Pi_{\mathcal{K}}(x) - \Pi_{\mathcal{K}}(z)\rangle \\ &\leq - \langle z - x, \Pi_{\mathcal{K}}(x) - \Pi_{\mathcal{K}}(z)\rangle \\ &= \langle x - z, \Pi_{\mathcal{K}}(x) - \Pi_{\mathcal{K}}(z)\rangle . \end{aligned}$$

Usando la desigualdad de Cauchy-Schwarz

$$\langle x - z, \Pi_{\mathcal{K}}(x) - \Pi_{\mathcal{K}}(z)\rangle \leq ||x - z||\, ||\, \Pi_{\mathcal{K}}(x) - \Pi_{\mathcal{K}}(z)||.$$

Se concluye

$$\begin{aligned} ||\Pi_{\mathcal{K}}(x) - \Pi_{\mathcal{K}}(z)||^2 &\leq ||x - z||\, ||\, \Pi_{\mathcal{K}}(x) - \Pi_{\mathcal{K}}(z)|| \\ ||\Pi_{\mathcal{K}}(x) - \Pi_{\mathcal{K}}(z)|| &\leq ||x - z||. \end{aligned}$$

□

Teorema B.2. *Sea $f : \mathbb{R}^d \to \mathbb{R}$ es convexa con gradiente $\beta -$ Lipschitz. Entonces la función ψ definida como*

$$\psi(w) = w - \eta \nabla f(w)$$

es una contracción si $\eta \leq 2/\beta$.

Demostración.

$$\begin{aligned} &||w - \eta \nabla f(w) - w' + \eta \nabla f(w')||^2 = ||w - w' - \eta(\nabla f(w) - \nabla f(w'))||^2 \\ &\leq ||w - w'||^2 - 2\eta \langle \nabla f(w) - \nabla f(w'), w - w'\rangle + \eta^2 ||\nabla f(w) - \nabla f(w')||^2. \end{aligned}$$

Ahora, por el lema de *Baillon-Haddad corolario 18.16 en* [BC11], f cumple

$$\langle \nabla f(w) - \nabla f(w'), w - w'\rangle \geq \frac{1}{\beta} ||\nabla f(w) - \nabla f(w')||^2;$$

entonces tenemos

$$\begin{aligned}
&||w - \eta\nabla f(w) - w' + \eta\nabla f(w')||^2 \\
&\leq ||w - w'||^2 - 2\eta\frac{1}{\beta}||\nabla f(w) - \nabla f(w')||^2 + \eta^2||\nabla f(w) - \nabla f(w')||^2 \\
&= ||w - w'||^2 + \left(\eta^2 - 2\frac{\eta}{\beta}\right)||\nabla f(w) - \nabla f(w')||^2.
\end{aligned}$$

Ahora, el gradiente es β-Lipschitz, entonces

$$\begin{aligned}
&||w - \eta\nabla f(w) - w' + \eta\nabla f(w')||^2 \\
&\leq ||w - w'||^2 + \left(\eta^2 - 2\frac{\eta}{\beta}\right)\beta^2||w - w'||^2 \\
&= \left(1 + \eta^2\beta^2 - 2\eta\beta\right)||w - w'||^2 \\
&= (1 - \eta\beta)^2\,||w - w'||^2.
\end{aligned}$$

Finalmente llegamos a

$$||\psi(w) - \psi(w')|| \leq |1 - \eta\beta|\,||\, w - w'||.$$

Considerando $\eta \leq 2/\beta$; $\eta\beta - 1 \leq 1$. Por tanto ψ es una contracción. □

Bibliografía

[Sha48] C. Shannon. A Mathematical Theory of Communication. *The Bell System Technical Journal, vol. 27* (1948).

[S K51] R. A. Leibler S. Kullback. On Information and Sufficiency. *The Annals of Mathematical Statistics vol. 22* (1951).

[Ren61] A. Renyi. On Measures of Entropy and Information. *Berkeley Symp. on Math. Statist. and Prob.* (1961).

[Roy68] H. L. Royden. *Real Analysis second edition.* The Macmillan Company, 1968.

[LV87] Friederich Liese e Igor Vajda. Convex statistical distances. *Teubner* (1987).

[Vil03] C. Villani. *Topics in Optimal Transportation.* American Mathematical Society, 2003.

[Nes04] Yurii Nesterov. *Introductory Lectures on Convex Optimization.* Springer Science, 2004.

[Gra09] Guillermo Grabinsky. *Teoría de la Medida.* Facultad de Ciencias, UNAM, 2009.

[Erv10] Tim van Erven. Rényi Divergence and Kullback-Leibler Divergence. *IEEE* (2010).

[BC11] Heinz H. Bauschke y Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces.* Springer Science, 2011.

[Fel+18] Vitaly Feldman, Ilya Mironov, Kunual Talwar y Abhradeep Thakurta. Privacy Amplification by Iteration. *IEEE 59th Annual Symposium on Foundations of Computer Science* (2018).

Índice alfabético