FACULTAD DE MATEMÁTICAS

ALGORITMOS DIFERENCIALMENTE PRIVADOS PARA OPTIMIZACIÓN CONVEXA ESTOCÁSTICA SOBRE LA BOLA NUCLEAR

por

JUAN PABLO FLORES MELLA

Tesis presentada a la Facultad de Matemáticas
de la Pontificia Universidad Católica de Chile,
para optar al grado de Magíster en Matemáticas

Profesor guía: Cristóbal Guzmán Paredes

Septiembre, 2022

Santiago, Chile

Resumen

En este trabajo investigamos métodos de resolución privada y aproximada del problema de optimización convexa y estocástica sobre el espacio de matrices restringido a la bola nuclear. Todo esto con el fin de hallar una cota de error que conjeturamos óptima $\tilde{O}\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d}}{\varepsilon n}\right)$. Creemos que el problema puede ser resuelto con esta tasa de error por el algoritmo de Frank-Wolfe estocástico, el cual requiere resolver una subrutina de programación semidefinida para su implementación; esto se traduce en el problema de maximizar (o minimizar) una forma cuadrática $x^\top Ax$.
Debido a esto, exploramos dos métodos para resolver aproximada y privadamente esta subrutina semidefinida: El primero consiste en muestrear una solución del problema usando una nueva variante del mecanismo exponencial. Como veremos, la aleatoriedad de tal muestreo implicará la privacidad de la subrutina.
El segundo, por su parte, consiste en implementar una versión privada del algoritmo de Oja. La privacidad, en este caso, vendrá dada por la perturbación de cada iterado mediante ruido Gaussiano.

Finalmente, nuestra indagación concluye que el primer método –el del muestreo– no satisface el criterio de ser una buena solución aproximada. En cambio, el segundo método sí logra dar una solución privada y aproximada, aunque –creemos– subóptima, pues no es la conjeturada.

Índice general

Capítulo 1

Introducción

La presente tesis tiene como propósito indagar nuevos métodos para resolver el problema de optimización (convexa) estocástica y privada (DP-SCO, por su sigla en inglés) de una función convexa, Lipschitz y suave, $f : \mathbb{R}^{d_1 \times d_2} \times \mathcal{Z} \to \mathbb{R}$, sobre el espacio de matrices $\mathbb{R}^{d_1 \times d_2}$ restringido a la bola nuclear unitaria. La bola nuclear, en particular, suscita interés debido a que es una relajación para problemas que involucran matrices de bajo rango. La restricción de «bajo rango» sobre las matrices es una restricción no-convexa –pues el rango es un objeto discreto–, la «convexificación» que resulta de ella es la norma nuclear [11].

Un ámbito donde aparecen las matrices de bajo rango es el de las matrices de recomendación –por ejemplo recomendación de películas en plataformas de *streaming*, para ponernos en un caso concreto– donde se detallan las preferencias de distintos usuarios. Una hipótesis usual es que las preferencias de la población se pueden aproximar por una matriz simple, lo que se traduce en que esta sea de bajo rango. Trabajos como el de Jain et al. [14] aprovechan la heurística de la bola nuclear para resolver privadamente el problema de *completación de matrices*: A partir del conocimiento de la puntuación de unos cuantos usuarios, con respecto a unas pocas películas, encontrar la matriz de recomendación «completa» (donde están las valoraciones de todos los usuarios respecto a todas las películas), la cual, por las consideraciones anteriores, se asume de bajo rango.

Comenzaremos la tesis con un breve repaso de privacidad diferencial y de algunas herramientas geométricas para obtener concentración de la medida mediante la desigualdad de log-Sobolev. Esto último requiere calcular gradientes y Hessianos en el sentido Riemanniano, por lo que añadimos un apéndice con propiedades básicas de ellos y cómo obtenerlos a partir de gradientes Euclidianos. A lo largo de esta presentación, asumimos conocimientos básicos de geometría, funciones y optimización convexa.

Desmenucemos la temática de la tesis –optimización (convexa) estocástica y privada–, para dar una mejor idea de qué es lo que se trata. Comencemos con la optimización estocástica.

Entendemos por optimización convexa [8] al área que busca estudiar y resolver los problemas del tipo

$$\begin{aligned} \text{mín} \quad & f(x) \\ \text{s.a} \quad & x \in \mathcal{X}, \end{aligned}$$

donde $f : \mathcal{X} \to \mathbb{R}$ es una función convexa y $\mathcal{X}$ es un conjunto convexo y cerrado. Salvo por contadas excepciones, encontrar una solución exacta para este tipo de problemas es imposible, mas esto no es el fin del camino, pues existen diversos métodos para encontrar soluciones aproximadas tan «buenas»[1] como se requiera.

La variante estocástica busca exactamente lo mismo:

$$\begin{aligned} \text{mín} \quad & F_{\mathcal{D}}(x) \\ \text{s.a} \quad & x \in \mathcal{X}, \end{aligned}$$

donde $F_{\mathcal{D}}$ es una función convexa y $\mathcal{X}$ es un conjunto convexo y cerrado, con la diferencia que la manera en la que accedemos a la función $F_{\mathcal{D}}$ es distinta. En el caso estocástico no trabajamos directamente con la función $F_{\mathcal{D}}$, sino que con una familia de funciones $f(\cdot, z)$, donde $z \in \mathcal{Z}$ es un parámetro, que se relacionan con $F_{\mathcal{D}}$ mediante la fórmula

$$F_{\mathcal{D}}(x) = \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}\left[f(x, \mathbf{z})\right],$$

donde $\mathcal{D}$ es una distribución sobre $\mathcal{Z}$.
Puede que, en general, no tengamos conocimiento directo de la distribución $\mathcal{D}$ y, por ello, no podamos calcular $F_{\mathcal{D}}(x) = \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}\left[f(x, \mathbf{z})\right]$. Por lo cual nuestro acceso a la función $F_{\mathcal{D}}$ vendrá dado por una muestra $S = (z_1, \ldots, z_n) \sim \mathcal{D}^n$, la cual asumimos i.i.d. En este escenario podemos encontrar, por ejemplo, los problemas de regresión lineal, donde la función f es cuadrática [20].

Mencionemos, por último, la componente privada del problema. Buscamos que nuestro problema sea privado en el sentido de *privacidad diferencial*; esto significa que, dadas bases de datos «similares», esperamos que el resultado obtenido en los procesos en paralelo sea similar.
La teoría de privacidad diferencial no está diseñada específicamente para los algoritmos de optimización, mas su versatilidad permite incorporarla al proceso, de manera que garantice la privacidad del resultado.

Organización del texto

La tesis se divide en dos grandes bloques: Preliminares y Privacidad en la bola nuclear. A esto se le suma, por completitud, un apéndice de cálculo en variedades Riemannianas.

1. **Preliminares.** El capítulo de preliminares se subdivide, a su vez, en tres partes: La primera, de privacidad, introduce los conceptos y propiedades básicas de privacidad-diferencial. A esto

[1] Aquí «bueno» significa que $f(x) - f(x^*)$ sea pequeño, donde x^* es el minimizador de la función.

se le suman algunos de los mecanismos de privacidad más conocidos (relevantes para distintas partes de esta tesis) y los teoremas de composición, que cuantifican la privacidad de un algoritmo ensamblado con distintas partes diferencialmente-privadas.
La segunda parte contiene las técnicas de concentración de la medida que ocuparemos. Aquí se introducen conceptos tales como condición de curvatura-dimensión (también conocida como criterio de Bakry-Émery) y la desigualdad de log-Sobolev. La presencia de esta última se traduce en desigualdades de concentración sub-gaussiana para la medida.
Finalmente, en la tercera parte, se plantea formalmente el problema que estamos interesados en resolver, así como una reducción de este.

2. **Privacidad en la bola nuclear.** El capítulo de privacidad en la bola nuclear, por su parte, se subdivide en la presentación de dos técnicas: La primera, basada en un nuevo análisis del mecanismo exponencial propuesto por Minami et al. [18], busca –a grandes rasgos– encontrar soluciones mediante muestreos en la esfera unitaria, haciendo uso de una medida de probabilidad perturbada, conocida como distribución de Bingham en la esfera. Esta técnica, sin embargo, falla en dar una solución al problema.
La segunda técnica consiste en la versión del algoritmo de Oja presentado en Jain et al. [14]. Este algoritmo «soporta» la presencia de ruido en la matriz con la que queremos trabajar, cuantificando el error adicional en el que se incurre debido a la presencia de este. Esta técnica sí entrega una solución al problema, aunque la estimamos subóptima.

3. **Apéndice.** En el apéndice aparecen los teoremas y definiciones de geometría y cálculo en variedades Riemannianas, que si bien no influyen directamente en la construcción de una solución, sí aparecen –en un nivel más técnico– en los diversos cálculos que deben llevarse a cabo para el análisis de la privacidad de la primera técnica.

Literatura relacionada

Mencionamos a continuación algunos de los trabajos que abordan el problema de DP-SCO en el escenario en que $f(\cdot, z)$ es una función suave. En el contexto de los espacios $\mathcal{X} = \ell^p$: $p = 2$ (o sea, el caso Euclidiano) es abordado por Bassily et al. [3], donde logran alcanzar la tasa óptima mediante una variante de SGD; $p = 1$ es estudiado tanto por Bassily et al. [4], donde obtienen una tasa de error casi óptima mediante una variante privada del algoritmo de Frank-Wolfe estocástico, como por Asi et al. [1], donde alcanzan la tasa óptima usando localización iterativa en conjunto con descenso reflejado. Los casos $p \in (1, 2)$ y $p \in (2, \infty]$ son estudiados en Bassily et al. [4].

En el contexto matricial podemos encontrar los artículos de Kapralov y Talwar [15] y Jain et al. [14]. El primero presenta un algoritmo basado en el mecanismo exponencial para encontrar una aproximación de rango k y ε-DP de una matriz A. El segundo ($\mathcal{X} = \left\{Y \in \mathbb{R}^{d_1 \times d_2} : \|Y\|_{\text{Nuc}} \leq k\right\}$), por su parte, estudia el problema de completación de matrices (descrito brevemente arriba), pero con una variante distinta de privacidad diferencial, conocida como *joint DP*.

Capítulo 2

Preliminares

2.1. Privacidad Diferencial

2.1.1. El premio de Netflix

El 2 de Octubre de 2006, Netflix anunció el *Netflix prize*, un concurso dotado de un premio de un millón de dólares y que tenía como fin mejorar el sistema de recomendación de series y películas de la compañía. Para ayudar a los participantes en el diseño del modelo predictivo, la compañía liberó una base de datos (el *training data set*) «anonimizada» con las preferencias desde 1999 hasta Diciembre de 2005 de 480.189 suscriptores de Netflix [19]. Un ganador fue elegido en Septiembre de 2009 y se realizaron planes para una segunda versión del concurso. Sin embargo, a los pocos meses, Netflix debió abandonar sus pretensiones y lo canceló: cuatro suscriptores demandaron a Netflix alegando que la compañía había infringido la *Video Privacy Protection Act*.

La publicación en 2006 de la base de datos por parte de Netflix levantó sospechas acerca de la potencial violación a la privacidad de los suscriptores dentro de esta (de la base de datos). De acuerdo con Netflix, los datos no contenían información identificadora, solo la valoración de películas del usuario y las fechas en las que las valoró. Podemos pensar el registro de un usuario como un vector fila del estilo

$$(\text{ID anonimizada}, \text{Valoración película 1}, \text{Fecha de valoración}, \text{Valoración película 2}, \text{Fecha valoración}, \ldots).$$

Incluso si alguien conociera todas las valoraciones de un usuario y las fechas en las que las realizó, esta persona –según Netflix– no podría reidentificarlo en la base de datos, pues solo se tomó una pequeña muestra del total y, además, los datos fueron perturbados.
Sin embargo, esto no bastó. En el año 2007, Narayanan y Shmatikov [19] probaron que con el uso de un poco de *información pública*, era posible reidentificar a una fracción sustancial de suscriptores dentro de la base de datos liberada por Netflix[1]. En particular, mostraron que uno puede usar

[1] En el estudio que realizaron se ponen en distintos escenarios de error en los que se puede incurrir. Obteniendo resultados tan sorprendentes como: Un adversario que conoce dos valoraciones (exactas) de un suscriptor objetivo y

otras bases de datos públicas, en este caso IMDb (Internet Movie Data base), para efectivamente reidentificar a una porción sustantiva de miembros de la base de datos.

Así, a partir de la información de los gustos de unas pocas películas de una determinada persona, podemos acceder a un vasto catálogo de películas que ha visto, así como sus opiniones (si le puso un 1 o un 5) de ellas. De estos datos podemos inferir información potencialmente sensible, como: la orientación sexual de la persona, sus preferencias políticas, su religión, etc. ¡Una clara violación a su privacidad!

2.1.2. Privacidad diferencial

A lo largo de esta sección supondremos que hay un *conservador confiable* (o curador) que tiene acceso a la base de datos tal cual es, una suerte de intermediario de esta. Para acceder a la información en la base de datos, debemos dirigir nuestras preguntas al conservador y él las responderá de manera «privada».
Irremediablemente, para mantener la privacidad de la base de datos, la respuesta del conservador no podrá ser fidedigna, debe contener imprecisiones. Por otro lado, para que la respuesta nos sea útil, las imprecisiones no pueden ser mayúsculas. Un ambiente en silencio nos permite entender claramente el mensaje, pero quizás al nivel de comprometer la privacidad de los participantes; un ambiente en el que todos hablan no nos permite entender nada: Debemos encontrar un balance entre *privacidad* y *precisión*.

Definición 2.1.1 (Privacidad diferencial «pura»)**.** Decimos que un algoritmo aleatorizado, $\mathcal{M} : \mathcal{Z}^n \to \mathcal{R}$, es ε-DP (ε-Diferencialmente Privado) si para todo par de base de datos S, $S' \in \mathcal{Z}^n$ que difieren en a lo más una entrada, y para todo evento (conjunto medible) $\mathcal{O} \subseteq \mathcal{R}$ se tiene que:

$$\mathbb{P}\left[\mathcal{M}(S) \in \mathcal{O}\right] \leq \exp\left(\varepsilon\right) \cdot \mathbb{P}\left[\mathcal{M}(S') \in \mathcal{O}\right] \tag{2.1}$$

Llamaremos *mecanismo* a los algoritmos aleatorizados que buscan asegurar privacidad.

Observación 2.1.1. En la práctica ε suele tomarse como una constante (no depende de la dimensión del problema) ≤ 1.

A raíz de esta definición, introducimos el siguiente concepto:

Definición 2.1.2 (Bases de datos vecinas)**.** Dos bases de datos $S, S' \in \mathcal{Z}^n$ se dicen *vecinas*, denotado $S \simeq S'$, si estas difieren en a lo más un dato.

Intuitivamente, la definición de privacidad diferencial nos dice que si $\mathcal{M}$ es diferencialmente privado, entonces a un analista le será muy difícil poder distinguir si el participante z_i está o no en la base de datos, a partir del resultado del algoritmo. Otra manera de pensarlo es que si $\mathcal{M}$ es

además conoce las fechas en que las realizó, con un margen de error de 3 días, puede lograr reidentificar al 68 % de los miembros de la base de datos.

diferencialmente privado, la información aportada por el participante z_i al mecanismo no es determinante en el resultado que este entrega (aporta, pero no tanto).

La siguiente propiedad nos dice que los algoritmos ε-DP son inmunes al *post-procesamiento*; es decir, una vez el algoritmo entrega su resultado, un analista no puede ingeniárselas para hacer este resultado menos privado. La demostración es sencilla y puede encontrarse en Dwork y Roth [10].

Teorema 2.1.1 (Post-procesamiento [10])**.** Sea $\mathcal{M} : \mathcal{Z}^n \to \mathcal{R}$ un algoritmo ε-DP y sea $f : \mathcal{R} \to \mathcal{R}'$ un mapeo aleatorio. Entonces

$$f \circ \mathcal{M} : \mathcal{Z}^n \to \mathcal{R}'$$

es ε-DP.

2.1.3. Ejemplos de mecanismos diferencialmente privados

Hasta ahora solo hemos visto la definición de mecanismo ε-DP y un par de propiedades que estos tienen. Pasemos a ver casos concretos de ellos, no sin antes realizar la siguiente observación técnica:

Cómo probar que un mecanismo es ε-DP

Para probar que un mecanismo $\mathcal{M}$ es ε-diferencialmente privado no es necesario probar la desigualdad (2.1) para todo evento $\mathcal{O}$:

- Si la distribución es discreta: Basta probar

$$\mathbb{P}\left[\mathcal{M}(S) = r\right] \leq \exp\left(\varepsilon\right) \mathbb{P}\left[\mathcal{M}(S') = r\right] \qquad \forall r \in \mathcal{R}$$

- Si la distribución es absolutamente continua con respecto a una medida λ: Basta probar

$$p(r) \leq \exp\left(\varepsilon\right) q(r) \qquad \forall r \in \mathcal{R},$$

donde $p(r) = \dfrac{d\mathbb{P}_{\mathcal{M}(\mathcal{S})}}{d\lambda}(r)$ y $q(r) = \dfrac{d\mathbb{P}_{\mathcal{M}(\mathcal{S}')}}{d\lambda}(r)$.

Mecanismo Laplaciano

El mecanismo Laplaciano es una manera de privatizar funciones $f : \mathcal{Z}^n \to \mathbb{R}^k$ dependientes de los datos y que entregan resultados numéricos. Supongamos que nuestros datos, S, son obtenidos mediante un estudio de mercado en el que participan 1000 personas, y que busca saber cuáles películas estrenadas en 2022 –digamos que se han estrenado 10– han sido vistas por cada persona encuestada y cuál de ellas ha sido su favorita; o sea, cada uno de los datos de S sería una fila del tipo:

$$\left(\text{ID}, \text{¿Vio película 1?}, \ldots, \text{¿Vio película 10?}, \text{¿Película favorita?}\right).$$

Una pregunta que podría hacerse con esta información es: ¿Cómo se distribuyen las películas favoritas en esta base de datos? Podemos formalizar esta pregunta mediante la función $f : \mathcal{Z}^{1000} \to \mathbb{R}^{10}$

$$f(S) = (\#S \text{ cuya favorita es la película } 1, \ldots, \#S \text{ cuya favorita es la película } 10\,) \tag{2.2}$$

El mecanismo Laplaciano nos dice que podemos responder privadamente esta pregunta si es que perturbamos lo suficiente la respuesta.
Introduzcamos ahora los conceptos relevantes para la definición del mecanismo Laplaciano.

Definición 2.1.3 (Distribución de Laplace)**.** La distribución de Laplace centrada en 0 y escalada por b es la distribución con función de densidad

$$\text{Lap}(x \mid b) = \frac{1}{2b} \exp\left(-\frac{|x|}{b}\right).$$

La media de esta distribución es $\mu = 0$ y su varianza, $\sigma^2 = 2b^2$.
Cuando una variable aleatoria Y tiene densidad $\text{Lap}(x \mid b)$, escribimos $Y \sim \text{Lap}(b)$.

Definición 2.1.4 (Sensibilidad)**.** Sea $f : \mathcal{Z}^n \to \mathbb{R}^k$ una función de los datos y sea $\|\cdot\|$ una norma de $\mathbb{R}^k$. Definimos la sensibilidad de f con respecto a $\|\cdot\|$ como

$$\Delta_{\|\cdot\|} f = \sup_{S, S' \in \mathcal{Z}^n, S \simeq S'} \|f(S) - f(S')\|.$$

Cuando $\|\cdot\| = \|\cdot\|_p$, en vez de usar $\Delta_{\|\cdot\|_p}$, usaremos Δ_p para alivianar la notación.

Dependiendo del mecanismo con el que estemos trabajando, puede que sea útil definir la sensibilidad de una función de manera ligeramente diferente (véase, por ejemplo, el mecanismo exponencial). Sin embargo, todas ellas intentan capturar lo mismo: ver cuánto puede afectar la presencia de un dato en el resultado final del algoritmo.

Definición 2.1.5 (Mecanismo Laplaciano)**.** Dada cualquier $f : \mathcal{Z}^n \to \mathbb{R}^k$ dependiente de los datos, definimos el mecanismo Laplaciano como:

$$\mathcal{M}_L(S, f, \varepsilon) = f(S) + (Y_1, \ldots, Y_k),$$

donde Y_i son variables aleatorias i.i.d. que distribuyen $\text{Lap}(\Delta_1 f/\varepsilon)$.

Teorema 2.1.2 ([10])**.** El mecanismo Laplaciano es ε-DP.

