PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
ESCUELA DE INGENIERIA

COTAS DE RIESGO EN OPTIMIZACIÓN CONVEXA ESTOCÁSTICA MEDIANTE ESTIMACIÓN DEL DESEMPEÑO COMPUTACIONAL DE PEOR CASO

PATRICIO ULLOA BALDASSARE

Tesis presentada a la Dirección de Investigación y Postgrado
como parte de los requisitos para optar al grado de
Magíster en Ciencias de la Ingeniería

Profesor Supervisor:
CRISTÓBAL GUZMÁN PAREDES

Santiago de Chile, Septiembre 2021

PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
ESCUELA DE INGENIERIA

COTAS DE RIESGO EN OPTIMIZACIÓN CONVEXA ESTOCÁSTICA MEDIANTE ESTIMACIÓN DEL DESEMPEÑO COMPUTACIONAL DE PEOR CASO

PATRICIO ULLOA BALDASSARE

Miembros del Comité:
CRISTÓBAL GUZMÁN PAREDES
CARLOS SING-LONG COLLAO
MARIO BRAVO GONZÁLEZ
CLAUDIO MOURGUES ÁLVAREZ

DocuSigned by: Cristóbal Guzmán 7EE5729B2A224CF...
DocuSigned by: Carlos Sing Long 3574ABD6A96E44F...
DocuSigned by: Mario Bravo 16C9C95946B1454...
DocuSigned by: Claudio Mourgues 1F399B17887E48D...

Tesis presentada a la Dirección de Investigación y Postgrado
como parte de los requisitos para optar al grado de
Magíster en Ciencias de la Ingeniería

Santiago de Chile, Septiembre 2021

A mi familia y amigos

AGRADECIMIENTOS

Primero, quisiera agradecer a mi familia y amigos por su apoyo constante e incondicional. Gracias por acompañarme en todos mis proyectos y motivarme a seguir adelante. Gracias por escucharme, por su confianza, su cariño y por inspirarme a ser mejor cada día.

También agradecerle a mi profesor supervisor, Cristóbal Guzmán, por presentarme la oportunidad de hacer esta investigación. Gracias por tu atenta disposición, por las enseñanzas y por guiarme en el desarrollo de la tesis. Eres un profesor ejemplar, cuyo apoyo y *expertise* me permitieron aprovechar al máximo esta experiencia.

Además, darle las gracias a la comunidad del IMC por su cercanía y amabilidad. También destacar el esfuerzo colectivo por promover la participación de los estudiantes dentro y fuera de clases, e involucrarlos en el desarrollo de los programas. Ambas razones han permitido generar una (pequeña) gran comunidad y un excelente ambiente para promover la investigación.

Por último, agradecer el apoyo financiero a esta investigación, recibido a través del proyecto FONDECYT Iniciación 11160939.

INDICE GENERAL

INDICE DE FIGURAS

INDICE DE TABLAS

RESUMEN

Se considera el problema de optimización convexa estocástica, que permite una representación general de una diversa gama de aplicaciones en aprendizaje automático, estadística e investigación de operaciones, entre otros. En el estudio de tal problema, se analiza la complejidad de generalizar el conocimiento obtenido a través de una muestra de datos con comportamiento desconocido, minimizando un problema convexo definido por la muestra. Debido a no realizar supuestos sobre tales datos, los resultados comunes entregan cotas asintóticas que garantizan órdenes de convergencia de los errores inducidos por aproximaciones al óptimo a través de métodos de primer orden, descartando una cuantificación exacta del peor caso alcanzable en la práctica. En este trabajo, se plantea la hipótesis que es posible recuperar cotas de generalización ayudándose de un problema computacional que calcula el rendimiento de peor caso de tanto el error de optimización como la estabilidad algorítmica de un método de primer orden. Se proponen problemas semidefinidos que permiten la representación exacta de ambas métricas, para distintos tipos de métodos. Luego, basándose en una implementación de los modelos desarrollados, se procura inferir expresiones simbólicas de los resultados obtenidos a través de un proceso heurístico, recuperando información que puede traducirse en demostraciones de garantías de generalización. Se encontró una representación matricial de dos tipos de métodos de primer orden, aprovechando la estructura de las actualizaciones. También se realizó el proceso heurístico para el método de punto proximal, encontrando casos donde es posible su correcta utilización y otros donde aparecen limitaciones prácticas que imposibilitan producir una demostración rigurosa.

Palabras Claves: Optimización convexa estocástica, Minimización de riesgo empírico, Exceso de riesgo, Estabilidad algorítmica, Problema de estimación de rendimiento, Optimización semidefinida, Métodos de primer orden, Análisis asistido por computador.

ABSTRACT

We consider the stochastic optimization problem, which allows a general representation of a diverse range of applications in machine learning, statistics and operations research, among others. To study this problem, we analyze the complexity of generalizing knowledge obtained through a data sample, whose behaviour is unknown, by minimizing a sample-defined convex problem. Due to not imposing assumptions on such data, common findings yield asymptotic bounds which guarantee a certain convergence rate of errors induced by approximation to optima through first-order methods, dismissing an exact cuantification of the precise worst-case attained in practice. In this thesis, we claim it is possible to retrieve generalization bounds aided by a computational problem, which computes a first-order method's worst-case performance of both optimization error and algorithmic stability. We propose semidefinite programs which allow the exact worst-case representation of such measures, for different types of methods. Then, based on the implementation of the developed models, we propose deriving symbolic expressions from the achieved results through a heuristic process, retrieving information which can be reformulated into a generalization guarantee proof. We found a matrix representation for two distinct types of first-order methods, taking advantage of the update's structure. Besides, we applied the heuristic process to the proximal point method, obtaining both cases where it may be used correcly and where practical limitations arise, which preclude us from deriving a rigorous proof.

Keywords: Stochastic convex optimization, Empirical risk minimization, Excess risk, Algorithmic stability, Performance estimation problem, Semidefinite optimization, First-order methods, Computer-aided analysis.

1. INTRODUCCIÓN

El problema de optimización convexa estocástica (SCO) es un modelo de gran importancia en Estadística y *Machine Learning*. Este programa permite representar diversos modelos de regresión y de aprendizaje supervisado, garantizando la aprendibilidad tras imponer supuestos de convexidad de la clase de hipótesis y nociones de regularidad de la pérdida, restricciones que son naturales a estas aplicaciones (Shalev-Shwartz y Ben-David, 2014).

Así, un problema SCO (1.1) se caracteriza a través de una función de pérdida f sobre un espacio producto $\mathcal{X} \times \mathcal{Z}$, que se desea optimizar en un conjunto factible $\mathcal{X}$ convexo en el espacio euclídeo $(\mathbb{R}^d, ||\cdot||)$ y una distribución desconocida $\mathcal{D}$ sobre $\mathcal{Z}$. Además, las pérdidas $f(\cdot, \xi)$ son convexas para todo $\xi \in \mathcal{Z}$. En general, se suponen características adicionales de regularidad en la primera componente de f, tales como Lipschitz-continuidad, suavidad y/o convexidad fuerte. Se asume que la distribución anterior solo puede ser accedida mediante el muestreo de n datos aleatorios i.i.d. $\{\boldsymbol{\xi_i}\}_{i\in[n]}$. Se llama Riesgo de población a la función $F(x) := \mathop{\mathbb{E}}_{\boldsymbol{\xi}\sim\mathcal{D}}[f(x, \boldsymbol{\xi})]$, objetivo a minimizar en el problema (1.1).

$$\min_{x\in\mathcal{X}} F(x) \tag{1.1}$$

Debido a la imposibilidad de conocer $\mathcal{D}$ en la práctica y consecuentemente de resolver (1.1), el acercamiento común al problema anterior consiste en la utilización del funcional de Riesgo empírico (1.2) definido por la muestra $\mathbf{S} = (\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_n)$ como una aproximación del valor esperado en (1.1). Utilizando información de primer orden de las pérdidas $f(\cdot, \xi)$ (dado $\xi \in \mathcal{Z}$) entregada por algún oráculo de primer orden, se requiere derivar aproximaciones a la respuesta óptima de (1.1) mediante soluciones aproximadas al óptimo del problema de minimización de riesgo empírico (1.3).

$$F_{\mathbf{S}}(x) := \frac{1}{n}\sum_{\boldsymbol{\xi}\in\mathbf{S}} f(x, \boldsymbol{\xi}) \tag{1.2}$$

$$\min_{x\in\mathcal{X}} F_{\mathbf{S}}(x) \tag{1.3}$$

Una aplicación de los problemas de aprendizaje (1.1) y (1.3) es el problema de regresión lineal, donde se busca predecir los parámetros de dependencia lineal de una variable respuesta $\boldsymbol{b}$ con respecto a variables predictoras $\boldsymbol{a}$ en la bola unitaria, dada una muestra de ejemplos aleatorios de la forma $\boldsymbol{\xi}_i = (\boldsymbol{a}_i, \boldsymbol{b}_i) \in \mathcal{X} \times \mathbb{R}$. Se considera la clase de hipótesis $\mathcal{X} = \mathcal{B}_d(0,1)$ y la función de pérdida cuadrática $f(x, \boldsymbol{\xi}) = \frac{1}{2}(\langle x, \boldsymbol{a}\rangle - \boldsymbol{b})^2$. Shalev-Shwartz y Ben-David (2014) muestran que utilizando lo anterior, el problema de minimización de riesgo empírico (1.3) se puede reescribir como un problema de mínimos cuadrados sobre la muestra $\mathbf{S}$ con función objetivo (1.4), teniendo por óptimo a algún parámetro en $\{x \in \mathbb{R}^d : ||x|| \leq 1\}$ que minimiza el error cuadrático medio a los datos. Asimismo, el problema SCO se puede reescribir como aquel que minimiza, en valor esperado, la pérdida cuadrática de un nuevo dato a la predicción, como se expresa en (1.5).

$$F_{\mathbf{S}}(x) = \frac{1}{2n} \sum_{i \in [n]} \left(\langle x, \boldsymbol{a}_i\rangle - \boldsymbol{b}_i\right)^2 \tag{1.4}$$

$$\min_{||x|| \leq 1} \mathop{\mathbb{E}}_{(\boldsymbol{a},\boldsymbol{b}) \sim \mathcal{D}} \left[\frac{1}{2}(\langle x, \boldsymbol{a}\rangle - \boldsymbol{b})^2\right] \tag{1.5}$$

Sin embargo, en (1.5) y en el problema SCO general (1.1), poder minimizar sobre la muestra conocida no es *per se* una garantía de una buena generalización a toda la población proveniente de $\mathcal{D}$. En general, se requieren supuestos adicionales sobre el método de descenso utilizado y/o las propiedades de la función de pérdida para poder lograr garantías fuertes sobre el problema (1.1). Una descomposición común en la literatura de aprendizaje de máquinas, es aquella donde se analizan los fenómenos de generalización, optimización y aproximación separadamente (Shalev-Shwartz y Ben-David, 2014), balanceando la complejidad atribuible a cada uno de estos errores mediante la elección de parámetros convenientes. Para el problema de generalización resultante, una técnica ampliamente utilizada es relacionar esta capacidad de aprendizaje sobre la población completa a través de propiedades de estabilidad algorítmica de los métodos utilizados. Esta propiedad de los algoritmos de aprendizaje es una cuantificación (según distintas nociones) del máximo de pérdida o de distancia entre trayectorias generadas por un algoritmo cuando se altera marginalmente la muestra provista.