Teorema 2.1.3 ([10])**.** Sea $f : \mathcal{Z}^n \to \mathbb{R}^k$ y sea $y = \mathcal{M}_L(S, f, \varepsilon)$. Luego, para todo $\gamma \in (0, 1]$,

$$\mathbb{P}\left[\|f(S) - y\|_\infty \geq \log\left(\frac{k}{\gamma}\right) \cdot \left(\frac{\Delta_1 f}{\varepsilon}\right)\right] \leq \gamma.$$

Ejemplo 2.1.1. Usando el mismo ejemplo de (2.2): $\Delta_1 f = 1$, pues si reemplazamos uno de los participantes del estudio por uno nuevo, lo «peor» que podría pasar es que su película favorita del 2022 sea distinta de la del participante removido. Por lo tanto, $f(S) + (Y_1, \ldots, Y_{10})$, donde los Y_i son i.i.d $\sim \text{Lap}(1/\varepsilon)$, es ε-DP. Además, con 95 % de probabilidad se tiene un error de a lo más $5.3/\varepsilon$ en cada coordenada.

Report Noisy Max

Consideremos el mismo ejemplo de antes y supongamos que esta vez queremos saber cuál ha sido la película más vista del 2022. Una manera responder esto usando el mecanismo Laplaciano es la siguiente: Calcular la función $f : \mathcal{Z}^{1000} \to \mathbb{R}^{10}$ definida por

$$f(S) = (\#S \text{ que vio la película } 1, \ldots, \#S \text{ que vio la película } 10)\,.$$

Como una persona puede ver múltiples películas, la sensibilidad de f será $\Delta_1 f = 10$ (esto se obtiene si, por ejemplo, cambiamos una persona que no ha visto ninguna película por otra que las ha visto todas). Siguiendo el procedimiento del mecanismo Laplaciano, sumamos a cada coordenada ruido $\sim \text{Lap}(10/\varepsilon)$ y extraemos la mayor de ellas.

Como el vector ruidoso $y = \mathcal{M}_L\,(S, f, \varepsilon)$ es ε-DP y el proceso de calcular la máxima coordenada es independiente de los datos, tenemos, por post-procesamiento (Teorema 2.1.1), que esta solución es ε-DP. Sin embargo, la cantidad de ruido que tuvimos que inyectar para obtener privacidad es considerablemente más grande que la necesaria. Esto se debe a que dimos más información de la que realmente necesitábamos entregar: No es necesario que aquel que quiere saber cuál es la película más vista sepa cuántas personas han visto las otras películas.

Para casos como este definimos *Report Noisy Max*.

Definición 2.1.6 (Report Noisy Max)**.** Sean $f_1, \ldots, f_k : \mathcal{Z}^n \to \mathbb{R}$ funciones de los datos tales que $\Delta_1 f_i \leq \Delta$, para todo $i \in [k]$ y sean $Y_1, \ldots, Y_k$ i.i.d. $\sim \text{Lap}\,(\Delta/\varepsilon)$.[2] Report Noisy Max es el mecanismo que:

1. Calcula $y_i = f_i(S) + Y_i$ para cada $i \in [k]$.
2. Entrega el máximo y_i junto a su etiqueta i.

Teorema 2.1.4 ([10])**.** Report Noisy Max es ε-DP.

Teorema 2.1.5 ([10])**.**

$$\mathbb{P}\left[\max_{i\in[k]} f_i(S) - \max_{i\in[k]} y_i \geq \log\left(\frac{k}{\gamma}\right)\cdot\left(\frac{\Delta}{\varepsilon}\right)\right] \leq \gamma$$

y

$$\mathbb{E}\left[\max_{i\in[k]} f_i(S) - \max_{i\in[k]} y_i\right] = O\left(\Delta \log(k)/\varepsilon\right).$$

Ejemplo 2.1.2. En el ejemplo del inicio, cada $f_i(S) = \#S$ que vio la película i y cada una de ellas tiene sensibilidad $\Delta_1 f_i = 1 = \Delta$. Por ende, al aplicar Report Noisy Max a estas funciones, entregamos un resultado ε-DP y que con 95 % de probabilidad entrega un resultado que discrepa en, a lo más, $(21.2)/\varepsilon$.

[2] Notar que los f_i son funciones real-valuadas. Por lo que la norma ℓ_1 es solo el valor absoluto.

Mecanismo Exponencial

El mecanismo exponencial es un método que garantiza ε-DP eligiendo aleatoriamente (dentro de un conjunto factible) la respuesta que queremos entregar. La aleatoriedad, sin embargo, no es uniforme: no todas las respuestas tienen la misma chance de ser elegidas. La manera en que el mecanismo exponencial escoge su respuesta es recogiendo una muestra aleatoria, pero «cargada» hacia las respuestas más útiles.

En Bassily et al. [6], por ejemplo, usan el mecanismo exponencial para resolver privada y aproximadamente el problema de *minimización de riesgo empírico*:

$$\min_{\theta \in \mathcal{X}} \mathcal{L}\left(\theta, S\right),$$

donde $S = (z_1, \ldots, z_n)$ es una base con n datos, $\mathcal{X} \subseteq \mathbb{R}^d$ es un conjunto convexo, $\mathcal{L}(\theta, S) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta, z_i)$ y $\ell(\cdot, z_i)$ es una *función de pérdida* que es L-Lipschitz para cada $i \in [n]$.

En el artículo recién citado muestran que si θ^{priv} es muestreado en $\mathcal{X}$ mediante una distribución proporcional a $\exp\left(-\frac{\varepsilon}{2L\|\mathcal{X}\|_2}\mathcal{L}(\theta, S)\right)$, donde $\|\mathcal{X}\|_2$ es el diámetro (Euclidiano) del conjunto $\mathcal{X}$, y si $\theta^* \in \text{argmín}_{\theta\in\mathcal{X}}\mathcal{L}(\theta, S)$, entonces θ^{priv} es ε-DP y, además,

$$\mathbb{E}\left[\mathcal{L}(\theta^{\text{priv}}, S) - \mathcal{L}(\theta^*, S)\right] = O\left(\frac{L\left\|\mathcal{X}\right\|_2 d}{\varepsilon n}\right),$$

donde la esperanza es tomada con respecto a la aleatoriedad del algoritmo.

Definición 2.1.7 (Mecanismo exponencial)**.** Dado algún rango $\mathcal{R}$, una medida μ en $\mathcal{R}$ y una función de puntuación (score/utility function) $u : \mathcal{Z}^n \times \mathcal{R} \to \mathbb{R}$, definimos el *mecanismo exponencial* $\mathcal{M}_E\left(S, u, \mathcal{R}\right)$ como el mecanismo que muestrea $r \in \mathcal{R}$ de acuerdo a la densidad $\propto \exp\left(\frac{\varepsilon u(S,x)}{2\Delta(u)}\right) d\mu(x)$, donde $\Delta(u) := \sup_{r\in\mathcal{R}} \sup_{S\simeq S'} |u(S, r) - u(S', r)|$.

Observación 2.1.2. Para que el mecanismo exponencial esté bien definido, necesitamos que $\int_{\mathcal{R}} \exp\left(\frac{\varepsilon u(S,x)}{2\Delta u}\right) d\mu(x) < \infty$. Cosa que ocurre si $\mathcal{R}$ es finito o si es un compacto, u es continua y μ es medida de Radon.

Teorema 2.1.6 ([10])**.** El mecanismo exponencial es ε-DP.

2.1.4. Privacidad diferencial aproximada

En algunas ocasiones es conveniente relajar la noción de ε-DP, pues esta es una comparación entre «absolutos». Supongamos que tenemos un mecanismo $\mathcal{M}$ que falla en ser ε-DP y asumamos $\varepsilon = 1$. ¿Qué significa esto? Que existen $S, S' \in \mathcal{Z}^n$ y un evento $\mathcal{O}$ tal que

$$\frac{\mathbb{P}\left[\mathcal{M}(S) \in \mathcal{O}\right]}{\mathbb{P}\left[\mathcal{M}(S') \in \mathcal{O}\right]} > \exp\left(1\right) = e.$$

Supongamos que este cociente es grande. Por ejemplo, supongamos que

$$\frac{\mathbb{P}\left[\mathcal{M}(S) \in \mathcal{O}\right]}{\mathbb{P}\left[\mathcal{M}(S') \in \mathcal{O}\right]} = \exp\left(15\right).$$

Si el resto de cocientes está acotado por $\exp(15)$, entonces, por definición, el mecanismo sería 15-DP. Sin embargo, este nivel de privacidad, en la práctica, puede[3] ser inútil. No obstante, hay una porción de aquellos mecanismos que vale la pena rescatar, aquellos donde el cociente puede ser grande, pero con una muy baja probabilidad. Estos son los mecanismos que captura la privacidad diferencial aproximada.

Definición 2.1.8 (Privacidad diferencial aproximada)**.** Sean $\varepsilon > 0$ y $\delta \geq 0$. Decimos que un algoritmo aleatorizado, $\mathcal{M} : \mathcal{Z}^n \to \mathcal{R}$, es (ε, δ)-DP si para todo $S \simeq S' \in \mathcal{Z}^n$ y todo evento $\mathcal{O} \subseteq \mathcal{R}$ se tiene que:

$$\mathbb{P}[\mathcal{M}(S) \in \mathcal{O}] \leq \exp(\varepsilon) \cdot \mathbb{P}[\mathcal{M}(S') \in \mathcal{O}] + \delta \tag{2.3}$$

Observación 2.1.3. Notar que un mecanismo ε-DP es $(\varepsilon, 0)$-DP.

Observación 2.1.4. Para no generar violaciones catastróficas en la privacidad, se pide $\delta \ll 1/n$, donde n es la cantidad de datos. Una elección usual es tomar $1/\delta$ como una función superpolinomial en n.

Cómo probar que un algoritmo es (ε, δ)-DP

Probar que un algoritmo, $\mathcal{M}$, es (ε, δ)-DP es más complicado que probar que es ε-DP. De la misma definición es evidente que ya no basta verificar que el cociente de la funciones de probabilidad –en el caso discreto– o de las funciones de densidad –en el caso absolutamente continuo– esté acotado por $\exp(\varepsilon)$; debemos abordarlo de una manera distinta.

Con fines ilustrativos, supongamos que $\mathcal{M} : \mathcal{Z}^n \to \mathcal{R}$ tiene una distribución discreta.

Sea $S \in \mathcal{Z}^n$ y definamos $p_S(x) = \mathbb{P}[\mathcal{M}(S) = x]$. Probar que $\mathcal{M}$ es ε-DP es equivalente, como ya mencionamos, a probar que $p_S(x)/p_{S'}(x) \leq \exp(\varepsilon)$ para todo $S \simeq S'$ y todo $x \in \mathcal{R}$. Tomando logaritmos, tenemos que esto es equivalente a probar que

$$\log\left(\frac{p_S(x)}{p_{S'}(x)}\right) \leq \varepsilon \qquad \forall S \simeq S' \quad \forall x \in \mathcal{R}.$$

Usando esta «versión» de ε-DP, podemos formular en mayor concordancia la demostración de (ε, δ)-DP. Puede probarse (ver por ejemplo la sección 3.5.1 de [10]) que $\mathcal{M}$ es (ε, δ)-DP si y solo si

$$\log\left(\frac{p_S(x)}{p_{S'}(x)}\right) \leq \varepsilon \text{ con probabilidad (al menos) } 1-\delta \quad \forall S \simeq S' \quad \forall x \in \mathcal{R}.$$

Formalicemos esto para incluir también el caso absolutamente continuo.

Definición 2.1.9. Sea $\mathcal{M} : \mathcal{Z}^n \to \mathcal{R}$ un algoritmo aleatorizado y sean $S, S' \in \mathcal{Z}^n$. Definimos la *variable aleatoria de pérdida de privacidad* como

- $\mathcal{L}_{S,S'} = \log\left(\frac{\mathbb{P}[\mathcal{M}(S)=t]}{\mathbb{P}[\mathcal{M}(S')=t]}\right)$, $t \sim \mathcal{M}(S)$, si $\mathcal{M}$ tiene distribución discreta.

[3] Recalquemos el *puede*, pues como veremos en el desarrollo de esta subsección, para casos en que la discrepancia de los cocientes ocurre con baja frecuencia (por no decir casi nunca), el mecanismo sí es útil.

- $\mathcal{L}_{S,S'} = \log\left(\frac{d\mathbb{P}_{\mathcal{M}(S)}(t)}{d\mathbb{P}_{\mathcal{M}(S')}(t)}\right)$, $t \sim \mathcal{M}(S)$, si $\mathcal{M}$ tiene distribución continua (o sea, es absolutamente continua con respecto a una medida λ).

En caso de que los soportes de $\mathcal{M}(S)$ y $\mathcal{M}(S')$ no coincidan, definimos $\mathcal{L}_{S,S'}$ en ese punto como $+\infty$ o $-\infty$, según corresponda.

Teorema 2.1.7 ([10])**.** Sea $\mathcal{M} : \mathcal{Z}^n \to \mathcal{R}$ un algoritmo aleatorizado. Luego

(a) $\mathcal{M}$ es ε-DP si y solo si $\mathcal{L}_{S,S'} \leq \varepsilon$ casi seguramente para todo $S \simeq S'$.

(b) $\mathcal{M}$ es (ε, δ)-DP si y solo si $\mathbb{P}_{\mathcal{M}(S)}\left[\mathcal{L}_{S,S'} > \varepsilon\right] \leq \delta$ para todo $S \simeq S'$.

El mecanismo Gaussiano

El mecanismo Gaussiano es un mecanismo similar al Laplaciano, donde a un vector armado a partir de los datos, $f(S) \in \mathbb{R}^k$, le sumamos ruido, solo que esta vez ruido Gaussiano. A diferencia del mecanismo Laplaciano, este nuevo mecanismo obtiene (ε, δ)-DP en vez de ε-DP. Por completitud, antes de introducir formalmente el mecanismo, recordemos la distribución Gaussiana.

Definición 2.1.10 (Distribución Gaussiana)**.** La distribución Gaussiana (o normal) centrada en 0 y con varianza σ^2 es la distribución cuya densidad es

$$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right).$$

Cuando una variable aleatoria, Y, tiene esta distribución, escribimos $Y \sim \mathcal{N}(0, \sigma^2)$.

Definición 2.1.11 (Mecanismo Gaussiano)**.** Sean $\varepsilon, \delta \in (0, 1)$. Dada cualquier $f : \mathcal{Z}^n \to \mathbb{R}^k$ dependiente de los datos, definimos el mecanismo Gaussiano como

$$\mathcal{M}_G(S, f, \varepsilon, \delta) = f(S) + (Y_1, \ldots, Y_k),$$

donde los Y_i son i.i.d $\sim \mathcal{N}(0, \sigma^2)$, con $\sigma = 2\log(1.25/\delta)\Delta_2 f/\varepsilon$.

Observación 2.1.5. Notar que, a diferencia del mecanismo Laplaciano, en el mecanismo Gaussiano usamos Δ_2 en vez de Δ_1.

Teorema 2.1.8 ([10])**.** El mecanismo Gaussiano es (ε, δ)-DP.

Una pregunta que puede surgir tras la introducción del mecanismo Gaussiano es ¿por qué usar un mecanismo que nos garantiza una noción de privacidad más débil ((ε, δ)-DP), cuando el mecanismo Laplaciano nos garantiza una más fuerte (ε-DP)? La respuesta a esto viene dada por las distintas sensibilidades que se usan. Recordemos que las normas $\|\cdot\|_1$ y $\|\cdot\|_2$ son equivalentes con constante $1/\sqrt{k}$ y 1; con esto queremos decir que, para todo $x \in \mathbb{R}^k$,

$$\frac{1}{\sqrt{k}} \|x\|_1 \leq \|x\|_2 \leq \|x\|_1.$$

Estas constantes son, de hecho, ajustadas. Así que una función $f : \mathcal{Z}^n \to \mathbb{R}^k$ podría tener $\Delta_2 f = 1$ y $\Delta_1 f = \sqrt{k}$. Si la dimensión k es grande, la adición de ruido vía el mecanismo Laplaciano puede ser sustancial.

2.1.5. Teoremas de composición

Pasemos a ver cómo afecta a la privacidad cuando combinamos (o componemos) diferentes mecanismos diferencialmente privados. La intuición nos dice que al ir combinando información privada, deberíamos obtener algo menos privado. Esto queda confirmado si es que calculamos repetidamente, por ejemplo, $\mathcal{M}_L(S, f, \varepsilon)$. La ley de los grandes números nos dice que el promedio de todos estos resultados convergerá *casi seguramente* al valor real, $f(S)$; o sea, una cálculo repetido de un mismo estadístico «disminuye» la privacidad de este.
Los teoremas de composición abordan este fenómeno cuantificando la degradación de los parámetros de privacidad al componer varios mecanismos privados.

Teorema 2.1.9 (Composición simple [10])**.** Sean $\mathcal{M}_1, \mathcal{M}_2 : \mathcal{Z}^n \to \mathcal{R}$ dos mecanismos $(\varepsilon_1, \delta_1)$-DP y $(\varepsilon_2, \delta_2)$-DP, respectivamente. Luego $(\mathcal{M}_1, \mathcal{M}_2)$ es $(\varepsilon_1 + \varepsilon_2, \delta_1 + \delta_2)$-DP.

Teorema 2.1.10 (Composición avanzada [10])**.** Sean $\varepsilon, \delta' > 0$ y $\delta \geq 0$. La clase de mecanismos (ε, δ)-DP satisface $(\varepsilon', k\delta + \delta')$ bajo composición *k-fold* adaptativa, donde

$$\varepsilon' = \sqrt{2k \log(1/\delta')}\varepsilon + k\varepsilon(e^{\varepsilon} - 1).$$

Observación 2.1.6. Por composición k-fold adaptativa queremos decir que al i-ésimo mecanismo en la composición le es permitido usar la información obtenida en los $\mathcal{M}_1, \ldots, \mathcal{M}_{i-1}$ mecanismos previos. O sea, puede usar el resultado de estos $i - 1$ mecanismos previos en su input.

Observación 2.1.7. Notar que en la composición simple el factor ε es amplificado por k, mientras que en composición avanzada este mismo factor solo es amplificado por $\tilde{O}(\sqrt{k})$.

Corolario 2.1.1 ([10])**.** Sean $\varepsilon \in (0, 1)$ y $\delta' > 0$ parámetros de privacidad. Para asegurar $(\varepsilon, k\delta + \delta')$-DP en la composición de k mecanismos, es suficiente que cada uno de ellos sea $\left(\frac{\varepsilon}{2\sqrt{2k \log(1/\delta')}}, \delta\right)$-DP.

2.2. Concentración de la medida

El siguiente preliminar tiene como objetivo hacer un rápido repaso de las herramientas del Γ-cálculo, donde Γ es el *Carré du Champ* (ver Definición 2.2.1). Estas, junto a la presencia de la desigualdad de curvatura, nos llevarán a la desigualdad de log-Sobolev y, en última instancia, a desigualdades de concentración sub-gaussiana necesarias para el análisis de privacidad que nos proponemos.

2.2.1. Definiciones

A continuación presentamos la definición pertinente para trabajar sobre variedades suaves ($\mathcal{C}^\infty$) y compactas. Los resultados no necesitan de la estructura de variedad y pueden enunciarse con mucha mayor generalidad, pero para propósito de esta tesis esto no será necesario.
Por completitud, presentamos la –algo extensa– siguiente definición.

Definición 2.2.1 (Tripleta Markoviana completa en una variedad compacta)**.** Dada una variedad Riemanniana, compacta y $\mathcal{C}^\infty$, M, definimos la *tripleta Markoviana completa en una variedad compacta* –desde ahora, simplemente tripleta Markoviana– como una tripleta (M, μ, Γ) compuesta por un espacio de medida $(M, \mathcal{F}, \mu)$, la clase $\mathcal{C}^\infty(M)$ de funciones suaves (funciones con derivadas de todo orden) sobre la variedad y un operador $\Gamma : \mathcal{C}^\infty(M) \times \mathcal{C}^\infty(M) \to \mathcal{C}^\infty(M)$ bilineal y simétrico, conocido como el *operador de Carré du Champ*, el cual típicamente será $\Gamma(f, g) = \nabla f \cdot \nabla g$, donde el producto interno es el producto interno de la variedad y ∇ es el gradiente Riemanniano. Esta tripleta debe satisfacer:

1. La medida μ es σ-finita.
2. $\Gamma(f) := \Gamma(f, f) \geq 0$ para toda $f \in \mathcal{C}^\infty(M)$.
3. Para toda $f \in \mathcal{C}^\infty(M)$, existe una constante $C(f)$ (C depende de f) tal que para todo $g \in \mathcal{C}^\infty(M)$,
$$\left| \int_M \Gamma(f, g) d\mu \right| \leq C(f) \|g\|_{L^2(\mu)}.$$
4. Si $f \in \mathcal{C}^\infty(M)$ y $\Gamma(f) = 0$, entonces f es constante.
5. L es un operador lineal en $\mathcal{C}^\infty(M)$ definido mediante la fórmula de integración por partes; o sea, para todo $f, g \in \mathcal{C}^\infty(M)$:
$$\int_M gLf d\mu = -\int_M \Gamma(f, g) d\mu.$$
Un ejemplo de operador que satisface las propiedades de L (para $\Gamma(f, g) = \nabla f \cdot \nabla g$) es el operador de Laplace-Beltrami, Δ_M.
6. Γ y L satisfacen la siguiente fórmula:
$$\Gamma(f, g) = \frac{1}{2} \left[L(fg) - fL(g) - gL(f) \right].$$
7. Para todo $f \in \mathcal{C}^\infty(M)$, $\int_M Lf d\mu = 0$. Claramente usando esta propiedad y la anterior, obtenemos la Propiedad 5.[4]
8. Para toda función suave $\Psi : \mathbb{R}^k \to \mathbb{R}$ y cualesquiera funciones $f_1, \ldots, f_k, g \in \mathcal{C}^\infty(M)$, se tiene que
$$\Gamma\left(\Psi(f_1, \ldots, f_k), g\right) = \sum_{i=1}^k \partial_i \Psi(f_1, \ldots, f_k) \Gamma(f_i, g).$$
y que
$$L\left(\Psi(f_1, \ldots, f_k)\right) = \sum_{i=1}^k \partial_i \Psi(f_1, \ldots, f_k) Lf_i + \sum_{i,j=1}^k \partial^2_{i,j} \Psi(f_1, \ldots, f_k) \Gamma(f_i, f_j).$$

[4]Esta redundancia se debe a que tenemos dos maneras de comenzar: Con un carré du Champ, Γ, obtener L mediante la Propiedad 5 y luego verificar el cumplimiento de las otras propiedades, o mediante un operador L que satisface 7, definir Γ mediante 6 y verificar el resto de las propiedades.

9. $L(\mathcal{C}^\infty(M)) \subseteq \mathcal{C}^\infty(M)$.

10. El semigrupo simétrico cuyo generador infinitesimal es L, $\mathbf{P} = (P_t)_{t\geq 0}$, y que está definido en $\mathcal{C}^\infty(M)$, es Markoviano; esto es, si $f \geq 0$, entonces

$$P_t f \geq 0 \qquad \forall t \geq 0$$

y

$$P_t(1) = 1 \qquad \forall t \geq 0.$$

11. Para toda $f \in \mathcal{C}^\infty(M)$ se tiene que $P_t f \in \mathcal{C}^\infty(M)$ para todo $t \geq 0$.

Lema 2.2.1.

(a) Sea $\Delta_{\mathbb{S}^{d-1}}$ el operador de Laplace-Beltrami en la esfera. Luego

$$\Delta_{\mathbb{S}^{d-1}}(fg) - f\Delta_{\mathbb{S}^{d-1}}(g) - g\Delta_{\mathbb{S}^{d-1}}(f) = 2\nabla f \cdot \nabla g \tag{2.4}$$

(b) Sea $L = \Delta_{\mathbb{S}^{d-1}} - \nabla W \cdot \nabla$ y sea

$$\Gamma(f,g) = \frac{1}{2}\left[L(fg) - fL(g) - gL(f)\right].$$

Luego $\Gamma(f,g) = \nabla f \cdot \nabla g$.

Demostración. La parte *(a)* no es más que una instanciación de la Propiedad A.2.2, parte (a).

Parte (b):

$$\begin{aligned} L(fg) - fL(g) - gL(f) &= (\Delta_{\mathbb{S}^{d-1}}(fg) - f\Delta_{\mathbb{S}^{d-1}}(g) - g\Delta_{\mathbb{S}^{d-1}}(f)) \\ &\quad - (\nabla W \cdot \nabla(fg) - f\nabla W \cdot \nabla g - g\nabla W \cdot \nabla f) \\ &= \Delta_{\mathbb{S}^{d-1}}(fg) - f\Delta_{\mathbb{S}^{d-1}}(g) - g\Delta_{\mathbb{S}^{d-1}}(f) \\ &= 2\nabla f \cdot \nabla g, \end{aligned}$$

donde el segundo paréntesis de la primera igualdad se cancela por regla de la cadena (Propiedad A.2.2), y la tercera igualdad se obtiene de (2.4). ■

Teorema 2.2.1. $\left(\mathbb{S}^{d-1}, \dfrac{e^{-W(x)}dx}{\int_{\mathbb{S}^{d-1}} e^{-W(z)}dz}, \Gamma\right)$, con dx la medida de probabilidad uniforme en la esfera, $\Gamma(f,g) = \nabla f \cdot \nabla g$ y $W \in \mathcal{C}^\infty(M)$, es una tripleta Markoviana completa.

Demostración. 1. Se sigue del hecho de que dx es la medida de probabilidad uniforme en la esfera y W es continua sobre un compacto.

2. y 4. se siguen directamente del hecho de que $\Gamma(f) = |\nabla f|^2$.

6. Sea $L = \Delta_{\mathbb{S}^{d-1}} - \nabla W \cdot \nabla$. La propiedad se sigue por Lema 2.2.1.

7. Dado $f \in \mathcal{C}^\infty(M)$, debemos probar que

$$\int_{\mathbb{S}^{d-1}} (\Delta_{\mathbb{S}^{d-1}}(f) - \nabla W \cdot \nabla f)\, e^{-W} dx = 0.$$

Al aplicar la Propiedad A.2.2 *(a)* y *(b)*, obtenemos

$$\begin{aligned}\Delta_{\mathbb{S}^{d-1}}(fe^{-W}) &= f\Delta_{\mathbb{S}^{d-1}}(e^{-W}) + 2\nabla f \cdot \nabla(e^{-W}) + e^{-W}\Delta_{\mathbb{S}^{d-1}}(f) \\ &= f\Delta_{\mathbb{S}^{d-1}}(e^{-W}) - 2(\nabla f \cdot \nabla W)e^{-W} + e^{-W}\Delta_{\mathbb{S}^{d-1}}(f) \\ &= (\Delta_{\mathbb{S}^{d-1}}(f) - \nabla W \cdot \nabla f)\, e^{-W} + f\Delta_{\mathbb{S}^{d-1}}(e^{-W}) - (\nabla f \cdot \nabla W)\, e^{-W}. \qquad (2.5)\end{aligned}$$

Luego al integrar

$$\begin{aligned}0 &= \int_{\mathbb{S}^{d-1}} \Delta_{\mathbb{S}^{d-1}}(fe^{-W})dx \\ &= \int_{\mathbb{S}^{d-1}} \left[(\Delta_{\mathbb{S}^{d-1}}(f) - \nabla W \cdot \nabla f)\, e^{-W} + f\Delta_{\mathbb{S}^{d-1}}(e^{-W}) - (\nabla f \cdot \nabla W)\, e^{-W}\right] dx \\ &= \int_{\mathbb{S}^{d-1}} (\Delta_{\mathbb{S}^{d-1}}(f) - \nabla W \cdot \nabla f)\, e^{-W} dx + \int_{\mathbb{S}^{d-1}} \left[f\Delta_{\mathbb{S}^{d-1}}(e^{-W}) - (\nabla f \cdot \nabla W)\, e^{-W}\right] dx \\ &= \int_{\mathbb{S}^{d-1}} (\Delta_{\mathbb{S}^{d-1}}(f) - \nabla W \cdot \nabla f)\, e^{-W} dx - \int_{\mathbb{S}^{d-1}} \nabla f \cdot \nabla(e^{-W})dx - \int_{\mathbb{S}^{d-1}} (\nabla f \cdot \nabla W)\, e^{-W} dx \\ &= \int_{\mathbb{S}^{d-1}} (\Delta_{\mathbb{S}^{d-1}}(f) - \nabla W \cdot \nabla f)\, e^{-W} dx + \int_{\mathbb{S}^{d-1}} (\nabla f \cdot \nabla W)\, e^{-W} dx - \int_{\mathbb{S}^{d-1}} (\nabla f \cdot \nabla W)\, e^{-W} dx \\ &= \int_{\mathbb{S}^{d-1}} (\Delta_{\mathbb{S}^{d-1}}(f) - \nabla W \cdot \nabla f)\, e^{-W} dx,\end{aligned}$$

donde la primera igualdad se sigue de la Propiedad A.2.2 *(c)*, la segunda de (2.5), la cuarta de integración por partes (Propiedad A.2.2 *(d)*) y la quinta de Propiedad A.2.2 *(b)*.

5. Se sigue por 6 y 7.

3. Debemos probar que, dados $f, g \in \mathcal{C}^\infty(M)$, existe una constante $C(f)$ que solo depende de f tal que

$$\int_{\mathbb{S}^{d-1}} \Gamma(f,g)e^{-W} dx \leq C(f)\, \|g\|_{L^2(e^{-W}dx)}.$$

Para esto notar que:

$$\begin{aligned}\int_{\mathbb{S}^{d-1}} \Gamma(f,g)e^{-W} dx &= -\int_{\mathbb{S}^{d-1}} g\, (\Delta_{\mathbb{S}^{d-1}}(f) - \nabla W \cdot \nabla f)\, e^{-W} dx \\ &= \int_{\mathbb{S}^{d-1}} g\, (-\Delta_{\mathbb{S}^{d-1}}(f) + \nabla W \cdot \nabla f)\, e^{-W} dx \\ &\leq \|g\|_{L^2(e^{-W}dx)} \cdot \|(-\Delta_{\mathbb{S}^{d-1}}(f) + \nabla W \cdot \nabla f)\|_{L^2(e^{-W}dx)} \\ &= C(f)\, \|g\|_{L^2(e^{-W}dx)},\end{aligned}$$

donde la primera igualdad se debe a integración por partes (Propiedad 5) y la tercera se obtiene mediante una aplicación de Cauchy-Schwarz.

8. Se sigue por Propiedad A.2.2 y regla de la cadena.

9. Evidente.

10. y 11: El semigrupo cuyo generador infinitesimal es $\Delta_{\mathbb{S}^{d-1}}$ es conocido como el semigrupo Browniano esférico, el cual satisface las Propiedades 10 y 11. La medida perturbada $e^{-W}dx$ modifica este semigrupo de acuerdo a las reglas del capítulo 1.15 de Bakry et al. [2]. Este semigrupo modificado continúa conservando las Propiedades 10 y 11.

■

Definición 2.2.2 (Operador Γ_2)**.** El operador $\Gamma_2 : \mathcal{C}^\infty(M) \times \mathcal{C}^\infty(M) \to \mathcal{C}^\infty(M)$ es un operador bilineal definido mediante la fórmula

$$\Gamma_2(f,g) = \frac{1}{2}\left[L\Gamma(f,g) - \Gamma(f,Lg) - \Gamma(Lf,g)\right].$$

De manera análoga a los casos anteriores, $\Gamma_2(f) := \Gamma_2(f,f)$.

Debido a que la acción del Hessiano Riemanniano sobre gradientes de funciones suaves siempre puede escribirse como

$$\begin{aligned}\mathrm{Hess}(f)(\nabla g, \nabla h) = \frac{1}{2}\left[\nabla g \cdot (\nabla f \cdot \nabla h) + \nabla h \cdot (\nabla f \cdot \nabla g)\right.\\ \left. -\nabla f \cdot (\nabla g \cdot \nabla h)\right],\end{aligned} \tag{2.6}$$

definimos el siguiente «Hessiano» abstracto:

Definición 2.2.3 (Hessiano)**.** El hessiano, H, es un operador trilineal $H : \mathcal{C}^\infty(M) \times \mathcal{C}^\infty(M) \times \mathcal{C}^\infty(M) \to \mathcal{C}^\infty(M)$ definido por la fórmula

$$H(f)(g,h) = \frac{1}{2}\left[\Gamma\left(g, \Gamma(f,h)\right) + \Gamma\left(h, \Gamma(f,g)\right) - \Gamma\left(f, \Gamma(g,h)\right)\right].$$

Observación 2.2.1. Claramente si $\Gamma(f,g) = \nabla f \cdot \nabla g$, entonces $\mathrm{Hess}(f)(\nabla g, \nabla h) = H(f)(g,h)$.

Definición 2.2.4 (Condición de curvatura-dimensión)**.** Una tripleta Markoviana, (M,μ,Γ), satisface la condición de curvatura-dimensión $CD(\rho,n)$, donde $\rho \in \mathbb{R}$ y $n \in [1,\infty]$ si para toda $f \in \mathcal{C}^\infty(M)$:

$$\Gamma_2(f) \geq \rho\Gamma(f) + \frac{1}{n}(Lf)^2.$$

Si (M,μ,Γ) satisface $CD(\rho,\infty)$, decimos que satisface la condición de curvatura.

Para el planteamiento de algunas desigualdades de concentración, necesitaremos definir «la constante de Lipschitz» de una función f.

Definición 2.2.5. Dada $f \in \mathcal{C}^\infty(M)$, definimos la seminorma $\|\cdot\|_{\mathrm{Lip}}$ como

$$\|f\|_{\mathrm{Lip}} := \|\Gamma(f)\|_\infty^{1/2}.$$

Observación 2.2.2. En el contexto de concentración de la medida se suele llamar Lipschitz a las funciones f que satisfacen $\|f\|_{\mathrm{Lip}} < \infty$. Nosotros, sin embargo, no suscribiremos a esta convención. Así que no se debe confundir $\|f\|_{\mathrm{Lip}}$ con la constante de Lipschitz de una función f (ver definición (2.9)).

2.2.2. La desigualdad de log-Sobolev

La desigualdad de log-Sobolev (LSI por su sigla en inglés) nos servirá como punto de apoyo para establecer la desigualdad de concentración que probará la privacidad del mecanismo. En esta sección siempre asumiremos que μ es una medida de probabilidad.

A continuación presentaremos la definición de LSI, cómo obtenerla mediante $CD(\rho, n)$ y cómo obtener una desigualdad de concentración a partir de ella.

Comenzaremos con la definición de entropía que, para nuestros fines, ayudará a abreviar algunas expresiones.[5]

Definición 2.2.6 (Entropía)**.** Sea $(M, \mathcal{F}, \mu)$ un espacio de medida y sea $f : M \to \mathbb{R}$ una función integrable, positiva y tal que $\int_M f\,|\ln(f)|\,d\mu < \infty$. Definimos la entropía de f con respecto a μ como:

$$\mathrm{Ent}_\mu(f) = \int_M f\ln(f)d\mu - \left(\int_M f d\mu\right)\ln\left(\int_M f d\mu\right),$$

con la convención $0\ln(0) = 0$.

Observación 2.2.3. En virtud de la desigualdad de Jensen, junto al hecho de que $\phi(x) = x\ln x$ es estrictamente convexa, se tiene, al ser μ medida de probabilidad, que $\mathrm{Ent}_\mu(f) \geq 0$, con igualdad si y solo si f es constante.

Definición 2.2.7 (Desigualdad de log-Sobolev)**.** Una tripleta Markoviana completa (M, μ, Γ) satisface la desigualdad de log-Sobolev si existe una constante $C > 0$ tal que para toda función $f \in \mathcal{C}^\infty(M)$, se tiene que

$$\mathrm{Ent}_\mu\left(f^2\right) \leq 2C\int_M \Gamma(f)d\mu.$$

En tal caso, diremos que (M, μ, Γ) satisface $LS(C)$.

La desigualdad de log-Sobolev para una medida ν puede obtenerse mediante comparación con una medida μ de la que ya conocemos satisface $LS(C)$. Esto resulta de particular interés cuando ν es una «perturbación» de μ. El siguiente teorema, conocido como principio de perturbación de Holley-Stroock, cuantifica el cambio de la constante de log-Sobolev de la medida perturbada frente a la de la medida original.

Teorema 2.2.2 (Holley-Stroock [2])**.** Sea (M, μ, Γ) una tripleta Markoviana completa que satisface $LS(C)$. Si $\nu = e^{-W}d\mu$, entonces (M, ν, Γ) satisface $LS\left(e^{\sup W(x) - \inf W(x)}C\right)$.

En el siguiente capítulo discutiremos brevemente por qué en nuestro caso este resultado no es suficiente y cuán mejor puede llegar a ser nuestra constate de log-Sobolev cuando usamos la condición de curvatura.

[5] Para ver su relación con entropía relativa y la información de Fisher, revisar el sección 5.1.1 de Bakry et al. [2]

La desigualdad de concentración que buscamos se obtiene mediante un método llamado *argumento de Herbst*. Presentamos a continuación este método –enunciado como lema– y mostramos cómo ocuparlo para demostrar la desigualdad deseada.

Lema 2.2.2 (Argumento de Herbst [2])**.** Si (M, μ, Γ) satisface $LS(C)$ para algún $C > 0$, entonces para toda f integrable con $\|f\|_{\text{Lip}} \leq 1$ y todo $s \in \mathbb{R}$, se tiene que

$$\int_M \exp(sf)\, d\mu \leq \exp\left(s \int_M f d\mu + \frac{Cs^2}{2}\right).$$

Teorema 2.2.3 (Concentración [2])**.** Sea (M, μ, Γ) una tripleta Markoviana completa que satisface $LS(C)$ para algún $C > 0$. Luego, para toda función $f \in \mathcal{C}^\infty(M)$ y todo $r > 0$, se tiene que

$$\mu\left(f \geq \int_M f d\mu + r\right) \leq \exp\left(-\frac{r^2}{2C\|f\|^2_{\text{Lip}}}\right).$$

Demostración. Sea $s > 0$. Luego

$$\begin{aligned}
\mu\left(f \geq \mathbb{E}_\mu[f] + r\right) &= \mu\left(e^{sf} \geq e^{s\mathbb{E}_\mu[f] + sr}\right) \\
&\leq \frac{\mathbb{E}_\mu\left[e^{sf}\right]}{e^{s\mathbb{E}_\mu[f] + sr}} \\
&\leq \frac{e^{s\mathbb{E}_\mu[f] + \frac{Cs^2}{2}\|f\|^2_{\text{Lip}}}}{e^{s\mathbb{E}_\mu[f] + sr}} \\
&= e^{-sr + \frac{Cs^2}{2}\|f\|^2_{\text{Lip}}},
\end{aligned}$$

donde la primera desigualdad es la desigualdad de Markov y la segunda es el argumento de Herbst. Optimizando la variable s –cosa que es fácil de hacer, puesto que es una cuadrática– obtenemos que

$$\mu\left(f \geq \mathbb{E}_\mu[f] + r\right) \leq \exp\left(-\frac{r^2}{2C\|f\|^2_{\text{Lip}}}\right).$$

■

La condición de curvatura $CD(\rho, \infty)$ es equivalente –entre otras cosas– a la desigualdad de log-Sobolev *local*, la cual queda formalizada en el siguiente lema.

Lema 2.2.3 (Desigualdad de log-Sobolev local [2])**.** Sea (M, μ, Γ) un triple de Markov completo con semigrupo $\mathbf{P} = (P_t)_{t \geq 0}$. Son equivalentes:

(a) (M, μ, Γ) satisface $CD(\rho, \infty)$ para algún $\rho \in \mathbb{R}$.

(b) Para toda función positiva $f \in \mathcal{C}^\infty(M)$ y para todo $t \geq 0$,

$$P_t(f \ln(f)) - P_t(f) \ln(P_t(f)) \leq \frac{1 - e^{-2\rho t}}{2\rho} P_t\left(\frac{\Gamma(f)}{f}\right).$$

Teorema 2.2.4 (LSI bajo $CD(\rho, \infty)$ [2])**.** Sea (M, μ, Γ) un triple de Markov completo y $\rho > 0$. Si (M, μ, Γ) satisface $CD(\rho, \infty)$, entonces satisface $LS(\frac{1}{\rho})$; es decir,

$$\text{Ent}_\mu\left(f^2\right) \leq \frac{2}{\rho} \int_M \Gamma(f) d\mu \qquad \forall f \in \mathcal{C}^\infty(M).$$

Demostración. El teorema se sigue del Lema 2.2.3 y del hecho (ver, por ejemplo, sección 3.1.9 de Bakry et al. [2]) de que si $f \in \mathcal{C}^\infty(M)$, entonces

$$\lim_{t\to\infty} P_t f = \int_M f d\mu. \tag{2.7}$$

Por Lema 2.2.3, se tiene que

$$P_t\left(f^2\ln(f^2)\right) - P_t(f^2)\ln\left(P_t\left(f^2\right)\right) \leq \frac{1-e^{-2\rho t}}{2\rho} P_t\left(\frac{\Gamma(f^2)}{f^2}\right).$$

Usando la Propiedad 8 de la tripleta Markoviana (definición 2.2.1) con $\Psi(x) = x^2$, se tiene que la expresión anterior es igual a

$$P_t\left(f^2\ln(f^2)\right) - P_t(f^2)\ln\left(P_t\left(f^2\right)\right) \leq \frac{1-e^{-2\rho t}}{2\rho} P_t\left(\frac{4f^2\Gamma(f)}{f^2}\right).$$

Finalmente, aplicando (2.7) a la expresión anterior y notando que $\lim_{t\to\infty} e^{-2\rho t} = 0$, se concluye el resultado.

■

Corolario 2.2.1 ([2])**.** Sea $d\mu = e^{-W}dx$ sobre la variedad Riemanniana $(M, \mathfrak{g})$, donde dx es la medida (de probabilidad) Riemanniana de la variedad. Si $L = \Delta_M - \nabla W \cdot \nabla$ y

$$\mathrm{Ric}(L) := \mathrm{Ric}_{\mathfrak{g}} + \nabla\nabla W \geq \rho\mathfrak{g},$$

donde $\mathrm{Ric}_{\mathfrak{g}}$ es la curvatura de Ricci de la variedad, entonces (M, μ, Γ) satisface $LS\left(\frac{1}{\rho}\right)$.