Las garantías teóricas convencionales en SCO entregan órdenes de complejidad (medido, por ejemplo, en tiempo o evaluaciones de gradiente) para los errores de optimización y estabilidad de algoritmos de primer orden en términos de parámetros del problema, como aquellas presentes en Hardt, Recht, y Singer (2016) y Bassily, Feldman, Guzmán, y Talwar (2020). Estos resultados derivan cotas de complejidad asintóticas superiores o inferiores, alcanzando en algunos casos el mismo orden de complejidad entre ambos límites. En cualquier caso, tener ambas cotas para la complejidad medida no necesariamente permite obtener una representación del peor caso real de manera exacta.

Por otra parte, el uso del problema de estimación de rendimiento (PEP, por su nombre en Inglés) asistido por computador ha facilitado la representación de la complejidad de primer orden de métodos de descenso en optimización convexa determinista de manera exacta (Taylor, Hendrickx, y Glineur, 2017b), permitiendo extensiones incluso al caso no convexo (Taylor, Hendrickx, y Glineur, 2017a) y estocástico (Taylor y Bach, 2019).

1.1. Objetivos de investigación

En este trabajo, se planteó la hipótesis de que es posible obtener cotas no asintóticas de optimización y estabilidad para los problemas (1.1) y (1.3) utilizando representaciones semidefinidas positivas (SDP) de PEP. Estos problemas modelarían de manera exacta el peor caso de una ejecución de métodos de primer orden, permitiendo además medir la uniforme estabilidad.

De acuerdo con la hipótesis anterior, se propuso el objetivo general de aplicar la metodología de trabajo para PEP a modelos que representan explícitamente los peores casos de optimización y/o estabilidad en el contexto de minimización de riesgo empírico. En vista de tal objetivo, se plantearon los siguientes tres objetivos de investigación específicos, correspondientes a tres pasos distintos de aplicación de la metodología asistida por computador:

1. Desarrollar versiones semidefinidas de PEP aplicables a los métodos estudiados en el contexto de SCO, incorporando en la estructura de tales problemas representaciones exactas de métricas para la estabilidad y optimización.
2. Usando la implementación computacional los modelos semidefinidos propuestos, establecer comparaciones entre peor caso real y cotas de la teoría.
3. Obtención de garantías teóricas basadas en los resultados computacionales logrados, utilizando la metodología heurística para PEP.

Adicionalmente, se buscaron garantías teóricas complementarias a los antecedentes del problema sin hacer un análisis asistido computacionalmente, de manera de complementar resultados teóricos previos.

1.2. Metodología

Se modelan problemas de optimización tipo PEP para representar el rendimiento en el peor caso de los problemas de minimización de riesgo empírico y/o estabilidad de manera exacta. A continuación se describe una secuencia de los pasos a seguir, de acuerdo con los objetivos planteados:

1. Se plantea un PEP que simula dos trayectorias de un método de primer orden generadas por muestras vecinas, utilizando la distancia en norma entre trayectorias vecinas como métrica de la estabilidad uniforme y el *gap* de optimalidad del riesgo empírico como métrica del error de optimización, en función de la respuesta de cada método. Además se muestra que tales PEP pueden ser reescritos como problemas de optimización semidefinida para las elecciones de método de descenso, métrica y restricciones planteadas; representando estas expresiones en la estructura del problema semidefinido.
2. Se calcula el dual asociado a la versión semidefinida de cada PEP, probando que se tiene dualidad fuerte en todo caso de interés mediante un certificado de la condición de Slater en el problema dual.
3. Se implementan los modelos semidefinidos anteriores utilizando la librería de optimización `mosek.fusion` para el lenguaje de programación Python 3.7. Esta librería permite la resolución de problemas semidefinidos mediante ejecuciones del *solver* de optimización MOSEK versión 9.2.47, basado en el método de punto interior. Se implementa el problema primal, se ejecuta y compara el ajuste del cálculo de peor caso con cotas teóricas asintóticas conocidas de la literatura y/o probadas en este trabajo.
4. Se abstraen resultados teóricos de algunos casos utilizando la metodología PEP de Taylor et al. (2017b) y Taylor et al. (2017a). Para esto se implementa el problema dual, ejecutando y añadiendo restricciones sucesivamente, de manera heurística, a modo de simplificar los valores asignados a las variables duales por la ejecución del *solver* para problemas semidefinidos.

5. Una vez que se obtienen expresiones simbólicas factibles para todas las variables duales, se ponderan con las restricciones primales respectivas para generar un esquema de demostración de cota superior, el que puede ser formalizado en una demostración de cota superior basada en desigualdades de convexidad. Se nota que tales asignaciones simbólicas exactas para las variables duales pueden ser subóptimas o difíciles de obtener. En su defecto, se analizan tales dificultades y se conjetura si la cota superior puede ser mejorada mediante este proceso.

1.3. Contribuciones

En esta investigación se aplica la metodología de trabajo para el problema de estimación de rendimiento en la obtención de garantías de convergencia en SCO, cuantificando exactamente métricas de los errores inducidos por la optimización del riesgo empírico y la estabilidad uniforme en sus peores casos. En particular, se desarrollan modelos semidefinidos equivalentes al peor caso real de aplicaciones de optimizar el riesgo empírico y estabilidad uniforme de argumentos mediante métodos *batch*, incluyendo representaciones separadas y conjuntas de ambas fuentes de error. Además, se desarrollan modelos semidefinidos equivalentes al problema de peor caso de aplicar métodos incrementales, condicionados por una conjetura.