Demostración. Si $L = \Delta_M - \nabla W \cdot \nabla$, entonces, por Lema 2.2.1, $\Gamma(f,g) = \nabla f \cdot \nabla g$. Usemos esto para describir de manera concreta al operador Γ_2:

$$\begin{aligned}
\Gamma_2(f) &= \frac{1}{2}\left[L\Gamma(f) - 2\Gamma(f, Lf)\right] \\
&= \frac{1}{2}\left[(\Delta_M - \nabla W\cdot\nabla)\left(|\nabla f|^2\right) - 2\nabla f\cdot\nabla\left(\Delta_M f - \nabla W\cdot\nabla f\right)\right] \\
&= \frac{1}{2}\left[\Delta_M\left(|\nabla f|^2\right) - \nabla W\cdot\nabla\left(|\nabla f|^2\right) - 2\nabla f\cdot\nabla(\Delta_M f) + 2\nabla f\cdot\nabla\left(\nabla W\cdot\nabla f\right)\right] \\
&= \frac{1}{2}\Big[2\nabla f\cdot\nabla(\Delta_M f) + 2\left|\nabla\nabla f\right|^2 + 2\mathrm{Ric}_{\mathfrak{g}}(\nabla f, \nabla f) - \nabla W\cdot\nabla\left(|\nabla f|^2\right) \\
&\quad -2\nabla f\cdot\nabla(\Delta_M f) + 2\nabla f\cdot\nabla\left(\nabla W\cdot\nabla f\right)\Big] \\
&= |\nabla\nabla f|^2 + \mathrm{Ric}_{\mathfrak{g}}(\nabla f, \nabla f) + \frac{1}{2}\left[2\nabla f\cdot\nabla\left(\nabla W\cdot\nabla f\right) - \nabla W\cdot\nabla\left(|\nabla f|^2\right)\right] \\
&= |\nabla\nabla f|^2 + \mathrm{Ric}_{\mathfrak{g}}(\nabla f, \nabla f) + \nabla\nabla W(\nabla f, \nabla f),
\end{aligned}$$

donde en el paso de la tercera a la cuarta igualdad usamos el Teorema A.2.1 y de la quinta a la sexta, usamos (2.6).
De la expresión anterior se sigue que si $\mathrm{Ric}(L)(\nabla f, \nabla f) \geq \rho\left|\nabla f\right|^2$, entonces

$$\begin{aligned}
\Gamma_2(f) &\geq \mathrm{Ric}(L)(\nabla f, \nabla f) \\
&\geq \rho\left|\nabla f\right|^2 \\
&= \rho\Gamma(f).
\end{aligned}$$

Lo que implica que (M, μ, Γ) satisface $CD(\rho, \infty)$ y, por ende, $LS(1/\rho)$. ■

Observación 2.2.4. En el caso en que M es un abierto de $\mathbb{R}^d$, el corolario anterior nos dice que $e^{-W}dx$ satisface $LS(1/\rho)$ cuando $W(x)$ es $\mathcal{C}^2$ y ρ-fuertemente convexa.

Corolario 2.2.2 ([2])**.** $\left(\mathbb{S}^{d-1}, \frac{e^{-W(x)}dx}{\int_{\mathbb{S}^{d-1}} e^{-W(z}dz}, \Gamma\right)$, donde $\Gamma(f,g) = \nabla f \cdot \nabla g$ y $W \in \mathcal{C}^\infty(\mathbb{S}^{d-1})$, satisface $\mathrm{LS}\left(\frac{1}{d-2+\lambda}\right)$, donde $\lambda \in \mathbb{R}$ es tal que

$$\nabla\nabla W(x)\left(\nabla f(x), \nabla f(x)\right) \geq \lambda \left|\nabla f(x)\right|^2.$$

Demostración. La demostración es una instanciación del corolario anterior junto con al hecho de que

$$\mathrm{Ric}_{\mathbb{S}^{d-1}}(\nabla f, \nabla f) = (d-2)\left|\nabla f\right|^2.$$

■

Es posible mejorar la constante de log-Sobolev cuando consideramos la condición de curvatura-dimensión, $CD(\rho, n)$, $\rho > 0$, $n > 1$, pero para probarla es necesaria un poco más de maquinaria. A continuación enunciamos –sin demostración– este resultado.

Teorema 2.2.5 (LSI bajo $CD(\rho, n)$ [2])**.** Sea (M, μ, Γ) un triple de Markov completo y sean $\rho > 0$, $n > 1$. Si (M, μ, Γ) satisface $CD(\rho, n)$, entonces también satisface $LS\left(\frac{n-1}{\rho n}\right)$.

2.3. Presentación del problema

Usaremos Sym_d para denotar al espacio vectorial de las matrices simétricas (con entradas reales) de $d \times d$, y Sym_d^+ para las matrices semidefinidas positivas.

Definición 2.3.1 (Norma nuclear y norma de operador)**.** Llamamos *norma nuclear* a la norma

$$\|\cdot\|_{\mathrm{Nuc}} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}_{\geq 0}$$

definida como

$$\begin{aligned}\|A\|_{\mathrm{Nuc}} &= \sum_{i=1}^{\min\{d_1,d_2\}} \sigma_i(A) \\ &= \mathrm{Tr}\left(\sqrt{A^\top A}\right),\end{aligned}$$

donde los $\sigma_i(A)$ denotan los valores singulares de A y $\sqrt{A^\top A}$ denota la raíz cuadrada de la matriz $A^\top A$.[6]

Llamamos, por su parte, *norma de operador* a la norma

$$\|\cdot\|_{\mathrm{op}} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}_{\geq 0}$$

[6] Recordar que $A^\top A$ es una matriz semidefinida positiva y toda matriz de este tipo tiene una (única) raíz cuadrada semidefinida positiva.

definida por

$$\begin{aligned}\|A\|_{\text{op}} &= \max_{i=1,\dots,\min\{d_1,d_2\}} |\sigma_i(A)| \\ &= \max_{x\in\mathbb{S}^{d_2-1}} \|Ax\|_2\end{aligned}$$

Observación 2.3.1. Los valores singulares de una matriz simétrica son los valores absolutos de sus valores propios. En el caso en que la matriz sea además semidefinida positiva, podemos omitir el valor absoluto.

La razón por la que introducimos la norma nuclear junto a la norma de operador es que una es la norma dual de la otra.

Propiedad 2.3.1. $\|\cdot\|_{\text{Nuc}}$ es la norma dual de $\|\cdot\|_{\text{op}}$ y viceversa.

2.3.1. Planteamiento del problema

Estamos interesados en resolver de manera privada el problema estocástico

$$\begin{aligned}\text{mín} \quad & \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[f(X,\mathbf{z})] \\ \text{s.a.} \quad & \|X\|_{\text{Nuc}} \leq 1\end{aligned} \tag{2.8}$$

donde asumimos que $f : \mathbb{R}^{d_1\times d_2} \times \mathcal{Z} \to \mathbb{R}$ es:

- Convexa.
- L_0-Lipschitz con respecto a $\|\cdot\|_{\text{Nuc}}$:

$$|f(X,z) - f(Y,z)| \leq L_0 \|X-Y\|_{\text{Nuc}} \qquad \left(\forall X,Y \in \mathbb{R}^{d_1\times d_2}\right)(\forall z\in\mathcal{Z}). \tag{2.9}$$

- L_1-suave con respecto a $\|\cdot\|_{\text{Nuc}}$:

$$\|\nabla f(X,z) - \nabla f(Y,z)\|_{\text{op}} \leq L_1 \|X-Y\|_{\text{Nuc}} \qquad \left(\forall X,Y \in \mathbb{R}^{d_1\times d_2}\right)(\forall z\in\mathcal{Z}).$$

Además, asumimos que $\mathcal{Z}$ es un espacio de parámetros y $\mathcal{D}$ es una distribución en $\mathcal{Z}$.
Dos propiedades básicas de este escenario son las siguientes.

Teorema 2.3.1. Sea $(\mathcal{X}, \|\cdot\|)$ un espacio normado y $\mathcal{Z}$ un espacio de parámetros con distribución $\mathcal{D}$.

- Si $f : \mathcal{X}\times\mathcal{Z} \to \mathbb{R}$ es L_0-Lipschitz para todo $z\in\mathcal{Z}$, entonces $\|\nabla f(x,z)\|_* \leq L_0$ para todo $x\in\mathcal{X}$ y todo $z\in\mathcal{Z}$, donde $\|\cdot\|_*$ es la norma dual a $\|\cdot\|$.
- Si $f : \mathcal{X}\times\mathcal{Z} \to \mathbb{R}$ es L_1-suave para todo $z\in\mathcal{Z}$, entonces $\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[f(x,\mathbf{z})]$ es L_1-suave.

Observación 2.3.2. En nuestro problema, la Lipschitzianidad de la función f junto al teorema anterior implican que el gradiente de la función es de espectro acotado para todo $X\in\mathbb{R}^{d_1\times d_2}$ y todo $z\in\mathcal{Z}$. En efecto, debido a que la norma dual de $\|\cdot\|_{\text{Nuc}}$ es $\|\cdot\|_{\text{op}}$, tenemos que $\|\nabla f(X,z)\|_{\text{op}} \leq L_0$.

Conjeturamos que el Problema (2.8) puede ser resuelto con una tasa óptima $\tilde{O}\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d}}{\varepsilon n}\right)$ –donde d es la dimensión, n es la cantidad de datos y ε es el parámetro de privacidad– mediante una versión privada del algoritmo de Frank-Wolfe estocástico [21]. Este algoritmo requiere[7] que resolvamos privadamente:

$$\max_{\|Y\|_{\text{Nuc}} \leq 1} \langle -\nabla F_B(X, z), Y \rangle,$$

donde $\langle U, V \rangle = \operatorname{Tr}\left(U^\top V\right)$, B es un lote de datos de S (o sea, un subconjunto) y $\nabla F_B(X, z) = \frac{1}{|B|} \sum_{z \in B} \nabla f(X, z)$.

Observación 2.3.3. Usamos la notación ∇ para $\nabla F_B(X, z)$ debido a que este es un *estimador* del gradiente de $F_{\mathcal{D}}(X) = \mathbb{E}_{\mathbf{z} \sim \mathcal{D}}\left[f(X, \mathbf{z})\right]$, y no debe ser confundido con un gradiente real.

La «Lipschitzianidad» de la función $f(\cdot, z)$ implica que para todo $X \in \mathbb{R}^{d_1 \times d_2}$ y todo $z \in \mathcal{Z}$, se tenga que $\|\nabla f(X, z)\|_{\text{op}} \leq L_0$. Esto, a su vez, implica que

$$\|\nabla F_B(X, z)\|_{\text{op}} = \left\| \frac{1}{|B|} \sum_{b \in B} \nabla f(X, z) \right\|_{\text{op}} \leq \frac{1}{|B|} \cdot |B| \cdot L_0 = L_0.$$

Sin embargo, el hecho de que $\nabla F_B(X, z)$ se construya mediante gradientes de funciones, le da licencia a $\nabla F_B(X, z)$ para «tomar cualquier forma», salvo la restricción de que sus valores propios estén acotados por L_0. Es por esto que nos enfocamos en el caso aparentemente más general –pero notacionalmente más liviano– de resolver aproximadamente y de forma privada

$$\max_{\|Y\|_{\text{Nuc}} \leq 1} \langle A(S), Y \rangle, \tag{2.10}$$

donde $A(S)$ es una matriz construída a partir de los datos, con $\|A(S)\|_{\text{op}} \leq L_0$.

Resolución de problemas de manera privada

En la subsección anterior hicimos alusión a "resolver privadamente un problema"sin dar más detalles al respecto. En este apartado esperamos hacer explícito qué entendemos por aquello, para no dejar lugar a dudas.

Dado un determinado problema –por ejemplo (2.8)– decimos que puede ser *resuelto de manera privada* si existe un algoritmo $\mathcal{M} : \mathcal{Z}^n \to \mathcal{R}$ que entregue una solución[8] a este y, además, que sea diferencialmente privado (el algoritmo $\mathcal{M}$); ya sea en la variante «pura» (Definición 2.1.1) o

[7] Una breve explicación de este requerimiento: El algoritmo de Frank-Wolfe no privado (algoritmo 1) necesita que en cada iteración resolvamos un subproblema lineal. Una manera de privatizar este algoritmo (o sea, de que su *output* sea diferencialmente privado) es privatizar el resultado de este subproblema lineal. Para después obtener privacidad en el resultado final del algoritmo, ensamblamos las distintas partes privadas y apelamos al teorema de composición avanzada (Teorema 2.1.10).

[8] Qué se considera una solución al problema varía, naturalmente, de problema a problema. En el caso de un problema de optimización convexo, pediremos que si x^* es el minimizador real, entonces $f(\mathcal{M}(S)) - f(x^*)$ sea pequeño.

«aproximada» (Definición 2.1.8). Así, en el caso del Problema (2.8), resolveremos privadamente el problema si diseñamos un algoritmo diferencialmente privado, $\mathcal{M}$, tal que

$$F_{\mathcal{D}}(\mathcal{M}(S)) - F_{\mathcal{D}}(X^*) \leq \tau,$$

en alta probabilidad, o bien

$$\mathbb{E}_{S\sim\mathcal{D}^n}\left[F_{\mathcal{D}}(\mathcal{M}(S))\right] - F_{\mathcal{D}}(X^*) \leq \tau,$$

donde X^* es el minimizador real y τ es algún criterio que se espera satisfacer.

2.3.2. Miremos el lado positivo

Siempre es posible plantear el problema para el caso semidefinido positivo sin perder generalidad en el análisis asintótico, que es el que nos interesa. O sea, siempre podemos plantear el Problema (2.8) como

$$\begin{aligned} \text{mín} \quad & \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[f(X,\mathbf{z})] \\ \text{s.a.} \quad & \|X\|_{\text{Nuc}} \leq 1, \end{aligned} \tag{2.11}$$

pero con $X \in \text{Sym}_d^+$, pagando el costo de agregar un factor constante a la tasa de error (constante que es irrelevante en el análisis asintótico). Para esto, primero debemos realizar un argumento de simetrización: En vez de trabajar con la matriz $A \in \mathbb{R}^{d_1\times d_2}$, trabajaremos con su versión «simetrizada»

$$\bar{A} = \begin{pmatrix} 0 & A \\ A^\top & 0 \end{pmatrix},$$

la cual tiene entre sus valores propios a los valores singulares de A.

Propiedad 2.3.2. Sea $A \in \mathbb{R}^{d_1\times d_2}$, sean $\sigma_1, \ldots, \sigma_r$ sus valores singulares no-negativos y sea $\bar{A}$ definida como arriba. Luego los valores propios no-negativos de $\bar{A}$ son σ_i y $-\sigma_i$ para todo $i \in [r]$. Además, los vectores propios asociados a σ_i son de la forma $\begin{pmatrix} u \\ v \end{pmatrix}$ y los vectores propios asociados a $-\sigma_i$ son de la forma $\begin{pmatrix} -u \\ v \end{pmatrix}$, donde u y v son un par de vectores singulares izquierdo y derecho, respectivamente, asociado a σ_i.

Demostración. Sea $A = \sum_{i=1}^r \sigma_i u_i {v_i}^\top$ la descomposición de Schmidt de A. Un cálculo sencillo nos muestra que $\begin{pmatrix} u_i \\ v_i \end{pmatrix}$ y $\begin{pmatrix} -u_i \\ v_i \end{pmatrix}$ son vectores propios de $\bar{A}$ asociados a σ_i y $-\sigma_i$, respectivamente.

Sea ahora $\begin{pmatrix} u \\ v \end{pmatrix}$ vector propio de $\bar{A}$ asociado a $\lambda \neq 0$. Entonces

$$\begin{aligned} Av &= \lambda u \\ A^\top u &= \lambda v. \end{aligned}$$

Lo que implica que

$$A^\top A v = \lambda^2 v.$$

Por ende, $|\lambda|$ es un valor singular no-negativo de A. Además, u y v es un par singular izquierdo-derecho de A asociado a $|\lambda|$. ■

Segundo, «hacemos positiva» a la matriz $\bar{A} \in \mathrm{Sym}_{d_1+d_2}$ usando el hecho de que $\left\|\bar{A}\right\|_{\mathrm{op}}$ es de espectro acotado.

Propiedad 2.3.3. Sea $A \in \mathrm{Sym}_d$ tal que $\|A\|_{\mathrm{op}} \leq L$. Luego $(A + L \cdot I) \in \mathrm{Sym}_d^+$.

Demostración. Sea $x \in \mathbb{S}^{d-1}$. Luego

$$\begin{aligned} x^\top (A + L \cdot I) x &= x^\top A x + L \\ &\geq -L + L \\ &= 0, \end{aligned}$$

donde la desigualdad se debe al hecho de que $\|A\|_{\mathrm{op}} \leq L$. ■

Desde ahora asumiremos que estamos trabajando con matrices semidefinidas positivas.

2.3.3. Reducción del problema

Independiente de la restricción de privacidad, el Problema (2.10) es simplemente un problema de programación semidefinida. Es bien sabido que cuando el conjunto factible es convexo y compacto, siempre se puede encontrar una solución a este problema en alguno de los puntos extremos.

Lema 2.3.1. Sean $A, B \in \mathrm{Sym}_d^+$ y sea $\mathrm{rank}(\cdot)$ la función de rango de una matriz. Luego

$$\mathrm{rank}(A + B) \geq \mathrm{máx}\left\{\mathrm{rank}(A), \mathrm{rank}(B)\right\}.$$

Demostración. La demostración se sigue del hecho de que, por el teorema núcleo-imagen,

$$\mathrm{rank}(Z) = d - \dim \ker(Z) \tag{2.12}$$

Como A y B son semidefinidas positivas, se tiene que

$$x^\top (A + B)\, x = x^\top A x + x^\top B x \geq 0 \qquad \forall x \in \mathbb{S}^{d-1}.$$

Además,

$$x^\top (A + B)\, x = 0 \Leftrightarrow x^\top A x = 0 \text{ y } x^\top B x = 0$$

y

$$x^\top Z x = 0 \Leftrightarrow x \in \ker(Z) \qquad \forall Z \in \mathrm{Sym}_d^+.$$

Por ende, $x \in \ker(A + B)$ si y solo si $x \in \ker(A) \cap \ker(B)$. De esto y de (2.12), se sigue que

$$\mathrm{rank}(A + B) \geq \mathrm{máx}\left\{\mathrm{rank}(A), \mathrm{rank}(B)\right\}.$$

■

Propiedad 2.3.4. Los puntos extremos de la bola nuclear, $\left\{X \in \mathrm{Sym}_d^+ : \|X\|_{\mathrm{Nuc}} \leq 1\right\}$, son las matrices $X \in \mathrm{Sym}_d^+$ de rango 1 con $\|X\|_{\mathrm{Nuc}} = 1$ y la matriz $\mathbf{0_{d\times d}}$. Las matrices, X, de rango 1 y $\|X\|_{\mathrm{Nuc}} = 1$ tienen la forma $X = xx^\top$, con $x \in \mathbb{S}^{d-1}$.

Demostración. Es un hecho conocido que las matrices de rango 1 en el cono Sym_d^+ tienen forma $\lambda xx^\top$, donde $\lambda > 0$ y $x \in \mathbb{S}^{d-1}$. Además,

$$\begin{aligned}\left\|\lambda xx^\top\right\|_{\mathrm{Nuc}} &= \lambda \left\|xx^\top\right\|_{\mathrm{Nuc}} \\ &= \lambda \left|x\right| \\ &= \lambda.\end{aligned}$$

Por lo que las matrices de rango 1 en Sym_d^+ y $\|X\|_{\mathrm{Nuc}} = 1$ tienen la forma $X = xx^\top$.

Es claro que la matriz $\mathbf{0_{d\times d}}$ es un punto extremo, ya que estamos en el cono de matrices semidefinidas positivas. Por lo que resta probar que las matrices del tipo $xx^\top$, con $x \in \mathbb{S}^{d-1}$, conforman el resto de los puntos extremos. Para esto, comencemos descartando que las matrices de rango $r \geq 2$ puedan ser puntos extremos.

Sea $A \in \left\{X \in \mathrm{Sym}_d^+ : \|X\|_{\mathrm{Nuc}} \leq 1\right\}$ de rango $r \geq 2$. Por teorema espectral,

$$A = \sum_{i=1}^{r} \lambda_i x_i {x_i}^\top,$$

donde los $\lambda_i > 0$, $\lambda_1 + \ldots + \lambda_r \leq 1$ y $x_i \in \mathbb{S}^{d-1}$ son los vectores propios de la matriz asociados a valores propios positivos. Luego

$$\begin{aligned}A &= \sum_{i=1}^{r} \lambda_i x_i {x_i}^\top \\ &= \sum_{i=1}^{r} \lambda_i x_i {x_i}^\top + (1 - \lambda_1 - \ldots - \lambda_r)\, \mathbf{0_{d\times d}},\end{aligned}$$

lo cual es una combinación convexa de a lo menos dos términos. Por ende, A no puede ser punto extremo de $\left\{X \in \mathrm{Sym}_d^+ : \|X\|_{\mathrm{Nuc}} \leq 1\right\}$.

Veamos ahora que las matrices de rango 1 de norma unitaria son puntos extremos. Partamos notando que

$$\lambda xx^\top = \lambda xx^\top + (1-\lambda)\mathbf{0_{d\times d}},$$

no puede ser punto extremos si $0 < \lambda < 1$, así que enfoquémonos en las matrices del tipo $xx^\top$, con $x \in \mathbb{S}^{d-1}$. Supongamos que existen matrices $A, B \in \left\{X \in \mathrm{Sym}_d^+ : \|X\|_{\mathrm{Nuc}} \leq 1\right\}$ tales que

$$xx^\top = \frac{A+B}{2}.$$

Por Lema 2.3.1,

$$1 = \mathrm{rank}(xx^\top) \geq \mathrm{máx}\left\{\mathrm{rank}(A), \mathrm{rank}(B)\right\}.$$

Por lo que A y B son de rango 1, o A es de rango 1 y $B = \mathbf{0_{d\times d}}$.

- Si A y B son de rango 1, entonces un cálculo simple nos muestra que $A = B = xx^\top$.
- Si A es de rango 1 y $B = \mathbf{0_{d\times d}}$, entonces $A = 2xx^\top$. Lo que contradice que $\|A\|_{\text{Nuc}} \leq 1$.

Por ende, $xx^\top$ es un punto extremo de $\left\{X \in \text{Sym}_d^+ : \|X\|_{\text{Nuc}} \leq 1\right\}$. ■

Con esto, el Problema (2.10) queda reducido a

$$\begin{aligned}\max_{\|Y\|_{\text{Nuc}}\leq 1} \langle A(S), Y\rangle &= \max_{\|Y\|_{\text{Nuc}}\leq 1} \text{Tr}\left(A(S)Y\right)\\ &= \max_{x\in\mathbb{S}^{d-1}} \text{Tr}\left(A(S)xx^\top\right)\\ &= \max_{x\in\mathbb{S}^{d-1}} x^\top A(S)x \qquad (2.13)\\ &= \|A(S)\|_{\text{op}}\end{aligned}$$

Nos apoyaremos en (2.13) para calcular privadamente (2.10).

2.3.4. ¿El máximo o el mínimo?

Usaremos esta subsección para explicitar algunos hechos ocupados en la implementación de la subrutina de Frank-Wolfe. En esta, como ya se mencionó, debemos minimizar

$$\text{Tr}\left(AX\right) \quad \text{s. a.} \quad X \in \text{Sym}_d^+ \wedge \|X\|_{\text{Nuc}} \leq 1.$$

Por (2.13), ya que el mínimo en un problema lineal también es alcanzado en los puntos extremos, nos basta encontrar una aproximación del vector propio asociado al menor valor propio de la matriz $A(S)$.

Sin embargo, métodos como el de Oja o el de muestreo vía mecanismo exponencial no están diseñados para calcular el menor vector propio, sino que el mayor, el principal. Esto realmente no es ningún problema, pues, como veremos, en el contexto de matrices semidefinidas positivas de espectro acotado, podemos «alterar» la matriz para que el máximo y el mínimo valor propio intercambien posiciones. En este proceso, además, el argmáx y el argmín también cambian de posición:

Propiedad 2.3.5. Sea $A \in \text{Sym}_d^+$ tal que $\|A\|_{\text{op}} \leq L$ y sea λ su mínimo valor propio. Luego

(a) $-A + L\cdot I \in \text{Sym}_d^+$.

(b) $-\lambda + L$ es el máximo valor propio de $-A + L\cdot I$.

(c) $u \in \text{argmáx}\left\{x^\top\left(-A(S) + L\cdot I\right)x : x \in \mathbb{S}^{d-1}\right\}$ si y solo si $u \in \text{argmín}\left\{x^\top A(S)x : x \in \mathbb{S}^{d-1}\right\}$.

Demostración. (a) Sea $x \in \mathbb{S}^{d-1}$. Luego

$$\begin{aligned}x^\top\left(-A + L\cdot I\right)x &= -x^\top Ax + L\left(x^\top Ix\right)\\ &= -x^\top Ax + L\left|x\right|^2\\ &= -x^\top Ax + L\\ &\geq -L + L\\ &= 0,\end{aligned}$$

donde la desigualdad se debe al hecho de que $\|A\|_{\text{op}} \leq L$. Por lo tanto, $(-A + L \cdot I) \in \text{Sym}_d^+$.

(b) Sea $x \in \mathbb{S}^{d-1}$. Luego

$$\begin{aligned} x^\top (-A + L \cdot I)\, x &= -x^\top A x + L \\ &\leq -\lambda + L, \end{aligned}$$

donde la desigualdad se debe a que λ es el mínimo valor propio de A. Además, si x es vector propio de A asociado a λ, entonces

$$(-A + L \cdot I)\, x = (-\lambda + L)\, x.$$

Por lo tanto, $(-\lambda + L)$ es el mayor valor propio de $(-A + L \cdot I)$.

(c) Es evidente a partir de lo anterior. ■

Así que, para propósitos del análisis, basta con que podamos calcular de manera privada y aproximada

$$\text{máx}\, \text{Tr}\,(AX) \quad \text{s. a} \quad X \in \text{Sym}_d^+ \wedge \|X\|_{\text{Nuc}} \leq 1.$$

2.3.5. En qué nos enfocaremos

Como ya fue mencionado, conjeturamos que el Problema (2.11) puede ser resuelto mediante el algoritmo de Frank-Wolfe estocástico. En su variante tradicional, es decir, determinista (ver Jaggi [13]), el algoritmo de Frank-Wolfe consiste en:

Algoritmo 1: Frank-Wolfe

Requiere: Conjunto compacto y convexo $K \subseteq \mathbb{R}^d$, cantidad de pasos T

1 Elegir $x^{(0)} \in K$

2 **for** $t = 0, \ldots, T$ **do**

3 $\quad s := \text{argmín}_{s \in K} \langle s, \nabla f(x^{(t)}) \rangle$

4 $\quad$ Actualizar $x^{(t+1)} := (1 - \eta_t)\, x^{(t)} + \eta_t s$, donde $\eta_t := \frac{2}{t+2}$

Entrega: $x^{(T+1)}$

Como podemos ver, el algoritmo posee una subrutina en la que se debe resolver un problema de minimización de un objetivo lineal sobre el compacto K. En el caso estocástico, el gradiente es reemplazado por un estimador insesgado del gradiente de $F_{\mathcal{D}}(X) = \mathbb{E}_{\mathbf{z} \sim \mathcal{D}}\,[f(X, \mathbf{z})]$. En nuestro caso, este será construido usando los datos y tomará la forma:

$$\begin{aligned} D_0 &= \sum_{i=1}^{n/2} \nabla f\left(X^{(0)}, z_i^0\right) \\ D_t &= (1 - \eta)\left[D_{t-1} + \left(\nabla f\left(X^{(t)}, z_t\right) - \nabla f\left(X^{(t-1)}, z_t\right)\right)\right] + \eta \nabla f\left(X^{(t)}, z_t\right), \end{aligned}$$

donde η es el paso del algoritmo, el cual se define más adelante.
De esta forma, la cantidad de iteraciones que hará el algoritmo estará limitada por la cantidad de

datos que se posean.

Presentamos a continuación un esquema del algoritmo que usaremos:

Algoritmo 2: Esquema de SFW en bola nuclear

Requiere: Conjunto de datos $S = (z_1, \dots, z_n) \in \mathcal{Z}^n$, parámetros de privacidad (ε, δ)

1 Establecer paso $\eta = \frac{\log(n)}{n}$

2 Elegir $X^{(0)} \in \left\{M \in \mathrm{Sym}_d^+ : \|M\|_{\mathrm{Nuc}} \leq 1\right\}$

3 Establecer un lote inicial $S_0 = \left(z_1^0, z_2^0, \dots, z_{n/2}^0\right)$ con $n/2$ datos

4 Calcular $D_0 = \frac{2}{n} \sum_{i=1}^{n/2} \nabla f\left(X^{(0)}, {z_i}^0\right)$

5 Calcular privadamente $V_0 = xx^\top$, con $x \in \mathbb{S}^{d-1}$, de manera que $\mathrm{Tr}\left(-D_0 V_0\right)$ aproxime a $\max_{\|Y\|_{\mathrm{Nuc}} \leq 1} \mathrm{Tr}\left(-D_0 Y\right)$.

6 $X^{(1)} \leftarrow (1-\eta) X^{(0)} + \eta V_0$

7 Establecer $\hat{S} = S \setminus S_0 = \left(z_1, \dots, z_{n/2}\right)$, los $n/2$ datos restantes

8 **for** $t = 1, \dots, n/2$ **do**

9 Calcular $\Delta_t\left(z_t\right) = \nabla f(X^{(t)}, z_t) - \nabla f(X^{(t-1)}, z_t)$

10 $D_t = (1-\eta)\left(D_{t-1} + \Delta_t(z_t)\right) + \eta \nabla f(X^{(t)}, z_t)$

11 Calcular privadamente $V_t = xx^\top$, con $x \in \mathbb{S}^{d-1}$, de manera que $\mathrm{Tr}\left(-D_t V_t\right)$ aproxime a $\max_{\|Y\|_{\mathrm{Nuc}} \leq 1} \mathrm{Tr}\left(-D_t Y\right)$.

12 $X^{(t+1)} \leftarrow (1-\eta) X^{(t)} + \eta V_t$

Entrega: $X^{(n/2+1)}$

La privacidad del algoritmo anterior vendrá dada por los pasos 5 y 11, en conjunto con el teorema de composición avanzada (Teorema 2.1.10). La precisión de este, a su vez, dependerá crucialmente de la precisión obtenida en estos pasos. Para ver esto último, referimos a los cálculos realizados en el Teorema 3.2 de Bassily et al. [4], los cuales, bajo una pequeña modificación en las constantes, nos dicen que

$$F_{\mathcal{D}}(X^{(n/2+1)}) - F_{\mathcal{D}}(X^*) \leq (1-\eta)^{\frac{n}{2}+1}\left(F_{\mathcal{D}}(X^{(0)}) - F_{\mathcal{D}}(X^*)\right) + 4\eta \sum_{t=0}^{n/2}(1-\eta)^t \left\|\nabla F_{\mathcal{D}}(X^{(n/2-t)}) - D_{\frac{n}{2}-t}\right\|_{\mathrm{op}} + 2L_1\eta^2 \sum_{t=0}^{n/2}(1-\eta)^t + \eta \sum_{t=0}^{n/2}(1-\eta)^t \alpha_{\frac{n}{2}-t}, \tag{2.14}$$

donde $X^* \in \operatorname{argmín}_{\|X\|_{\mathrm{Nuc}} \leq 1} F_{\mathcal{D}}(X)$ y α_t es el error en el que se incurre en la aproximación del paso 11 del t-ésimo iterado del algoritmo 2; o sea, $\alpha_t = \mathrm{Tr}\left(D_t V_t\right) - \mín_{\|Y\|_{\mathrm{Nuc}} \leq 1} \mathrm{Tr}\left(D_t Y\right)$.

El error en el que se incurre debido al uso del estimador de gradiente (segundo sumando en el lado derecho de la desigualdad (2.14)) está acotado por el análisis realizado en el Lema 3.4 de Bassily et al. [4], mientras que el error α_t es el objeto estudiado en esta tesis y es pensado con una cota uniforme (del mismo orden de error para todos los iterados).

Enunciamos a continuación dos resultados: el primero indica el error en el que se incurre cuando se usa el estimador de gradiente D_t, mientras que el segundo indica el nivel de privacidad que necesitamos de cada V_t para que el algoritmo 2 sea diferencialmente privado.

Teorema 2.3.2 (Basado en el Lema 3.4 de [4] y el Lema 6.2 de [5])**.** Sea $\mathcal{D}$ una distribución sobre $\mathcal{Z}$. Sea $S \sim \mathcal{Z}^n$ el conjunto de datos que introducimos al algoritmo 2. Luego:

(a) El estimador de gradiente, D_t, definido en el algoritmo 2 satisface

$$\mathbb{E}_{S\sim\mathcal{D}^n}\left[\left\|D_t - \nabla F_{\mathcal{D}}(X^{(t)})\right\|_{\text{op}}\right] \leq 2L_0\sqrt{\frac{2\kappa}{n}}(1-\eta)^t + 4\eta\sqrt{\kappa t}(2L_1 + L_0),$$

donde $\kappa = (2\log(d+2) - 1)\, e$.

(b) Para todo $p \in (0,1)$, con probabilidad al menos $1-p$, se tiene que para todo $t \in \{0,\ldots,n/2\}$

$$\left\|D_t - \nabla F_{\mathcal{D}}(X^{(t)})\right\|_{\text{op}} \leq 2\left[e\sqrt{\kappa} + \sqrt{3\log(n/p)}\right]\left(\frac{\sqrt{2}(1-\eta)^t L_0}{\sqrt{n}} + \eta\sqrt{t}\,(2L_1 + L_0)\right),$$

donde $\kappa = (2\log(d+2) - 1)\, e$.

Lema 2.3.2. Si en el algoritmo 2 cada V_t es $\left(\frac{\varepsilon}{2\sqrt{n\log(1/\delta)}}, (T+1)\delta\right)$-DP, donde $T \geq 0$ es algún número, entonces el algoritmo 2 es $(\varepsilon, n(T+1)\delta)$-DP.

Demostración. La demostración es consecuencia directa del teorema de composición avanzada. ■

Observación 2.3.4. La presencia del término T en el lema anterior es instrumental para el Teorema 3.2.2.

Una consecuencia de la desigualdad (2.14), junto con el Teorema 2.3.2, es el siguiente teorema. Nuevamente, los cálculos son referidos a [4].

Teorema 2.3.3.

(a) Si $\mathbb{E}[\alpha_t] \leq \alpha$ para todo $t \in \{0,\ldots,n/2\}$, entonces $X^{(n/2+1)}$ satisface

$$\begin{aligned}\mathbb{E}\left[F_{\mathcal{D}}(X^{(n/2+1)}) - F_{\mathcal{D}}(X^*)\right] \leq & e^{-\eta(\frac{n}{2}+1)}2L_0 + 4\eta\left(nL_0\sqrt{\frac{2\kappa}{n}}e^{-\eta\frac{n}{2}} + 4\sqrt{\kappa n}\,(2L_1 + L_0)\right) \\ & + 2L_1\eta + \alpha\end{aligned}$$

(b) Si $\alpha_t \leq \alpha$ para todo $t \in \{0,\ldots,n/2\}$, entonces, para todo $p \in (0,1)$, con probabilidad al menos $1-p$ se tiene que $X^{(n/2+1)}$ satisface

$$\begin{aligned}F_{\mathcal{D}}(X^{(n/2+1)}) - F_{\mathcal{D}}(X^*) \leq & e^{-\eta(\frac{n}{2}+1)}2L_0 + 4\eta\left(enL_0\sqrt{\frac{2\kappa}{n}}e^{-\eta\frac{n}{2}} + 2e\sqrt{\kappa n}\,(2L_1 + L_0)\right. \\ & \left. + \sqrt{3\log(n/p)}nL_0\sqrt{\frac{2}{n}}e^{-\eta\frac{n}{2}} + 2\sqrt{3\log(n/p)}\sqrt{n}\,(2L_1 + L_0)\right) + 2L_1\eta + \alpha.\end{aligned}$$

Una consecuencia del teorema anterior es que al tomar $\eta = \frac{\log(n)}{n}$, tal como en el algoritmo 2, se tiene el siguiente resultado:

Corolario 2.3.1. Sea $X^{(n/2+1)}$ el resultado entregado por el algoritmo 2.

(a) Si $\mathbb{E}[\alpha_t] \leq \alpha$ para todo $t \in \{0, \ldots, n/2\}$, entonces

$$\mathbb{E}\left[F_{\mathcal{D}}(X^{(n/2+1)}) - F_{\mathcal{D}}(X^*)\right] \leq \frac{2L_0}{\sqrt{n}} + \frac{4L_0\sqrt{2\kappa}\log(n)}{n} + \frac{16\left(2L_1 + L_0\right)\sqrt{\kappa}\log(n)}{\sqrt{n}} + \frac{2L_1\log(n)}{n} + \alpha$$

(b) Si $\alpha_t \leq \alpha$ para todo $t \in \{0, \ldots, n/2\}$, entonces, para todo $p \in (0,1)$, con probabilidad al menos $1-p$ se tiene que

$$\begin{aligned} F_{\mathcal{D}}(X^{(n/2+1)}) - F_{\mathcal{D}}(X^*) \leq & \frac{2L_0}{\sqrt{n}} + \frac{4eL_0\sqrt{2\kappa}\log(n)}{n} + \frac{8e\left(2L_1 + L_0\right)\sqrt{\kappa}\log(n)}{\sqrt{n}} \\ & + \frac{4L_0\sqrt{6\log(n/p)}\log(n)}{n} + \frac{8\left(2L_1 + L_0\right)\sqrt{3\log(n/p)}\log(n)}{\sqrt{n}} \\ & + \frac{2L_1\log(n)}{n} + \alpha \end{aligned}$$

Así, para obtener la cota deseada, debemos trabajar una cota apropiada para los α_t, ya sea en esperanza o en alta probabilidad.

Observación 2.3.5. El teorema y el corolario anterior no reflejan la posible dependencia de α en términos de η. Esta aparecerá, de hecho, mediante la sensibilidad $\Delta_{\mathrm{OP}}(D_t)$.

Teorema 2.3.4 (Lema 3.3 de [4])**.** $\Delta_{\mathrm{OP}}(D_t) \leq \max\left\{(1-\eta)^t \cdot \frac{2L_0}{n}, 2\eta\left(2L_1 + L_0\right)\right\}$.
En el caso particular en que $\eta = \frac{\log(n)}{n}$,

$$\Delta_{\mathrm{OP}}(D_t) \leq 2\eta\left(2L_1 + L_0\right).$$

Para finalizar esta subsección, repetimos una vez más la temática de la tesis: Queremos resolver privada y aproximadamente el problema de DP-SCO

$$\begin{aligned} \text{mín} \quad & \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[f(X, \mathbf{z})] \\ \text{s.a.} \quad & \|X\|_{\mathrm{Nuc}} \leq 1, \end{aligned}$$

donde $f(\cdot, z)$ es L_0-Lipschitz y L_1-suave, usando una variante privada del algoritmo de Frank-Wolfe estocástico.

Sin pérdida de generalidad, podemos asumir que las matrices en cada uno de los subproblemas son semidefinidas positivas. Además, debido al Teorema 2.3.2 y al corolario 2.3.1, el punto de interés radica en el cálculo privado de V_t y de la precisión α_t que se pueda obtener de cada uno de ellos. Por lo mismo, en el siguiente capítulo indagaremos técnicas que buscan precisamente esto: calcular privadamente cada $V_t = xx^\top$ y estudiar la precisión α_t que se obtiene de ellos.

Capítulo 3

Privacidad en la bola nuclear

3.1. Mecanismo exponencial para una cuadrática

Estamos interesados en, dada una matriz semidefinida positiva $A(S) \in \mathbb{R}^{d\times d}$, que se construye usando un lote de datos S, poder computar privadamente (y de manera aproximada) el vector propio principal unitario $x \in \mathbb{R}^d$ correspondiente a ella; esto es, un vector asociado al mayor valor propio $\lambda \geq 0$ de A. Específicamente, queremos que este cómputo sea realizado por una variación del mecanismo exponencial que nos permita obtener mayor precisión.

Recordemos que el mecanismo exponencial en su variante clásica [17] nos dice que, para poder obtener privadamente (ε-DP) un vector propio principal, nos basta con muestrear de una distribución

$$d\mu(x) \propto \exp\left(\frac{\varepsilon}{2\Delta(u_A)} u_A(x)\right) dx, \tag{3.1}$$

donde $u_A(\cdot)$ es una *función de puntuación*, una función que, tal como indica su nombre, indica qué tan «útil» (a mayor utilidad, mayor puntaje) es un determinado vector. La función de puntuación que usaremos es la función $u_A(x) = x^\top Ax$, la cuadrática en la esfera, puesto que los vectores que maximizan esta función son, precisamente, los vectores propios principales.

Naturalmente, si queremos que esta variante sea *más* precisa que la formulación clásica del mecanismo exponencial, debemos especificar cuál es nuestro criterio de precisión: Queremos un vector (aleatorio) $x \in \mathbb{S}^{d-1}$ tal que, si llamamos $\lambda \geq 0$ al valor propio principal de A,

$$\lambda - \mathbb{E}[x^\top Ax]$$

sea lo más pequeño posible.

Puede encontrarse una implementación eficiente de este algoritmo en Ge et al. [12].