Luego, se obtienen demostraciones a través de un proceso heurístico, permitiendo recuperar una garantía de cota superior para el error de optimización del PPM que alcanza exactamente el peor caso obtenido en la implementación computacional de PEP. Por último, se provee una idea de demostración para una garantía similar para la estabilidad uniforme, junto a evidencia computacional que soporta conjeturas sobre los pasos faltantes en la elaboración de una demostración rigurosa.

Por otra parte, se complementa la literatura de SCO con resultados que acotan superiormente los errores asociados a la optimización y la estabilidad uniforme de argumentos de los métodos expuestos. En específico, se presenta una garantía del error de optimización

para el método *IncrementalProx* que mejora la aplicación directa del análisis para métodos de gradiente estocástico proximal incremental de (Bertsekas, 2011); cotas superiores para la estabilidad del método SVRG de (Johnson y Zhang, 2013) bajo pérdidas suaves y fuertemente convexas, y para la estabilidad de métodos *minibatch*-Prox generalizados, bajo pérdidas Lipschitz-continuas.

1.4. Organización de la tesis

A continuación, se describe brevemente el contenido de cada uno de los capítulos siguientes.

En el Capítulo 2 se presentan los antecedentes teóricos de SCO, estabilidad algorítmica y el problema de estimación de rendimiento.

Los resultados principales obtenidos se exponen en los Capítulos 3-6. En el Capítulo 3 se muestra un resultado de estabilidad algorítmica para métodos proximales generalizados y en el Capítulo 4, para el método SVRG. En los Capítulos 5 y 6 se introduce el desarrollo de modelos de PEP incluyendo el fenómeno de estabilidad y los resultados de su posterior implementación; para métodos *batch* e incrementales, respectivamente.

El Capítulo 7 presenta una discusión de los resultados presentados en los cuatro capítulos anteriores, incluyendo posibles líneas de trabajo derivadas de ésta. El Capítulo 8 expone las conclusiones a las que se llegaron en cuanto a la pregunta de investigación y objetivos de trabajo planteados.

Además, se exponen resultados complementarios y demostraciones faltantes de SVRG en el Anexo A, de PEP para métodos *batch* en el Anexo B y para métodos incrementales en el Anexo C.

1.5. Resumen de antecedentes

Este trabajo reune antecedentes de tres áreas principales, que se describen brevemente a continuación:

Convergencia del error de optimización de ERM: Los análisis de convergencia convencionales están basados en resultados clásicos de peor caso para optimización convexa determinista en el caso *batch*, incluyendo cotas superiores para el método de punto proximal (Güler, 1991), para el método de subgradiente (Shor, 2012) y de gradiente, como se presenta en Beck (2017).

Adicionalmente, se utilizan resultados de convergencia para el método de punto proximal incremental (Bertsekas, 2011), SGD incremental (Nedic y Bertsekas, 2001; Bassily et al., 2020), SGD con promedio de iterados (Polyak y Juditsky, 1992) y SVRG, una versión de SGD acelerada mediante reducción de varianza (Johnson y Zhang, 2013). Estos métodos aleatorizados se basan en la utilización de un solo dato de la muestra en cada iteración y siguen uno de dos posibles esquemas de utilización de datos: mediante permutaciones o muestreo con reemplazo. Los primeros se basan en la minimización de funciones a través de (posiblemente múltiples) rondas de pasadas secuenciales por los datos, como en Nedic y Bertsekas (2001). Por otra parte, la actualización de los segundos se puede interpretar como una aproximación estocástica de gradientes de la función subyacente que se desea optimizar, obteniendo resultados de convergencia basados en garantías de aproximación estocástica (Robbins y Monro, 1951; Nemirovski y Yudin, 1978).

Generalización mediante Estabilidad Algorítmica: Basado en el trabajo de Bousquet y Elisseeff (2002), se utilizan resultados recientes de generalización apoyados en estabilidad uniforme de Bousquet, Klochkov, y Zhivotovskiy (2020) y del error de aproximación de Bassily et al. (2020). Además, se utilizan análisis de uniforme estabilidad para los métodos de aprendizaje antes mencionados: Para el caso suave (Hardt et al., 2016) y no-suave (Bassily et al., 2020) de métodos con oráculo tipo gradiente y para métodos proximales, inspirados en las 2 publicaciones anteriores, resultados de la teoría de análisis convexo de Rockafellar (1976) y el trabajo de Bassily, Feldman, Talwar, y Thakurta (2019) para algoritmos *minibatch* con muestreo-con-reemplazo.

Análisis asistido por computador: El problema de estimación de rendimiento fue planteado inicialmente por Drori y Teboulle (2014), pero este trabajo se basa en una representación semidefinida exacta del problema de peor caso lograda por Taylor et al. (2017b) y su posterior generalización para métodos proximales y múltiples funciones en Taylor et al. (2017a). Para el problema estocástico, Taylor y Bach (2019) utilizan potenciales para obtener cotas sobre el exceso de riesgo para métodos con acceso a oráculos ruidosos. Las tres publicaciones mencionadas proveen una metodología para la obtención de resultados asistidos por computador, empleada en esta investigación.

2. MARCO TEÓRICO

A continuación, se detallan los supuestos para los problemas (1.1) y (1.3) presentados en la sección anterior. Se considera el espacio Euclídeo $(\mathbb{R}^d, ||\cdot||)$ y la región factible de (1.1) como un conjunto $\mathcal{X} \subset \mathbb{R}^d$ convexo, cerrado y no-vacío.