Sobre la notación: A lo largo de este capítulo usamos ∇ para denotar tanto el gradiente Riemanniano en la esfera como el gradiente Euclidiano. La manera de distinguirlos es fijándose

si tiene o no una barra encima: ∇f indica el gradiente Riemanniano y $\nabla \bar{f}$ indica el gradiente Euclidiano de $\bar{f}$, donde $\bar{f}$ es una extensión suave de f (ver Definición A.1.4). A veces también se usa la notación «grad» para denotar el gradiente Euclidiano; esta es usada para casos en que se trabaja con una función concreta. Por ejemplo, en vez de escribir $\nabla\left(\overline{x^\top Ax}\right)$, escribimos grad $\left(x^\top Ax\right)$.
Por último, Γ, que denota el operador de Carré du Champ, siempre es (en esta sección) $\Gamma(f,g) = \nabla f \cdot \nabla g$. Como las funciones no tienen barra encima, quiere decir que los gradientes son en el sentido Riemanniano.

3.1.1. En busca de un nuevo mecanismo exponencial

Notemos que la densidad (3.1) del mecanismo exponencial se concentra, debido a la función de puntuación $x^\top Ax$, alrededor de los puntos de la esfera que maximizan esta función. Estos puntos son precisamente los vectores propios principales de A.
Una manera de aumentar esta concentración –para así aumentar la precisión– es aumentando el tamaño del factor ε que acompaña a u_A (usualmente un ε útil puede ser $\varepsilon \sim 0.1$). Como es de esperar, al aumentar el tamaño de este ε, el análisis del mecanismo exponencial nos llevará a concluir que hemos empeorado la privacidad del mecanismo. Debemos, entonces, cambiar la noción de privacidad que exigiremos al mecanismo: ya no será ε-DP, sino (ε,δ)-DP.

Usaremos a continuación las ideas de Minami et al. [18] para un análisis de (ε,δ)-DP del mecanismo exponencial. Como la letra ε es por convención el parámetro de privacidad, modificamos levemente la expresión (3.1) para recalcar que el parámetro que usaremos en la perturbación de la medida es distinto del parámetro de privacidad. La densidad que estudiaremos, entonces, es la siguiente:

$$dG_{\beta,S}(x) \propto \exp\left(\beta x^\top Ax\right) dx. \tag{3.2}$$

El siguiente teorema es una instanciación del Teorema 2.1.7, y constituye el primer paso para el análisis de (ε,δ)-DP del mecanismo exponencial. Los otros contenidos de esta sección son desarrollados para poder aplicar este teorema:

Teorema 3.1.1. Sean $\varepsilon,\delta > 0$ los parámetros de privacidad y suponga que

$$G_{\beta,S}\left\{\log \frac{dG_{\beta,S}(x)}{dG_{\beta,S'}(x)} \geq \varepsilon\right\} \leq \delta$$

para todo $S \simeq S'$. Entonces el vector x muestreado mediante la densidad (3.2) es (ε,δ)-DP.

Demostración. Tal como se mencionó, esta es una instancia del Teorema 2.1.7: $\mathbb{P}_{\mathcal{M}(S)} = G_{\beta,S}$ y $\mathcal{L}_{S,S'} = \log \frac{dG_{\beta,S}(x)}{dG_{\beta,S'}(x)}$, donde $x \sim G_{\beta,S}$ por hipótesis. ■

3.1.2. Cómo se aplicará el Teorema 3.1.1

En esta subsección enlistaremos los ingredientes que vamos a necesitar para poder obtener (ε,δ)-DP. En el camino nos encontraremos con algunos *Carré du Champ*, Γ. En la siguiente subsección nos encargaremos de dar cotas para estas expresiones.

1. Sea $D_{\mathrm{KL}}(G_{\beta,S}, G_{\beta,S'}) = \mathbb{E}_{G_{\beta,S}}\left[\log \frac{dG_{\beta,S}}{dG_{\beta,S'}}\right]$, la divergencia de Kullback-Leibler [9]. Es claro que si $D_{\mathrm{KL}}(G_{\beta,S}, G_{\beta,S'}) \leq R$, para algún $R \in \mathbb{R}$, entonces

$$\left\{\log \frac{dG_{\beta,S}}{dG_{\beta,S'}} \geq \varepsilon\right\} \subseteq \left\{\log \frac{dG_{\beta,S}}{dG_{\beta,S'}} \geq D_{\mathrm{KL}}(G_{\beta,S}, G_{\beta,S'}) + (\varepsilon - R)\right\}.$$

Y al aplicar monotonía de la medida, obtenemos que

$$G_{\beta,S}\left\{\log \frac{dG_{\beta,S}}{dG_{\beta,S'}} \geq \varepsilon\right\} \leq G_{\beta,S}\left\{\log \frac{dG_{\beta,S}}{dG_{\beta,S'}} \geq D_{\mathrm{KL}}(G_{\beta,S}, G_{\beta,S'}) + (\varepsilon - R)\right\}. \tag{3.3}$$

Para tratar (3.3), necesitaremos, por una parte, usar una desigualdad de concentración para el lado derecho, y por otra, acotar superiormente $D_{\mathrm{KL}}(G_{\beta,S}, G_{\beta,S'})$.
Ambos objetos serán trabajados mediante la misma propiedad de $G_{\beta,S}$: *la desigualdad de log-Sobolev* (LSI por sus siglas en inglés). Recordemos que (ver definición 2.2.7) una medida μ satisface LS(C), $C > 0$, si para toda función integrable f:

$$\int_\Omega f^2 \log f^2 d\mu - \left(\int_\Omega f^2 d\mu\right) \cdot \log\left(\int_\Omega f^2 d\mu\right) \leq 2C \int_\Omega |\nabla f|^2 \, d\mu \tag{3.4}$$

De momento asumamos que $G_{\beta,S}$ satisface LS($1/\rho$), sin prestar atención, por ahora, a qué es este objeto ρ.

2. Recordemos que (ver Teorema 2.2.3) la LS(C) para una medida μ nos entrega la siguiente desigualdad de concentración: Para toda función $f \in \mathcal{C}^\infty(M)$ tal que $\|f\|_{\mathrm{Lip}} < \infty$ y para todo $r \geq 0$,

$$\mu\left\{f \geq \int_E f d\mu + r\right\} \leq \exp\left(-\frac{r^2}{2C\|f\|^2_{\mathrm{Lip}}}\right).$$

Por lo que, si $\varepsilon > R$ (la cota superior de $D_{\mathrm{KL}}(G_{\beta,S}, G_{\beta,S'})$), y recordando que $\mathbb{E}_{G_{\beta,S}}\left[\log \frac{dG_{\beta,S}}{dG_{\beta,S'}}\right] = D_{\mathrm{KL}}(G_{\beta,S}, G_{\beta,S'})$, entonces

$$\begin{aligned} G_{\beta,S}\left\{\log \frac{dG_{\beta,S}}{dG_{\beta,S'}} \geq \varepsilon\right\} &\leq G_{\beta,S}\left\{\log \frac{dG_{\beta,S}}{dG_{\beta,S'}} \geq D_{\mathrm{KL}}(G_{\beta,S}, G_{\beta,S'}) + (\varepsilon - R)\right\} \\ &\leq \exp\left(-\frac{\rho}{2\left\|\log \frac{dG_{\beta,S}}{dG_{\beta,S'}}\right\|^2_{\mathrm{Lip}}} \cdot (\varepsilon - R)^2\right). \end{aligned} \tag{3.5}$$

3. Acotemos $D_{\mathrm{KL}}(G_{\beta,S}, G_{\beta,S'})$:

$$\begin{aligned}
D_{\mathrm{KL}}(G_{\beta,S}, G_{\beta,S'}) &= \mathbb{E}_{G_{\beta,S}}\left[\log \frac{dG_{\beta,S}}{dG_{\beta,S'}}\right] \\
&= \int_{\mathbb{S}^{d-1}} \log\left(\frac{dG_{\beta,S}}{dG_{\beta,S'}}\right) dG_{\beta,S} \\
&= \int_{\mathbb{S}^{d-1}} \log\left(\frac{dG_{\beta,S}}{dG_{\beta,S'}}\right)\left(\frac{dG_{\beta,S}}{dG_{\beta,S'}}\right) dG_{\beta,S'} \\
&= \int_{\mathbb{S}^{d-1}} \log\left(\frac{dG_{\beta,S}}{dG_{\beta,S'}}\right)\left(\frac{dG_{\beta,S}}{dG_{\beta,S'}}\right) dG_{\beta,S'} + \\
&\quad - \left(\int_{\mathbb{S}^{d-1}} \left(\frac{dG_{\beta,S}}{dG_{\beta,S'}}\right) dG_{\beta,S'}\right) \cdot \log\left(\int_{\mathbb{S}^{d-1}} \left(\frac{dG_{\beta,S}}{dG_{\beta,S'}}\right) dG_{\beta,S'}\right) \\
&\leq \frac{2}{\rho} \int_{\mathbb{S}^{d-1}} \Gamma\left(\sqrt{\frac{dG_{\beta,S}}{dG_{\beta,S'}}}\right) dG_{\beta,S'}, \qquad (3.6)
\end{aligned}$$

donde en la cuarta línea ocupamos que $\log\left(\int_{\mathbb{S}^{d-1}} \frac{dG_{\beta,S}}{dG_{\beta,S'}} dG_{\beta,S'}\right) = \log(1) = 0$ y en la quinta, la desigualdad de log-Sobolev.

Postergamos momentáneamente –para evitar ser demasiado reiterativos– la cota para esta integral. Esta y el cálculo de una cota para $\left\|\log \frac{dG_{\beta,S}}{dG_{\beta,S'}}\right\|_{\mathrm{Lip}}$ las desarrollaremos en la siguiente subsección.

3.1.3. Cotas del operador Γ

Cota para $\left\|\log \frac{dG_{\beta,S}}{dG_{\beta,S'}}\right\|^2_{\mathbf{Lip}}$

Comencemos recordando que

$$\begin{aligned}
\|f\|^2_{\mathrm{Lip}} &= \|\Gamma(f)\|_\infty \\
&= \sup_{x \in \mathbb{S}^{d-1}} |\nabla f(x)|^2 .
\end{aligned}$$

Por otra parte,

$$\begin{aligned}
\Gamma\left(\log \frac{dG_{\beta,S}}{dG_{\beta,S'}}\right) &= \left|\nabla\left(\log \frac{dG_{\beta,S}}{dG_{\beta,S'}}\right)\right|^2 \\
&= \left|\left(I - xx^\top\right) \operatorname{grad}\left(\overline{\log \frac{dG_{\beta,S}}{dG_{\beta,S'}}}\right)\right|^2 \\
&= \left|\left(I - xx^\top\right) \operatorname{grad}\left(\beta x^\top [A(S) - A(S')]\, x\right)\right|^2 \\
&= \beta^2 \left|\left(I - xx^\top\right)(2\,[A(S) - A(S')]\, x)\right|^2 \\
&= \beta^2 \left(4\,|[A(S) - A(S')]\, x|^2 - \left(x^\top [A(S) - A(S')] x\right)^2\right) \\
&\leq 4\beta^2 \|A(S) - A(S')\|^2_{\mathrm{op}} \\
&\leq 4\beta^2 \Delta_{\mathrm{OP}}(A)^2
\end{aligned}$$

3.1.4. Cota para $\Gamma\left(\sqrt{\frac{dG_{\beta,S}}{dG_{\beta,S'}}}\right)$

$$\begin{aligned}
\Gamma\left(\sqrt{\frac{dG_{\beta,S}}{dG_{\beta,S'}}}\right) &= \left|\nabla\left(\sqrt{\frac{dG_{\beta,S}}{dG_{\beta,S'}}}\right)\right|^2 \\
&= \left|\nabla\left(\exp\left(\frac{1}{2}\log\left(\frac{dG_{\beta,S}}{dG_{\beta,S'}}\right)\right)\right)\right|^2 \\
&= \left|\left(I - xx^\top\right)\operatorname{grad}\left(\exp\left(\frac{1}{2}\overline{\log}\left(\frac{dG_{\beta,S}}{dG_{\beta,S'}}\right)\right)\right)\right|^2 \\
&= \left|\left(I - xx^\top\right)\frac{1}{2}\exp\left(\frac{1}{2}\overline{\log}\left(\frac{dG_{\beta,S}}{dG_{\beta,S'}}\right)\right)\operatorname{grad}\left(\overline{\log}\left(\frac{dG_{\beta,S}}{dG_{\beta,S'}}\right)\right)\right|^2 \\
&= \left|\left(I - xx^\top\right)\frac{1}{2}\sqrt{\frac{dG_{\beta,S}}{dG_{\beta,S'}}}\operatorname{grad}\left(\overline{\log}\left(\frac{dG_{\beta,S}}{dG_{\beta,S'}}\right)\right)\right|^2 \\
&= \frac{1}{4}\cdot\frac{dG_{\beta,S}}{dG_{\beta,S'}}\left|\left(I - xx^\top\right)\operatorname{grad}\left(\overline{\log}\left(\frac{dG_{\beta,S}}{dG_{\beta,S'}}\right)\right)\right|^2 \\
&\leq \frac{dG_{\beta,S}}{dG_{\beta,S'}}\beta^2\Delta_{\mathrm{OP}}(A)^2
\end{aligned}$$

Por lo que (3.6) se transforma en

$$\begin{aligned}
D_{\mathrm{KL}}(G_{\beta,S}, G_{\beta,S'}) &\leq \frac{2}{\rho}\int_{\mathbb{S}^{d-1}}\Gamma\left(\sqrt{\frac{dG_{\beta,S}}{dG_{\beta,S'}}}\right)dG_{\beta,S'} \\
&\leq \frac{2\beta^2}{\rho}\int_{\mathbb{S}^{d-1}}\frac{dG_{\beta,S}}{dG_{\beta,S'}}\Delta_{\mathrm{OP}}(A)^2 dG_{\beta,S'} \\
&= \frac{2\beta^2}{\rho}\int_{\mathbb{S}^{d-1}}\Delta_{\mathrm{OP}}(A)^2 dG_{\beta,S} \\
&= \frac{2\beta^2}{\rho}\cdot\Delta_{\mathrm{OP}}(A)^2.
\end{aligned}$$

Reemplazando estas expresiones en (3.5), obtenemos que

$$G_{\beta,S}\left\{\log\frac{dG_{\beta,S}}{dG_{\beta,S'}} \geq \varepsilon\right\} \leq \exp\left(-\frac{\rho}{8\beta^2\left\|A(S) - A(S')\right\|_{\mathrm{op}}^2}\cdot\left(\varepsilon - \frac{2\beta^2}{\rho}\cdot\left\|A(S) - A(S')\right\|_{\mathrm{op}}^2\right)^2\right) \tag{3.7}$$

Al concatenar esta seguidilla de hechos, resulta el siguiente teorema:

Teorema 3.1.2. Si $\varepsilon, \delta \leq 1$ y $\beta \leq \frac{\sqrt{\rho/2}}{4}\cdot\frac{\varepsilon}{\sqrt{\log(1/\delta)}\Delta_{\mathrm{Op}}(A)}$, entonces el vector (aleatorio) x en la esfera muestreado mediante la distribución (3.2) satisface (ε, δ)-DP.

Demostración. La demostración es sencilla: Basta verificar que esta elección de β hace al lado derecho de (3.7) más pequeño que δ, luego concluimos usando el Teorema 3.1.1.

1. Comenzaremos viendo que la restricción elegida permite obtener la cota

$$\left(\varepsilon - \frac{2\beta^2}{\rho} \cdot \Delta_{\mathrm{Op}}(A)^2\right)^2 \geq \frac{\varepsilon^2}{4}. \tag{3.8}$$

En efecto, si

$$\beta \leq \frac{\sqrt{\rho\varepsilon}}{2\Delta_{\mathrm{Op}}(A)},$$

entonces (simplemente moviendo términos)

$$\left(\varepsilon - \frac{2\beta^2}{\rho} \cdot \Delta_{\mathrm{Op}}(A)^2\right) \geq \frac{\varepsilon}{2}.$$

Como

$$\frac{\sqrt{\rho/2}}{4} \cdot \frac{\varepsilon}{\sqrt{\log(1/\delta)}\Delta_{\mathrm{Op}}(A)} \leq \frac{\sqrt{\rho\varepsilon}}{2\Delta_{\mathrm{Op}}(A)},$$

pues ε y $\delta \leq 1$, (3.8) se cumple.

2. Veamos ahora que si

$$\beta \leq \frac{\sqrt{\rho/2}}{4} \cdot \frac{\varepsilon}{\sqrt{\log(1/\delta)}\Delta_{\mathrm{Op}}(A)},$$

entonces

$$\exp\left(-\frac{\rho}{8\beta^2\Delta_{\mathrm{Op}}(A)^2} \cdot \left(\varepsilon - \frac{2\beta^2}{\rho} \cdot \Delta_{\mathrm{Op}}(A)^2\right)^2\right) \leq \delta.$$

En efecto,

$$\exp\left(-\frac{\rho}{8\beta^2\Delta_{\mathrm{Op}}(A)^2} \cdot \left(\varepsilon - \frac{2\beta^2}{\rho} \cdot \Delta_{\mathrm{Op}}(A)^2\right)^2\right) \leq \exp\left(-\frac{\rho}{8\beta^2\Delta_{\mathrm{Op}}(A)^2} \cdot \frac{\varepsilon^2}{4}\right) \tag{3.9}$$

Por el paso 1.
Además, tal elección de β implica que (simplemente moviendo términos)

$$\log(1/\delta) \leq \frac{\rho}{8\beta^2\Delta_{\mathrm{OP}}(A)^2} \cdot \frac{\varepsilon^2}{4}$$

Lo que implica, a su vez, que

$$-\frac{\rho}{8\beta^2\Delta_{\mathrm{OP}}(A)^2} \cdot \frac{\varepsilon^2}{4} \leq \log(\delta) \tag{3.10}$$

Luego, usando (3.9) y (3.10), obtenemos que

$$\exp\left(-\frac{\rho}{8\beta^2\Delta_{\mathrm{Op}}(A)^2} \cdot \left(\varepsilon - \frac{2\beta^2}{\rho} \cdot \Delta_{\mathrm{Op}}(A)^2\right)^2\right) \leq \delta.$$

■

3.1.5. El parámetro de curvatura

Como se habrá podido notar, en las fórmulas recién presentadas aparece un número, ρ, del que no hemos especificado nada. Este número es no-negativo ($\rho \geq 0$) y corresponde a la curvatura de la esfera vista con la medida perturbada $G_{\beta,S}$, quedando cuantificado mediante la expresión (ver corolario 2.2.2):

$$\begin{aligned} \mathrm{Ric}_{\mathfrak{g}}(\nabla f, \nabla f) + \nabla\nabla W(\nabla f, \nabla f) &= (d-2)|\nabla f|^2 + \nabla\nabla W(\nabla f, \nabla f) \\ &\geq \rho |\nabla f|^2, \end{aligned} \tag{3.11}$$

para toda $f \in \mathcal{C}^\infty(\mathbb{S}^{d-1})$, y donde $W(x) = \beta x^\top A x$. Cabe aclarar que tanto los gradientes, como el hessiano en la expresión (3.11) son en el sentido Riemanniano.

Para tener un valor de ρ, debemos, entonces, calcular el hessiano de $W(x) = \beta x^\top A x$. Como subproducto de este resultado, obtendremos una cota superior para el valor de β.

Cálculo de $\nabla\nabla W(x)$

Siguiendo (2.6), calcularemos el hessiano mediante una iteración del Carré du Champ. En específico, haremos uso de la siguiente propiedad:

$$\nabla\nabla W(\nabla f, \nabla f) = \Gamma(f, \Gamma(W, f)) - \frac{1}{2}\Gamma(W, \Gamma(f)).$$

Para lo que sigue, debemos recordar que $\Gamma(f,g) = \nabla f(x)^\top \nabla g(x)$. Para ver cómo calcular los gradientes en el sentido Riemanniano, revisar el ejemplo A.1.1 del apéndice.
Haremos los cálculos de los Carré du Champ por separado y cada vez que se mencione una función con una barra encima –por ejemplo, $\bar{f}$–, queremos hacer alusión a una extensión suave de dicha función –en este caso, $\bar{f}$ es una extensión suave de f–. Comencemos de derecha a izquierda:

$$\begin{aligned} \Gamma(W, \Gamma(f)) &= \nabla W(x)^\top \nabla \Gamma(f) \\ &= \left(I - xx^\top\right) \nabla \bar{W}(x)^\top \left(I - xx^\top\right) \nabla \bar{\Gamma}(f) \end{aligned}$$

Por su parte,

$$\begin{aligned} \nabla \bar{\Gamma}(f) &= \nabla \left(\left| \left(I - xx^\top\right) \nabla \bar{f} \right|^2 \right) \\ &= 2\nabla^2 \bar{f}(x) \left[\left(I - xx^\top\right) \nabla \bar{f}(x) \right]. \end{aligned}$$

Y

$$\nabla \bar{W}(x) = 2\beta A x$$

Por lo que

$$\begin{aligned} \Gamma(W, \Gamma(f)) &= 4\beta \left[\left(I - xx^\top\right) Ax \right]^\top \left(I - xx^\top\right) \nabla^2 \bar{f}(x) \left[\left(I - xx^\top\right) \nabla \bar{f}(x) \right] \\ &= 4\beta x^\top A \left(I - xx^\top\right)^2 \nabla^2 \bar{f}(x) \left[\left(I - xx^\top\right) \nabla \bar{f}(x) \right] \\ &= 4\beta x^\top A \left(I - xx^\top\right) \nabla^2 \bar{f}(x) \left[\left(I - xx^\top\right) \nabla \bar{f}(x) \right], \end{aligned}$$

donde en la segunda línea ocupamos que tanto A como $\left(I - xx^\top\right)$ son matrices simétricas, y en la tercera línea, que, al ser $\left(I - xx^\top\right)$ una proyección, es idempotente.

Pasemos ahora al cálculo de la segunda expresión:

$$\begin{aligned}\Gamma\left(f, \Gamma\left(W, f\right)\right) &= \nabla f(x)^\top \nabla \Gamma\left(W, f\right) \\ &= \left[\left(I - xx^\top\right) \nabla \bar{f}(x)\right]^\top \left(I - xx^\top\right) \nabla \bar{\Gamma}\left(W, f\right) \\ &= \nabla \bar{f}(x)^\top \left(I - xx^\top\right) \nabla \bar{\Gamma}\left(W, f\right)\end{aligned}$$

Por su parte,

$$\begin{aligned}\Gamma\left(W, f\right) &= \nabla W(x)^\top \nabla f(x) \\ &= \left[\left(I - xx^\top\right) 2\beta Ax\right]^\top \left(I - xx^\top\right) \nabla \bar{f}(x) \\ &= 2\beta x^\top A \left(I - xx^\top\right) \nabla \bar{f}(x)\end{aligned}$$

Entonces

$$\nabla \bar{\Gamma}\left(W, f\right) = 2\beta \left(A \nabla \bar{f}(x) + \nabla^2 \bar{f}(x) Ax - \left(x^\top Ax\right) \nabla \bar{f}(x) - \left(x^\top Ax\right) \nabla^2 \bar{f}(x) x - 2\left(x^\top \nabla \bar{f}(x)\right) Ax\right).$$

Por lo que

$$\begin{aligned}\Gamma\left(f, \Gamma\left(W, f\right)\right) = 2\beta \big(&\nabla \bar{f}(x)^\top \left(I - xx^\top\right) A \nabla \bar{f}(x) + \nabla \bar{f}(x)^\top \left(I - xx^\top\right) \nabla^2 \bar{f}(x) Ax \\ &- \left(x^\top Ax\right) \nabla \bar{f}(x)^\top \left(I - xx^\top\right) \nabla \bar{f}(x) - \left(x^\top Ax\right) \nabla \bar{f}(x)^\top \left(I - xx^\top\right) \nabla^2 \bar{f}(x) x \\ &- 2\left(x^\top \nabla \bar{f}(x)\right) \nabla \bar{f}(x)^\top \left(I - xx^\top\right) Ax \big)\end{aligned}$$

Juntando todo esto, obtenemos que

$$\nabla \nabla W(x)\left(\nabla f(x), \nabla f(x)\right) = 2\beta \left(\frac{\nabla f(x)^\top}{|\nabla f(x)|} A \frac{\nabla f(x)}{|\nabla f(x)|} - x^\top Ax\right) |\nabla f(x)|^2 \tag{3.12}$$

Notar que esto es una resta de cuadráticas, donde ambas son evaluadas en vectores unitarios. Por lo que una cota inferior para esta expresión es:

$$\nabla \nabla W(x)\left(\nabla f(x), \nabla f(x)\right) \geq -2\beta\lambda \left|\nabla f(x)\right|^2, \tag{3.13}$$

donde $\lambda \geq 0$ es el valor propio principal de A.

Juntando (3.11) con (3.13), obtenemos que

$$\begin{aligned}\mathrm{Ric}_{\mathfrak{g}}(\nabla f, \nabla f) + \nabla \nabla W(\nabla f, \nabla f) &\geq \left(d - 2 - 2\beta\lambda\right) |\nabla f|^2 \\ &\geq \rho \left|\nabla f\right|^2\end{aligned} \tag{3.14}$$

De esta expresión, resulta que una cota factible para β es

$$\beta \leq \frac{d-2}{4\lambda}, \tag{3.15}$$

de la que se sigue que podemos tomar

$$\rho \geq \frac{d-2}{2}. \tag{3.16}$$

Observación 3.1.1. Notemos que el cálculo anterior no es más que una instanciación del corolario 2.2.2. Este nos permite obtener la desigualdad de log-Sobolev –su constante, en específico– mediante la condición de curvatura $CD(\rho, \infty)$. Sin embargo, como mencionamos en el preliminar de concentración de la medida, existe otro método para hacer este cálculo (demostrar que una tripleta Markoviana satisface log-Sobolev): el Teorema de Holley-Stroock (Teorema 2.2.2). El problema que surge con esta técnica y por qué es preferible –al menos en este estudio– el uso de la condición de curvatura viene dado por el tipo constante de log-Sobolev que entrega Holley-Stroock: esta constante es exponencial en β.
En efecto, la tripleta $(\mathbb{S}^{d-1}, dx, \Gamma)$, donde dx es la medida uniforme en la esfera y Γ es el carré du Champ usual, satisface $CD(d-2, \infty)$. Por lo que, en virtud del Teorema 2.2.4, satisface $LS\left(1/(d-2)\right)$. Por ende, Holley-Stroock nos garantiza que $\left(\mathbb{S}^{d-1}, \frac{e^{\beta x^\top A(S)x} dx}{\int_{\mathbb{S}^{d-1}} e^{\beta y^\top A(S)y} dy}, \Gamma\right)$ satisface $LS\left(e^{\beta\lambda}/(d-2)\right)$. Como queremos que β sea grande, para una mayor precisión, esta exponencial arruina rápidamente nuestras esperanzas.

3.1.6. Limitaciones del método

De la expresión obtenida en (3.14) se evidencia una limitación en β debido al análisis de privacidad obtenido mediante la desigualdad de log-Sobolev: El tamaño del parámetro β tiene un «techo», pues $2\beta\lambda$ no puede superar en magnitud a $d-2$. Esto es problemático, pues si ocupamos una mayor cantidad de datos, es esperable que la precisión que obtengamos sea más ajustada (esto es evidente en, por ejemplo, la precisión deseada $\tilde{O}\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d}}{\varepsilon n}\right)$, pues n representa la cantidad de datos). Sin embargo, al tener β este techo, nuestro mecanismo queda condenado a tener una precisión limitada y, en cierto punto, inmejorable con una mayor cantidad de datos.

Observación 3.1.2. El parámetro β no tendría este techo si en (3.13) hubiésemos obtenido $+2\beta\lambda \left|\nabla f(x)\right|^2$, en vez de la expresión con signo negativo. Esto ocurriría si es que la función $\beta x^\top A(S)x$ fuese *geodésicamente fuerte convexa*. Sin embargo, no es posible definir una función de tal tipo en toda la esfera.

Otra limitación con la que cuenta el mecanismo es la dificultad de dar una respuesta satisfactoria a la precisión de este. El trabajo de Kume y Walker [16] trata sobre la precisión de $\mathbb{E}_{G_{\beta,S}}\left[x^\top Ax\right]$ cuando $x \in \mathbb{S}^{d-1}$ es muestreado mediante la distribución $G_{\beta,S}$, mas solo cuando la dimensión es alta (lo cual no es problemático en nuestro caso) y los valores propios de βA no son comparables a d. Como β puede ser $O(d)$ (y, de hecho, tiene incentivos para tomar valores de ese orden para entregar una mayor precisión), esta restricción elimina la posibilidad de realizar un análisis de precisión con las técnicas de ese artículo.

Por otra parte, tenemos el trabajo de Kapralov y Talwar [15], el cual entrega la precisión de $\mathbb{E}_\nu\left[x^\top Ax\right]$, cuando $x \in \mathbb{S}^{d-1}$ es muestreado con respecto a $\nu \propto \exp\left(\varepsilon x^\top Ax\right)$; o sea, usando el mecanismo exponencial. Su análisis se basa en la estimación de áreas de casquetes esféricos de $\mathbb{S}^{d-1}$ y ocupa que para obtener una determinada precisión, el valor propio principal debe ser lo

suficientemente grande. En particular, si se desea

$$\mathbb{E}_\nu \left[x^\top A x\right] \geq \lambda(1-\gamma),$$

donde $\lambda \geq 0$ es el valor propio principal de A y $\gamma > 0$ es el error que se quiere obtener, entonces $\lambda = O\left(\frac{d\log(1/\gamma)}{\varepsilon\gamma}\right)$, restricción inaceptable en nuestro caso de estudio.

Ahora bien, el análisis hecho por [15] ocupa ε, una constante fija, así que podríamos preguntarnos qué ocurre cuando este es reemplazado por nuestro β, que tiene permitido crecer tanto como la dimensión. En este caso, emulando el argumento del Lema 3.1 de [15], obtenemos el siguiente resultado:

Teorema 3.1.3. Sea $\gamma \in (0,1)$ y sea $x \in \mathbb{S}^{d-1}$ muestreado mediante la distribución $G_{\beta,S}$. Luego

$$\mathbb{E}_{G_{\beta,S}} \left[x^\top A(S) x\right] \geq \lambda(1-2\gamma)\left(1 - e^{\frac{d-1}{2}\log(4/\gamma) - \beta\lambda\gamma}\right),$$

donde $\lambda \geq 0$ es el valor propio principal de $A(S)$.

Debido a (3.14), para que el resultado sea privado, β a lo más (de hecho, nunca podrá tomar este valor, siempre deberá ser más pequeño) puede ser $\beta = \frac{d-2}{2\lambda}$. Por lo que el análisis de precisión anterior es insuficiente para cualquier γ significativo.

Observación 3.1.3. Notemos que el análisis anterior es insuficiente. Con esto queremos decir que puede que no sea lo suficientemente refinado para capturar adecuadamente los parámetros ocupados en este problema. De ninguna forma este descarta que pueda realmente obtenerse una precisión significativa (aunque, hasta cierto punto, limitada, debido a la discusión con la que comienza esta subsección).

3.2. Algoritmo de Oja

A continuación presentamos un algoritmo que resuelve el subproblema semidefinido de minimización sobre la bola nuclear. Este es conocido como el algoritmo de Oja (usamos la versión presentada en Jain et al. [14]) y consiste en:

Algoritmo 3: SFW en bola nuclear - Algoritmo de Oja

Requiere: Matriz $A(S) \in \mathrm{Sym}_d^+$, parámetros de privacidad (ε, δ) tales que $\varepsilon \leq 2\log(1/\delta)$, número total de iteraciones T.

Establecer $\sigma = \frac{\log(1.25/\delta)\Delta_{\mathrm{OP}}(A)\sqrt{32T\log(1/\delta)}}{\varepsilon}$

Establecer $\eta = \frac{1}{T\sigma\sqrt{d}}$

Tomar $\bar{x}^{(0)} = g^{(0)} \sim \mathcal{N}\left(0, \sigma^2 I\right)$

Normalizar $x^{(0)} = \bar{x}^{(0)} / \left\|\bar{x}^{(0)}\right\|_2$

for $t = 1, \ldots, T$ **do**

  Tomar $g^{(t)} \sim \mathcal{N}\left(0, \sigma^2 I\right)$

  Tomar $x^{(t)} = \frac{x^{(t-1)} + \eta\left(Ax^{(t-1)} + g^{(t)}\right)}{\left\|x^{(t-1)} + \eta\left(Ax^{(t-1)} + g^{(t)}\right)\right\|_2}$

Entrega: $x^{(T)}$

Observación 3.2.1. La restricción $\varepsilon \leq 2\log(1/\delta)$ no es demandante si trabajamos en dimensiones altas, pues no queremos un ε demasiado grande (un buen ε puede ser 0.1 y, en general, pensaremos $0 < \varepsilon \leq 1$). Además, para que el resultado no produzca una grave violación en la privacidad de los participantes, debemos pedir que $\delta \ll 1/n$ (de hecho, usualmente debe pedirse que $1/\delta$ sea superpolinomial en la dimensión). Como la cota de error depende tanto de la dimensión como de la cantidad de datos, un error pequeño –que es lo deseable– necesita que la cantidad de datos sea comparable a la dimensión.

Lema 3.2.1. El algoritmo 3 es $(\varepsilon, (T+1)\delta)$-DP.

Demostración. Sean $\bar{x}^{(t)} = x^{(t-1)} + \eta A(S)x^{(t-1)}$ y $\bar{x}^{(t)'} = x^{(t-1)} + \eta A(S')x^{(t-1)}$. Luego

$$\begin{aligned}\Delta_2(\bar{x}^{(t)}) &= \left\|\bar{x}^{(t)} - \bar{x}^{(t)'}\right\|_2 \\ &= \eta \left\|\left(A(S) - A(S')\right) x^{(t-1)}\right\|_2 \\ &\leq \eta \left\|A(S) - A(S')\right\|_{\mathrm{op}} \\ &= \eta \Delta_{\mathrm{OP}}(A)\end{aligned}$$

Luego, por el Teorema 2.1.8, al sumar $\bar{g}^t \sim \mathcal{N}\left(0, \frac{32\eta \log(1.25/\delta)^2 \Delta_{\mathrm{OP}}(A)^2 T \log(1/\delta)}{\varepsilon}\right)$, garantizamos que

$$\bar{x}^{(t)} = x^{(t-1)} + \eta A(S)x^{(t-1)} + \bar{g}^{(t)}$$

sea $\left(\frac{\varepsilon}{2\sqrt{2T\log(1/\delta)}}, \delta\right)$-DP. Notar que $\bar{g}^t = \eta g^{(t)}$, donde $g^{(t)} \sim \mathcal{N}\left(0, \frac{32\log(1.25/\delta)^2 \Delta_{\mathrm{OP}}(A)^2 T \log(1/\delta)}{\varepsilon}\right)$. Usando esto y post-procesamiento (Teorema 2.1.1), obtenemos que

$$x^t = \frac{x^{t-1} + \eta\left(A(S)x^{(t-1)} + g^{(t)}\right)}{\left\|x^{t-1} + \eta\left(A(S)x^{(t-1)} + g^{(t)}\right)\right\|_2}$$

es $\left(\frac{\varepsilon}{2\sqrt{2T\log(1/\delta)}}, \delta\right)$-DP.

Finalmente, por teorema de composición avanzada (Teorema 2.1.10), el algoritmo de Oja es $(\varepsilon, T\delta + \delta)$. ■

Teorema 3.2.1 (Precisión del algoritmo de Oja (Teorema 3.3 de [14]))**.** El T-ésimo iterado del algoritmo 3, $x^{(T)}$, satisface, con probabilidad al menos $1 - 1/\mathrm{poly}(T)$,

$$x^{(T)\top} A(S)x^{(T)} \geq \left\|A(S)\right\|_{\mathrm{op}} - O\left(\frac{\sigma\sqrt{d}}{\sqrt{T}} + \frac{\left\|A(S)\right\|_{\mathrm{op}}}{T}\right),$$

donde $\mathrm{Poly}(T)$ es un polinomio en T.

Corolario 3.2.1. El T-ésimo iterado del algoritmo 3, $x^{(T)}$, satisface, con probabilidad al menos $1 - 1/\mathrm{poly}(T)$,

$$x^{(T)\top} A(S)x^{(T)} \geq \left\|A(S)\right\|_{\mathrm{op}} - O\left(\frac{\log(1.25/\delta)\Delta_{\mathrm{OP}}(A)\sqrt{d}\sqrt{32\log(1/\delta)}}{\varepsilon} + \frac{\left\|A(S)\right\|_{\mathrm{op}}}{T}\right).$$

Usemos este resultado para establecer una tasa de convergencia para el Algoritmo 2:

Teorema 3.2.2. El Algoritmo 2 entrega, con probabilidad al menos $1 - \left(p + \frac{n}{\text{poly}(T)}\right)$, una tasa de error $\tilde{O}\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d}}{\varepsilon\sqrt{n}}\right)$ cuando el paso 5 y el paso 11 son calculados mediante el algoritmo de Oja (algoritmo 3). Además, el resultado es $(\varepsilon, n\,(T+1)\,\delta)$-DP.

Demostración. 1. Notar que

$$\begin{aligned}\alpha_t &= \|D_t\|_{\text{op}} - \text{Tr}\,(D_t V_t) \\ &= \|D_t\|_{\text{op}} - \text{Tr}\left(D_t x(t)^{(T)} x(t)^{(T)\top}\right) \\ &= \|D_t\|_{\text{op}} - x(t)^{(T)\top} D_t x(t)^{(T)},\end{aligned}$$

donde $x(t)^{(T)}$ es lo entregado por el algoritmo de Oja cuando el input es D_t.

2. Por el Lema 3.2.1 y el corolario 3.2.1, cada $x(t)^{(T)}$ es $(\varepsilon', (T+1))$-DP y

$$\begin{aligned}\alpha_t &\leq O\left(\frac{\log(1.25/\delta)\Delta_{\text{OP}}(D_t)\sqrt{d}\sqrt{32\log(1/\delta)}}{\varepsilon'} + \frac{\|D_t\|_{\text{op}}}{T}\right) \\ &\leq O\left(\frac{(2L_1 + L_0)\log(1.25/\delta)\sqrt{32\log(1/\delta)}\sqrt{d}\log(n)}{\varepsilon' n} + \frac{\|D_t\|_{\text{op}}}{T}\right),\end{aligned}$$

donde $\Delta_{\text{OP}}(D_t)$ es reemplazado por el Teorema 2.3.4.

3. Por el Lema 2.3.2, para obtener $(\varepsilon, n(T+1)\delta)$-DP en el resultado final del algoritmo, basta que cada V_t sea $\left(\frac{\varepsilon}{2\sqrt{n\log(1/\delta)}}, (T+1)\delta\right)$. Por ende, ocupando $\varepsilon' = \frac{\varepsilon}{2\sqrt{n\log(1/\delta)}}$, obtenemos que el algoritmo 2 es $(\varepsilon, n(T+1)\delta)$-DP y que

$$\begin{aligned}\alpha_t &\leq O\left(\frac{(2L_1 + L_0)\log(1.25/\delta)\sqrt{32\log(1/\delta)}\sqrt{d}\log(n)}{\varepsilon' n} + \frac{\|D_t\|_{\text{op}}}{T}\right) \\ &\leq O\left(\frac{2\sqrt{32}(2L_1 + L_0)\log(1/\delta)^2\sqrt{d}\log(n)}{\varepsilon\sqrt{n}} + \frac{\|D_t\|_{\text{op}}}{T}\right) \\ &= O\left(\frac{(2L_1 + L_0)\log(1/\delta)^2\log(n)}{\varepsilon}\sqrt{\frac{d}{n}} + \frac{\|D_t\|_{\text{op}}}{T}\right),\end{aligned}$$

donde en la última fila ocupamos la «economía» de la notación O para omitir constantes universales.

4. Si tomamos $T \geq L_0 \cdot \sqrt{\frac{n}{d}}$, tenemos que

$$\alpha_t \leq O\left(\frac{(2L_1 + L_0)\log(1/\delta)^2\log(n)}{\varepsilon}\sqrt{\frac{d}{n}}\right). \tag{3.17}$$

5. Al usar (3.17) y el Teorema 2.3.1, parte (b), obtenemos que

$$\begin{aligned} F_{\mathcal{D}}(X^{(n/2+1)}) - F_{\mathcal{D}}(X^*) \leq & \frac{2L_0}{\sqrt{n}} + \frac{4eL_0\sqrt{2\kappa}\log(n)}{n} + \frac{8e\left(2L_1+L_0\right)\sqrt{\kappa}\log(n)}{\sqrt{n}} \\ & + \frac{4L_0\sqrt{6\log(n/p)}\log(n)}{n} + \frac{8\left(2L_1+L_0\right)\sqrt{3\log(n/p)}\log(n)}{\sqrt{n}} \\ & + \frac{2L_1\log(n)}{n} + O\left(\frac{(2L_1+L_0)\log(1/\delta)^2\log(n)}{\varepsilon}\sqrt{\frac{d}{n}}\right). \end{aligned}$$

6. El algoritmo falla si la cota en el Teorema 2.3.2, parte (b) falla, lo cual ocurre con probabilidad a lo más p; o si el algoritmo de Oja falla en alguno de sus iterados, lo que ocurre con probabilidad a lo más $1/\text{Poly}(T)$ en cada iterado.
   Como son $n/2+1$ iterados, una cota de unión nos dice que el algoritmo falla con probabilidad a lo más $p + \frac{n}{\text{Poly}(T)}$. Lo cual implica que es exitoso con probabilidad a lo menos $1 - \left(p + \frac{n}{\text{Poly}(T)}\right)$.

■

Observación 3.2.2. Que obtengamos $n(T+1)\delta$ en vez de δ no es problemático para la privacidad, pues $1/(n(T+1)\delta)$ sigue siendo superpolinomial en n cuando T es polinomial en n.