Dado $\mathcal{X} \subset \mathbb{R}^d$ convexo, cerrado y no-vacío y una función convexa $g : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$, esta función es M-Lipschitz si

$$|g(x) - g(y)| \leq M||x - y|| \quad (\forall x, y \in \mathcal{X}). \tag{2.1}$$

Adicionalmente, g es una función L-suave si cumple

$$||\nabla g(x) - \nabla g(y)|| \leq L||x - y|| \quad (\forall x, y \in \mathcal{X}). \tag{2.2}$$

Finalmente, se dice que g es una función μ-fuertemente convexa si cumple

$$g(y) \geq g(x) + \langle v, y - x\rangle + \frac{\mu}{2}||x - y||^2 \quad (\forall x, y \in \mathcal{X}; \forall v \in \partial g(x)). \tag{2.3}$$

Se denota la clase de funciones M-Lipschitz sobre $\mathcal{X}$ como $\mathcal{F}_M(\mathcal{X})$ y a la clase de funciones L-suaves y M-Lipschitz como $\mathcal{F}_{M,L}(\mathcal{X})$. Consideramos las generalizaciones respectivas de tales clases a dimensiones finitas como $\mathcal{F}_M := \bigcup_{d \geq d_0} \mathcal{F}_M(\mathbb{R}^d)$ y $\mathcal{F}_{M,L} := \bigcup_{d \geq d_0} \mathcal{F}_{M,L}(\mathbb{R}^d)$ para una cantidad d_0 que se especificará caso a caso. Además, se denota por $\mathcal{S}^{\mu}_{M,L}(\mathcal{X})$ a la clase de funciones M-Lipschitz, L-suaves y μ-fuertemente convexas.

En el caso de una función g no necesariamente diferenciable, se denota $\nabla g(x)$ como una elección arbitraria de subgradiente en el subdiferencial $\partial g(x)$. Por otra parte, en el contexto de optimización convexa determinística se denota por x_* al minimizador de una función si no existe ambigüedad respecto a ésta. En cambio, en el contexto de SCO, un minimizador de (1.1) se denota por x^* y un minimizador de (1.3) sobre una muestra aleatoria $\mathbf{S} \sim \mathcal{D}^n$ por $x^*_{\mathbf{S}}$.

Para problemas con regularización simple, se considera una pérdida f Lipschitz y suave. Se nota la pérdida μ-regularizada f_μ, definida por

$$f_\mu(\cdot, \xi) \equiv f(\cdot, \xi) + \frac{\mu}{2}\psi(\cdot),$$

donde ψ es un regularizador 1-fuertemente convexo. El riesgo y riesgo empírico asociados a la nueva pérdida se denotan por F^μ y F_S^μ, respectivamente.

En adelante, al tratar los problemas (1.1) y (1.3) se especificará que se está en el caso no-suave cuando las pérdidas son M-Lipschitz continuas, en el caso suave cuando las pérdidas además son L-suaves y en el caso fuertemente convexo si además son μ-fuertemente convexas. Los Supuestos 1a, 1b y 1c formalizan lo anterior.

Supuesto 1a. *Las pérdidas $f(\cdot, \xi)$ cumplen la condición de Lipschitz continuidad* (2.1) *para todo dato $\xi \in \mathcal{Z}$. Esto es,*

$$f(\cdot, \xi) \in \mathcal{F}_M(\mathcal{X}) \quad (\forall \xi \in \mathcal{Z}).$$

Supuesto 1b. *Las pérdidas $f(\cdot, \xi)$ cumplen las condiciones de Lipschitz continuidad* (2.1) *y suavidad* (2.2) *para todo dato $\xi \in \mathcal{Z}$. Esto es,*

$$f(\cdot, \xi) \in \mathcal{F}_{M,L}(\mathcal{X}) \quad (\forall \xi \in \mathcal{Z}).$$

Supuesto 1c. *Las pérdidas $f(\cdot, \xi)$ cumplen las condiciones de Lipschitz continuidad* (2.1)*, suavidad* (2.2) *y fuerte convexidad* (2.3) *para todo dato $\xi \in \mathcal{Z}$. Esto es,*

$$f(\cdot, \xi) \in \mathcal{S}_{M,L}^\mu(\mathcal{X}) \quad (\forall \xi \in \mathcal{Z}).$$

Además, se asume que los métodos utilizados son tales que las trayectorias generadas por su aplicación pueden ser acotadas por una bola de radio suficientemente grande, como se indica en el Supuesto 2.

Supuesto 2. *Existe un radio $r \gg 0$ suficientemente grande tal que las trayectorias generadas por un algoritmo $\mathcal{A}$ a partir de un iterado inicial $x_1 \in \overline{\mathcal{B}_d(0, R)}$ estén al interior de la bola $\mathcal{B}_d(0, r)$.*

Adicionalmente, para una función $g : \mathcal{X} \to \mathbb{R}$, la aplicación del operador proximal asociado a g con constante η se denota por

$$\mathbf{prox}_{\eta g}(x) := \arg\min_{z \in \mathcal{X}} \left\{ \eta g(z) + \frac{1}{2}||z - x||^2 \right\}.$$

Suponiendo un algoritmo $\mathcal{A}$ que accede a la muestra aleatoria $\mathbf{S}$ y la pérdida f para entregar una solución aproximada al problema (1.1) después de $T \geq n$ pasos, se puede medir qué tan buena es en términos del exceso de riesgo (2.4).

$$\varepsilon_{risk}(\mathcal{A}, \mathbf{S}) := F(\mathcal{A}(\mathbf{S})) - F(x^*) \tag{2.4}$$

Se nota que esta cantidad depende de la aleatorización de $\mathcal{A}$ y de la muestra $\mathbf{S}$ utilizada por el algoritmo. Como se mencionó en la sección anterior, Shalev-Shwartz y Ben-David (2014) acotan el exceso de riesgo por una suma de variables aleatorias correspondientes a los errores de generalización, optimización y aproximación; definidas en (2.5). Se hace notar que todas presentan dependencias de $\mathbf{S}$, pero el error de aproximación no depende de la aleatoriedad del algoritmo empleado.