3.3. Conclusiones

Tras la indagación de ambos métodos de resolución aproximada y privada del problema de programación semidefinida, podemos concluir que:

(a) El primer método –el mecanismo exponencial en la esfera– no logra su cometido. Es posible muestrear una «solución» privadamente, mas no garantizar que esta «solución» es una buena aproximación a la solución real del problema. En particular, vimos que el análisis de Kapralov y Talwar [15] queda corto al momento de estudiar la precisión de este muestreo y que el estudio de Kume y Walker [16] de la precisión no logra abarcar el caso en que $\beta = O(d)$.

Creemos que uno de los motivos del fracaso de esta técnica es la limitación que obtiene β en (3.14): Nunca podemos elegir β más grande que $O(d)$. Este problema es una consecuencia de los valores propios del Hessiano en (3.13). Una manera de rodear –para así evitar– el problema, puede ser trabajar directamente con la variedad Sym_d^+, en vez de la esfera. Aclaramos que esta es una propuesta puramente indagatoria, pues no conocemos artículos que trabajen en esta línea.

(b) El segundo método –el basado en el algoritmo de Oja– logra parcialmente su cometido. Entrega una aproximación útil y privada, pero no tan buena como esperábamos, pues es $\tilde{O}\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d}}{\varepsilon\sqrt{n}}\right)$, en lugar de $\tilde{O}\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d}}{\varepsilon n}\right)$. Creemos es posible mejorar este resultado si en lugar de usar composición avanzada (que nos hace pagar con un costo $\sqrt{n}$ adicional) para cuantificar la privacidad de los $n/2 + 1$ iterados entregados por Oja, usamos el análisis de árbol desarrollado en Asi et al. [1].

Apéndice A

Cálculo en variedades Riemannianas

A.1. Variedades incrustadas

Este apéndice está basado en la presentación de Boumal [7].

El análisis de privacidad, al realizarse en la esfera, necesita hacer uso de la teoría de variedades. En particular, de la teoría de variedades incrustadas. Presentamos a continuación un breve resumen con las definiciones y los resultados relevantes, entre los que se incluyen: definición de variedad incrustada Riemanniana, definición del diferencial y el gradiente de una función suave, y un método para calcular este último en el caso en que la métrica venga inducida por el espacio ambiente.

El motivo por el que escogemos a las variedades incrustadas es por la facilidad con que se pueden realizar los cálculos en ellas: no es necesario introducir la noción de carta local y gran parte de las computaciones se reducen al caso Euclidiano estándar.

De ahora en adelante cuando se haga referencia a un espacio vectorial E, se asume que es un espacio vectorial finito-dimensional, sobre $\mathbb{R}$ y dotado de un producto interno $\langle \cdot, \cdot \rangle$.

Definición A.1.1 (Variedad incrustada)**.** Sea M un subconjunto de un espacio vectorial E. Decimos que M es una variedad incrustada suave de E si se cumple una de las siguientes condiciones:

1. M es un subconjunto abierto de E. En tal caso, llamamos a M una subvariedad abierta.
2. Existe un entero (fijo) $k \geq 1$ tal que para todo $x \in M$ existe una vecindad $x \in U \subseteq E$ y una función suave $h : U \to \mathbb{R}^k$ tal que
   (a) Si $y \in U$, entonces $h(y) = 0$ si y solo si $y \in M$
   (b) $\text{rank} Dh(x) = k$.

Llamamos a tal E el espacio ambiente de M.

De ahora en adelante, cuando hablemos de «variedad incrustada» queremos decir «variedad incrustada suave».

Definición A.1.2. Sea $M \subseteq E$ e $I \subseteq \mathbb{R}$ un intervalo abierto que contiene al 0. Para cada $x \in M$ definimos

$$T_xM = \{c'(0) \mid c : I \to M \text{ es un camino suave alrededor de 0 y tal que } c(0) = x\}$$

Cuando M es una variedad incrustada y no solamente un conjunto, T_xM toma una forma especial.

Teorema A.1.1. Sea M una variedad incrustada de E y $x \in M$. Luego

- Si M es una subvariedad abierta, entonces $T_xM = E$.
- Si M no es subvariedad abierta, entonces $T_xM = \ker Dh(x)$, donde h es cualquier función suave que defina localmente a M alrededor de x.

Observación A.1.1. Cuando M es una variedad incrustada, T_xM es un subespacio vectorial de E. Además, su dimensión es independiente de x (pues el rango de $Dh(x)$ es constante para toda función que define a la variedad).

Definición A.1.3.

- Cuando M es una variedad incrustada, llamamos a T_xM el espacio tangente de M en x.
- $\dim M := \dim T_xM$.

Necesitamos estudiar los gradientes de funciones sobre la esfera. Por ello, necesitamos introducir los aspectos relevantes de las funciones definidas en variedades que permitan la existencia de tal objeto (el gradiente). Para propósitos de esta tesis, nos bastará con hablar de funciones suaves (infinitamente diferenciables).

Definición A.1.4. Sean M y M' variedades incrustadas en E y E', respectivamente. Una función $F : M \to M'$ es suave en $x \in M$ si existe una vecindad U de x en E y una función $\bar{F} : U \to E'$ la cual es suave (en el sentido usual) en una vecindad U de x en E y tal que

$$F(y) = \bar{F}(y) \qquad \forall y \in M \cap U.$$

Decimos que $\bar{F}$ es una extensión local suave de F alrededor de x. Decimos también que F es suave si es suave para todo $x \in M$.

La localidad de la extensión, de hecho, no es necesaria. Siempre podemos tomar una extensión suave de F que esté definida en toda la variedad.

Teorema A.1.2. Sean M y M' variedades incrustadas en E y E', respectivamente. Una función $F : M \to M'$ es suave si y solo si admite una extensión suave $\bar{F} : U \to E'$ en una vecindad U de M en E tal que

$$F = \bar{F}\,|_M\,.$$

Para poder hablar del gradiente de una función, debemos introducir antes el diferencial de esta. Recordaremos primero la definición del diferencial de una función suave en un espacio Euclidiano y veremos cómo se relaciona con la definición en variedades incrustadas.

Definición A.1.5 (Diferencial en espacio euclideano)**.** Sea $\bar{F} : U \subseteq E \to E'$ una función suave entre espacios vectoriales, restringida a un abierto U de E. Dado un punto $x \in U$, el diferencial de $\bar{F}$ en x es la transformación lineal $D\bar{F}(x) : E \to E'$ definida por:

$$D\bar{F}(x)[v] = \lim_{t \to 0} \frac{\bar{F}(x + tv) - \bar{F}(x)}{t} \tag{A.1}$$

Observación A.1.2. Para toda dirección $v \in E$ hay un $t > 0$ suficiente pequeño tal que el intervalo $[x, x + tv] \subseteq U$. Esto es una consecuencia de la local convexidad de los espacios Euclidianos; o en otras palabras, en cada punto $x \in U$ en un abierto U de un espacio euclideano, hay una bola dentro de U que lo contiene: $x \in B(x, r) \subseteq U$.

Observación A.1.3. Podemos representar la transformación lineal (A.1) mediante una matriz. Cuando $E = \mathbb{R}^d$ y la base fijada es la base canónica, la representación matricial de (A.1) es la matriz Jacobiana de $\bar{F}$.

Debemos definir ahora el diferencial para funciones suaves $F : M \to M'$. Sin embargo, aquí la Definición A.1.5 puede no hacer sentido. En efecto, si la variedad es «curvada» en el punto x, no habrán direcciones donde exista $t > 0$ tal que $[x, x + tv] \subseteq M$. El diferencial de funciones en variedades se definirá, entonces, usando curvas suaves en la variedad.

Definición A.1.6 (Diferencial en variedades incrustadas)**.** Sea $F : M \to M'$ una función suave entre variedades incrustadas. Definimos el diferencial de F en el punto $x \in M$ como la transformación lineal $DF(x) : T_x M \to T_{F(x)} M'$ definido por

$$DF(x)[v] = \left. \frac{\mathrm{d}}{\mathrm{d}t} F(c(t)) \right|_{t=0}, \tag{A.2}$$

donde c es una curva suave en M pasando por x en $t = 0$ a velocidad v.

El siguiente teorema nos dice cómo se relacionan los diferenciales Euclidianos y de variedades incrustadas. También nos dice que la definición anterior es independiente de la curva c y que el diferencial es, efectivamente, una transformación lineal.

Teorema A.1.3. Sean M y M' variedades incrustadas en E y E', respectivamente. Sea $F : M \to M'$ una función suave y sea $\bar{F} : U \subseteq E \to E'$ una extensión suave de F. Luego

$$DF(x) = D\bar{F}(x)\big|_{T_x M}.$$

Demostración. Sea $c : I \to M$ una curva suave que pasa por x en $t = 0$ a velocidad v; esto es,

$c(0) = x$ y $c'(0) = v$. Luego

$$\begin{aligned} DF(x)[v] &= \frac{\mathrm{d}}{\mathrm{d}t} F(c(t))\Big|_{t=0} \\ &= \frac{\mathrm{d}}{\mathrm{d}t} \bar{F}(c(t))\Big|_{t=0} \\ &= D\bar{F}(c(0))[c'(0)] \\ &= D\bar{F}(x)[v] \end{aligned}$$

■

En un espacio euclideano, el gradiente de una función suave $\bar{f} : U \subseteq E \to \mathbb{R}$ en un punto $x \in U$, denotado por $\operatorname{grad} \bar{f}(x)$, es la representación vectorial (garantizada por el Teorema de representación de Riesz) de la transformación lineal $D\bar{f}(x) : E \to E'$. La representación vectorial viene determinada por el producto interno con el que se encuentra dotado el espacio. Por ello, antes de poder introducir el gradiente, necesitaremos hablar de productos internos en el espacio tangente, el cual permite hablar de distancias entre puntos de la variedad y, entre otras cosas, nos permite definir el gradiente de una función suave.

Definición A.1.7.

- Un producto interno en T_xM es una función bilineal, simétrica y definida positiva $\langle \cdot, \cdot \rangle_x : T_xM \times T_xM \to \mathbb{R}$. Nos referimos a $\langle \cdot, \cdot \rangle_x$ como una métrica en M.
- Una métrica $\langle \cdot, \cdot \rangle_x$ en M es una métrica riemanniana si varía suavemente con x.
- Una variedad con una métrica riemanniana es una variedad riemanniana.

Los espacios tangentes T_xM de una variedad incrustada M son subespacios vectoriales del espacio ambiente E. Por ende, una manera de elegir un producto interno $\langle \cdot, \cdot \rangle_x$ para cada T_xM es simplemente restringiendo el producto interno $\langle \cdot, \cdot \rangle$ de E a cada subespacio T_xM.

Teorema A.1.4. Sea M una variedad incrustada de E y sea $\langle \cdot, \cdot \rangle$ la métrica Euclidiana en E. La métrica en M definida en cada x mediante la restricción de $\langle \cdot, \cdot \rangle$ al espacio T_xM; o sea,

$$\langle u, v \rangle_x = \langle u, v \rangle \qquad \forall u, v \in T_xM,$$

es una métrica riemanniana.

Definición A.1.8. Sea M una variedad incrustada en E. Cuando la métrica en M es la restricción de la métrica de E, decimos que M es una subvariedad riemanniana de E.

Por fin estamos en posición de definir el gradiente de una función suave $f : M \to \mathbb{R}$.

Definición A.1.9. Sea $f : M \to \mathbb{R}$ una función suave en una variedad riemanniana incrustada M. El gradiente Riemanniano de f, ∇f, queda definido por las siguientes identidades:

- $\nabla f(x) \in T_xM$.

- Para cada $x \in M$ se tiene que

$$Df(x)[v] = \langle v, \nabla f(x) \rangle_x \qquad \forall v \in T_x M.$$

El siguiente teorema nos dice que cuando M es una subvariedad riemanniana de E, podemos calcular $\nabla f(x)$ mediante grad $\bar{f}(x)$, donde $\bar{f}$ es una extensión suave de f. Para esto es bueno recordar que todo espacio vectorial finito-dimensional puede descomponerse mediante un subespacio vectorial y su complemento ortogonal; o sea,

$$E = T_x M \oplus T_x M^{\perp}.$$

Esto es lo mismo que decir que todo $w \in E$ se puede escribir de forma única como

$$w = v + u,$$

donde $v \in T_x M$ y $u \in T_x M^{\perp}$.

Podemos definir la proyección ortogonal, P_x, sobre $T_x M$ como aquel operador lineal que asigna a cada elemento w su parte correspondiente a $T_x M$. En el caso anterior,

$$P_x(w) = v.$$

Teorema A.1.5. Sea M una subvariedad riemanniana de E y sea $f : M \to \mathbb{R}$ una función suave. El gradiente Riemanniano de f viene dado por

$$\nabla f(x) = P_x \left(\text{grad } \bar{f}(x) \right),$$

donde $\bar{f}$ es cualquier extensión suave de f.

Ejemplo A.1.1. Sea $M = \mathbb{S}^{d-1} := \left\{ x \in \mathbb{R}^d : |x| = 1 \right\} \subset \mathbb{R}^d$.

- $\mathbb{S}^{d-1}$ es una variedad suave incrustada en $\mathbb{R}^d$. En efecto, usando la función (que es claramente suave) $h : \mathbb{R}^d \setminus \{0\} \to \mathbb{R}$ definida por

$$h(x) = x_1^2 + \ldots + x_d^2 - 1,$$

  tenemos que

  - $\mathbb{S}^{d-1} = h^{-1}(\{0\})$.
  - $\text{rank} Dh(x) = \text{rank} \left[2x_1 \ldots 2x_d \right] = 1$ para todo $x \in \mathbb{R}^d \setminus \{0\}$.

  Por lo que se satisface la segunda condición de la definición A.1.1.

- Para cada $x \in \mathbb{S}^{d-1}$

$$T_x \mathbb{S}^{d-1} = \left\{ z \in \mathbb{R}^d : x^{\top} z = 0 \right\}.$$

  Esto, pues, por Teorema A.1.1, $T_x \mathbb{S}^{d-1} = \ker \left[2x_1 \ldots 2x_d \right]$. Notar que, geométricamente, este es el hiperplano tangente a la esfera en el punto x.

- Claramente $\dim \mathbb{S}^{d-1} = d-1$ (su espacio tangente en cada punto es un hiperplano).
- Dado $f \in \mathcal{C}^\infty(\mathbb{S}^{d-1})$, el gradiente Riemanniano de f en x, $\nabla f(x)$, se puede calcular mediante una extensión suave, $\bar{f} : U \subseteq \mathbb{R}^d \to \mathbb{R}$, donde U es una vecindad de $\mathbb{S}^{d-1}$ en $\mathbb{R}^d$, usando la regla

$$\begin{aligned}\nabla f(x) &= P_x(\operatorname{grad} \bar{f}(x)) \\ &= (I - xx^\top) \operatorname{grad} \bar{f}(x).\end{aligned}$$

  Esto es una instanciación del Teorema A.1.5 junto con el hecho de que la proyección ortogonal al espacio $\left\{z \in \mathbb{R}^d : x^\top z = 0\right\}$ es

$$P_x(w) = (I - xx^\top)w,$$

  para todo $w \in \mathbb{R}^d$.

A.2. Hechos y definiciones misceláneas

Definición A.2.1 (Operador de Laplace-Beltrami)**.** Sea $(M, \mathfrak{g})$ una variedad Riemanniana. Definimos el operador de Laplace-Beltrami $\Delta_M : \mathcal{C}^\infty(M) \to \mathcal{C}^\infty(M)$ (en coordenadas locales) como

$$\Delta_M(f) = \frac{1}{\sqrt{\det(\mathfrak{g})_{kl}}} \frac{\partial}{\partial x^i} \left(\mathfrak{g}^{ij} \sqrt{\det(\mathfrak{g})_{kl}} \frac{\partial f}{\partial x^j} \right).$$

Si bien la definición anterior puede lucir complicada, es un hecho conocido que el operador de Laplace-Beltrami tiene una forma simplificada en la esfera.

Propiedad A.2.1 (Laplace-Beltrami en la esfera)**.** Sea $f \in \mathcal{C}^\infty(M)$ y sea $\bar{f}\mathbb{R}^d \to \mathbb{R}$ una extensión suave de f definida por

$$\bar{f}(x) = f\left(\frac{x}{|x|}\right).$$

Luego

$$\Delta_{\mathbb{S}^{d-1}}(f) = \Delta\left(\bar{f}\left(\frac{x}{|x|}\right)\right).$$

Las siguientes propiedades enlazan al operador de Laplace-Beltrami con el operador L definido en los preliminares:

Propiedad A.2.2. Sea $(M, \mathfrak{g})$ una variedad suave y compacta, $f_1, \ldots, f_k, f, g \in \mathcal{C}^\infty(M)$ y $\Psi : U \subseteq \mathbb{R}^d \to \mathbb{R}$ una función suave. Luego

(a)

$$\Delta_M\left(\Psi(f_1, \ldots, f_k) = \sum_{i=1}^k \partial_i \Psi\left(f_1, \ldots, f_k\right) \Delta_M(f_i) + \sum_{i,j=1}^k \partial_{i,j}^2 \Psi\left(f_1, \ldots, f_k\right) \nabla f_i \cdot \nabla f_j.\right.$$

(b)

$$\nabla\left(\Psi(f_1, \ldots, f_k)\right) \cdot \nabla g = \sum_{i=1}^k \partial_i \Psi(f_1, \ldots, f_k) \nabla f_i \cdot \nabla g.$$

(c)
$$\int_M \Delta_M(f)dx = 0.$$

(d) Fórmula de integración por partes
$$\int_M g\Delta_M(f)dx = -\int_M \nabla g \cdot \nabla f dx.$$

El siguiente teorema –conocido como fórmula de Bochner-Lichnerowicz– nos será útil en el cálculo del operador Γ_2:

Teorema A.2.1 (Fórmula de Bochner-Lichnerowicz [2])**.** Para toda $f \in \mathcal{C}^\infty(M)$,
$$\frac{1}{2}\Delta_M\left(|\nabla f|^2\right) = \nabla f \cdot \nabla\left(\Delta_M f\right) + |\nabla\nabla f|^2 + \mathrm{Ric}_{\mathfrak{g}}\left(\nabla f, \nabla f\right).$$

Bibliografía

[1] Asi, H., Feldman, V., Koren, T., and Talwar, K. (2021). Private stochastic convex optimization: Optimal rates in l1 geometry. In *International Conference on Machine Learning*, pages 393–403. PMLR.

[2] Bakry, D., Gentil, I., Ledoux, M., et al. (2014). *Analysis and geometry of Markov diffusion operators*, volume 103. Springer.

[3] Bassily, R., Feldman, V., Talwar, K., and Guha Thakurta, A. (2019). Private stochastic convex optimization with optimal rates. *Advances in neural information processing systems*, 32.

[4] Bassily, R., Guzmán, C., and Nandi, A. (2021). Non-euclidean differentially private stochastic convex optimization. In *Conference on Learning Theory*, pages 474–499. PMLR.

[5] Bassily, R., Guzmán, C., and Nandi, A. (2022). Non-euclidean differentially private stochastic convex optimization: Optimal rates in linear time.

[6] Bassily, R., Smith, A., and Thakurta, A. (2014). Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th annual symposium on foundations of computer science*, pages 464–473. IEEE.

[7] Boumal, N. (2020). An introduction to optimization on smooth manifolds. *Available online, May*, 3.

[8] Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

[9] Cover, T. M. and Thomas, J. A. (2006). Elements of information theory 2nd edition (wiley series in telecommunications and signal processing).

[10] Dwork, C., Roth, A., et al. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407.

[11] Fazel, M. (2002). *Matrix rank minimization with applications*. PhD thesis, PhD thesis, Stanford University.

[12] Ge, R., Lee, H., Lu, J., and Risteski, A. (2021). Efficient sampling from the bingham distribution. In *Algorithmic Learning Theory*, pages 673–685. PMLR.

[13] Jaggi, M. (2013). Revisiting frank-wolfe: Projection-free sparse convex optimization. In *International Conference on Machine Learning*, pages 427–435. PMLR.

[14] Jain, P., Thakkar, O. D., and Thakurta, A. (2018). Differentially private matrix completion revisited. In *International Conference on Machine Learning*, pages 2215–2224. PMLR.

[15] Kapralov, M. and Talwar, K. (2013). On differentially private low rank approximation. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*, pages 1395–1414. SIAM.

[16] Kume, A. and Walker, S. G. (2014). On the bingham distribution with large dimension. *Journal of Multivariate Analysis*, 124:345–352.

[17] McSherry, F. and Talwar, K. (2007). Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE.

[18] Minami, K., Arai, H., Sato, I., and Nakagawa, H. (2016). Differential privacy without sensitivity. *Advances in Neural Information Processing Systems*, 29.

[19] Narayanan, A. and Shmatikov, V. (2006). How to break anonymity of the netflix prize dataset. *arXiv preprint cs/0610105*.

[20] Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press.

[21] Zhang, M., Shen, Z., Mokhtari, A., Hassani, H., and Karbasi, A. (2020). One sample stochastic frank-wolfe. In *International Conference on Artificial Intelligence and Statistics*, pages 4012–4023. PMLR.