$$\varepsilon_{risk}(\mathcal{A}, \mathbf{S}) = \underbrace{F(\mathcal{A}(\mathbf{S})) - F_{\mathbf{S}}(\mathcal{A}(\mathbf{S}))}_{\varepsilon_{gen}(\mathcal{A},\mathbf{S})} + \underbrace{F_{\mathbf{S}}(\mathcal{A}(\mathbf{S})) - F_{\mathbf{S}}(x^*_{\mathbf{S}})}_{\varepsilon_{opt}(\mathcal{A},\mathbf{S})} + \underbrace{F_{\mathbf{S}}(x^*_{\mathbf{S}}) - F(x^*)}_{\varepsilon_{approx}(\mathbf{S})} \tag{2.5}$$

En pos de analizar el error de generalización, se introduce la noción de muestras vecinas, necesaria en el análisis del error de generalización mediante estabilidad de los métodos que se presentarán posteriormente.

Definición 2.1. *(Muestras vecinas) Se dice que dos muestras aleatorias* $\mathbf{S}, \mathbf{S}'$ *sobre* $\mathcal{Z}^n$ *son vecinas, denotado* $\mathbf{S} \simeq \mathbf{S}'$*, si difieren en a lo más uno de sus elementos. Es decir, existen un índice* $i \in [n]$ *y datos* $\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_n, \boldsymbol{\xi}'_i$ *tales que* $\mathbf{S} = (\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_n)$ *y* $\mathbf{S}' = (\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_{i-1}, \boldsymbol{\xi}'_i, \boldsymbol{\xi}_{i+1}, ..., \boldsymbol{\xi}_n)$.

Con esto es posible mayorar la diferencia, en términos de valores de la pérdida, de utilizar la respuesta de un algoritmo $\mathcal{A}$ para dos muestras vecinas, con el concepto de estabilidad uniforme:

Definición 2.2. *(Estabilidad uniforme) Se dice que un algoritmo $\mathcal{A}$ es γ-Uniformemente estable con respecto a una pérdida $f : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ si*

$$\sup_{\mathbf{S} \simeq \mathbf{S}', z} \mathop{\mathbb{E}}_{\mathcal{A}} [f(\mathcal{A}(\mathbf{S}), z) - f(\mathcal{A}(\mathbf{S}'), z)] \leq \gamma.$$

Alternativamente, se puede cuantificar la distancia entre trayectorias provenientes de muestras vecinas mediante la siguiente propiedad de estabilidad:

Definición 2.3. *(Estabilidad uniforme de argumentos) Se dice que un algoritmo $\mathcal{A}$ es γ-UAS (en inglés, Uniform Argument Stable) con respecto a una pérdida $f : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ si*

$$\sup_{\mathbf{S} \simeq \mathbf{S}'} \mathop{\mathbb{E}}_{\mathcal{A}} ||\mathcal{A}(\mathbf{S}) - \mathcal{A}(\mathbf{S}')|| \leq \gamma.$$

Cuando no exista ambigüedad respecto a la función de pérdida utilizada, se omitirá ésta y se hablará solo del tipo de estabilidad que un algoritmo cumple. Además, se puede apreciar que dado un algoritmo $\mathcal{A}$ y función $f \in \mathcal{F}_M(\mathcal{X})$, la propiedad de γ-UAS de $\mathcal{A}$ implica $M\gamma$-Unif. estabilidad, permitiendo conectar ambas definiciones anteriores (Hardt et al., 2016).

En los análisis teóricos de estabilidad posteriores se consideran dos trayectorias $\{x_t\}_{t \in [T+1]}$ e $\{y_t\}_{t \in [T+1]}$ asociadas respectivamente a las muestras $\mathbf{S} \simeq \mathbf{S}'$ que difieren, sin pérdida de generalidad, en la primera muestra enumerada. Este supuesto es válido para los métodos *batch*, donde la muestra completa se utiliza en cada iteración, y para los métodos incrementales con permutación aleatoria uniforme, donde el orden de los datos no importa debido a la invarianza bajo permutaciones. Sin embargo, estos resultados pueden extenderse fácilmente a métodos incrementales con permutación fija tras indicar una enumeración que tenga en cuenta la permutación del método, lo que no altera la garantía ni la idea de su demostración.

Se denota por δ_t a la variable aleatoria $||x_t - y_t||$ de los iterados generados por el algoritmo $\mathcal{A}$ sobre las muestras $\mathbf{S} \simeq \mathbf{S}'$. La diferencia entre las respuestas del algoritmo

para tales muestras se denota por la variable aleatoria

$$\delta(\mathcal{A}, \mathbf{S}, \mathbf{S}') = ||\mathcal{A}(\mathbf{S}) - \mathcal{A}(\mathbf{S}')||.$$

En vista de las consideraciones anteriores de estabilidad, se presenta el siguiente resultado de generalización para algoritmos aleatorizados que permite traducir la uniforme estabilidad de un método de descenso en una cota superior de alta probabilidad para el error de generalización con pérdidas Lipschitz.

Teorema 2.1. *(Bousquet et al., 2020) Sea el conjunto $\mathcal{X}$, f una función de pérdidas cumpliendo el Supuesto 1a y $\mathcal{A} : \mathcal{Z} \to \mathcal{X}$ un algoritmo aleatorizado γ-Uniformemente estable que cumple el Supuesto 2. Luego, existe una constante c tal que para toda distribución $\mathcal{D}$ sobre $\mathcal{Z}$ y $\theta \in (0, 1)$, se tiene*

$$\mathop{\mathbb{P}}_{\mathbf{S}\sim\mathcal{D}^n,\mathcal{A}}\left[|\varepsilon_{gen}(\mathcal{A}, \mathbf{S})| \geq c\left(\gamma \log(n) \log\left(\frac{1}{\theta}\right) + rM\sqrt{\frac{\log(1/\theta)}{n}}\right)\right] \leq \theta. \tag{2.6}$$

Adicionalmente, se expone el siguiente resultado que permite acotar con alta probabilidad el error de aproximación asociado a una muestra aleatoria $\mathbf{S}$.

Lema 2.1. *(Bassily et al., 2020) Sea un conjunto $\mathcal{X}$ r-acotado y una pérdida f cumpliendo el Supuesto 1a. Para todo $\theta \in (0, 1)$, con probabilidad al menos $1 - \theta$, el error de aproximación está acotado de la forma*

$$\varepsilon_{approx}(\mathbf{S}) \leq \frac{rM\sqrt{2\log(1/\theta)}}{\sqrt{n}}. \tag{2.7}$$

El Lema 2.1 utiliza la independencia de los datos que componen la muestra y una cota superior para el valor máximo de la pérdida. Si bien el resultado original no funciona para un conjunto $\mathcal{X}$ y pérdida f no-acotados ni depende de la elección del método, utilizar un método que cumpla el Supuesto 2 permite restringir el análisis del exceso de riesgo a un conjunto acotado de radio r y aplicar el Lema a la descomposición de riesgo (2.5) para el *setting* estudiado.

Alternativamente, un acercamiento en términos de cotas en valor esperado al problema facilita la descomposición del riesgo en la suma de esperanzas del error de optimización y de estabilidad, ya que el error de aproximación tiene un valor esperado no-positivo sobre $\mathbf{S}$ (Hardt et al., 2016). Tras tomar valor esperado a (2.5), se obtiene:

$$\mathop{\mathbb{E}}_{\mathbf{S},\mathcal{A}}\left[\varepsilon_{risk}(\mathcal{A},\mathbf{S})\right] \leq \mathop{\mathbb{E}}_{\mathbf{S},\mathcal{A}}\left[\varepsilon_{gen}(\mathcal{A},\mathbf{S})\right] + \mathop{\mathbb{E}}_{\mathbf{S},\mathcal{A}}\left[\varepsilon_{opt}(\mathcal{A},\mathbf{S})\right]. \tag{2.8}$$

Utilizando ambos resultados anteriores y la descomposición de (2.5) o (2.8), se reduce el problema de acotar superiormente el exceso de riesgo con alta probabilidad o en esperanza en obtener dos resultados: Una garantía de cota superior de convergencia en el peor caso para el error de optimización de un FOM y otra para la estabilidad del mismo, compatible con el supuesto del Teorema 2.1.

En las secciones siguientes se presentan cuatro tipos de iteración distintos para los casos Lipschitz y suave, junto con algunos resultados teóricos necesarios para obtener resultados de generalización en términos del número de iteraciones T y del tamaño n de la muestra. Además, presentan resultados conocidos para el error de optimización de un método estocástico acelerado en el caso fuertemente convexo. En las Secciones 2.1-2.5 y el Capítulo 3 se omiten dependencias que las cotas superiores e inferiores de los errores de optimización y estabilidad puedan tener con respecto a los otros parámetros del problema, tales como R, M y L. Asimismo, en estas secciones se omitirán tales dependencias al exponer órdenes de complejidad muestral, de iteraciones o de gradientes.

Además, se cuantifican las complejidades asociadas a aproximar la solución óptima del problema SCO mediante un método de primer orden en términos de un error ε. En particular, se presentan órdenes del tamaño de la muestra, el número de iteraciones y el cálculo total de subgradientes requeridos para garantizar un exceso de riesgo a lo más ε.

2.1. El método de (sub)gradiente

Se presenta la iteración de subgradiente (2.9) para una muestra S:

$$x_{t+1} := \Pi_{\mathcal{X}} \left[x_t - \eta \nabla F_S(x_t) \right]. \tag{2.9}$$

Primero, se considera el caso general de pérdidas M-Lipschitz. Dado que no se asume diferenciabilidad de las funciones, la actualización requiere de una elección arbitraria de subgradiente $\nabla F_S(x_t)$. Para el caso suave, la elección de subgradiente es única y correspondiente al gradiente de la función. En tal caso, se habla del "Método de Gradiente".

Algoritmo 1: $\mathcal{A}_{\text{GD}}$: Método de subgradiente.

Input: Muestra $S = (\xi_1, ..., \xi_n) \in \mathcal{Z}^n$, número de iteraciones T, paso η, $x_1 \in \mathcal{X}$ inicial ;

for $t = 1...T$ **do**
  Define $x_{t+1} := \Pi_{\mathcal{X}} \left[x_t - \eta \nabla F_S(x_t) \right]$;
end
return x_{T+1}

Siguiendo esta dinámica de descenso, se define el Algoritmo 1 que tiene como respuesta el último iterado. Sin embargo, la secuencia de estos no garantiza valores de la función no decrecientes y no es posible obtener un resultado de convergencia no trivial para el último iterado del caso supuesto. Así pues, se define el Algoritmo 2, que entrega por respuesta el promedio de los iterados después de cada iteración, en la literatura comúnmente referido como *model averaging*, $\bar{x}_T = \frac{1}{T} \sum_{t \in [T]} x_{t+1}$.

Algoritmo 2: $\mathcal{A}_{\text{AvGD}}$: Método de subgradiente promediado.

Input: Muestra $S = (\xi_1, ..., \xi_n) \in \mathcal{Z}^n$, número de iteraciones T, paso η, $x_1 \in \mathcal{X}$ inicial ;

for $t = 1...T$ **do**
  Define $x_{t+1} := \Pi_{\mathcal{X}} \left[x_t - \eta \nabla F_S(x_t) \right]$;
end
return $\frac{1}{T} \sum_{t=1}^{T} x_{t+1}$

Esta nueva respuesta del método permite utilizar el Lema 2.2 de convergencia para funciones convexas.

Lema 2.2. *(Shor, 2012) Sea una pérdida $f(\cdot, \xi) \in \mathcal{F}_M(\mathcal{X})$ para todo $\xi \in \mathcal{Z}$ (Supuesto 1a). Luego, el método de subgradiente (Algoritmo 2) con paso $\eta > 0$ garantiza*

$$F_S(\bar{x}_T) - F_S(x_S^*) \leq \frac{||x_1 - x_S^*||^2}{2\eta T} + \frac{M^2\eta}{2}.$$

En consecuencia, la elección de un paso $\eta \in \Theta\left(\frac{1}{\sqrt{T}}\right)$ para el Algoritmo 2 aplicado a una función con pérdidas Lipschitz continuas asegura un error de optimización $O\left(\frac{1}{\sqrt{T}}\right)$. Restringiendo la clase de pérdidas a funciones L-suaves es posible mejorar la convergencia en términos de valores de la función. Más aún, el Lema 2.3 garantiza convergencia para el último iterado cuando se asume suavidad de las funciones.

Lema 2.3. *(Beck, 2017) Sea $g \in \mathcal{F}_{M,L}(\mathcal{X})$. Luego, el método de gradiente con paso $\eta \leq \frac{1}{L}$ garantiza*

$$g(x_T) - \min_{x \in \mathcal{X}} g(x) \leq \frac{||x_1 - x^*||^2}{2\eta T}.$$

En particular, la convergencia del error de optimización del Algoritmo 1 sobre una pérdida cumpliendo el Supuesto 1b es de orden $O\left(\frac{1}{T}\right)$, tomando un paso constante $\eta \leq \frac{1}{L}$. Resta mostrar resultados de estabilidad algorítmica, partiendo por el caso no-suave.

Lema 2.4. *(Bassily et al., 2020) Sea una pérdida $f(\cdot, \xi) \in \mathcal{F}_M(\mathcal{X})$ para todo $\xi \in \mathcal{Z}$ (Supuesto 1a). Luego, el Algoritmo 2 cumple para todo par de muestras $\mathbf{S} \simeq \mathbf{S}'$*

$$\delta(\mathcal{A}_{AvGD}, \mathbf{S}, \mathbf{S}') \leq 2\eta M\sqrt{T} + \frac{4\eta MT}{n}.$$

Esta cantidad puede ser mejorada por un factor constante trabajando las sumatorias correspondientes a cada término, sin embargo, el mayor problema es la presencia del término $\Theta\left(\eta\sqrt{T}\right)$, pues este requiere fijar un largo de paso $\eta = o(T^{-\frac{1}{2}})$ para asegurar la convergencia de la estabilidad a cero, cuando n tiende a infinito.

Este término desaparece en el caso suave, donde se tiene la siguiente consecuencia directa de la no-expansividad de la regla de gradiente sobre una función suave con elección de paso suficientemente pequeño:

Afirmación 2.1. *(Hardt et al., 2016) Sea una pérdida $f(\cdot, \xi) \in \mathcal{F}_{M,L}(\mathcal{X})$ para todo $\xi \in \mathcal{Z}$ (Supuesto 1b). El Algoritmo 1 con paso $\eta \leq \frac{2}{L}$ tiene $\frac{2\eta MT}{n}$-Unif. estabilidad de argumentos.*

Unificando estos resultados de convergencia y estabilidad con la descomposición de riesgo en esperanza de (2.8) se concluye que el Algoritmo 1 con elección de *learning rate* $\eta = O(\sqrt{n}/T)$ garantiza, para pérdidas suaves, un exceso de riesgo en esperanza $O(1/\sqrt{n})$. Alternativamente, utilizando el Teorema 2.1 y Lema 2.1 con (2.5) se puede concluir que el algoritmo garantiza una aproximación del mismo orden con alta probabilidad, salvo términos polilogarítmicos, si cumple el Supuesto 2.

Por el contrario, para el caso no-suave, Bassily et al. (2020) concluyen que fijando $T = n^2$ iteraciones del Algoritmo 2 y paso $\eta = O\left(T^{-3/4}\right)$, una cota superior para el exceso de riesgo en valor esperado es del orden $O\left(T^{-\frac{1}{4}}\right)$ y con alta probabilidad, el exceso de riesgo será $\tilde{O}\left(T^{-\frac{1}{4}}\right)$. Asimismo, Amir, Koren, y Livni (2021) discuten que el método de subgradiente requiere $\Omega\left(\frac{1}{\varepsilon^4}\right)$ iteraciones para generalizar, concluyendo que incluso la cota para el exceso de riesgo es de orden ajustado.

2.2. El método de gradiente estocástico incremental

Se presenta la iteración de gradiente estocástico (2.10) para una muestra S,

$$x_{t+1} := \Pi_{\mathcal{X}}\left[x_t - \eta \nabla f_t(x_t)\right], \tag{2.10}$$

donde $f_t := f(\cdot, \xi_{i_t})$ corresponde a la pérdida asociada a la elección de un dato para la iteración t. La selección de índice suele ajustarse a uno de dos regímenes aleatorios. En el primero, se considera elección aleatoria uniforme con reemplazo en cada iteración. En el segundo, llamado incremental de permutación aleatoria se consideran T/n pasadas por

cada dato en S, donde la k-ésima ronda $\left(\text{para } k \in \left[\frac{T}{n}\right]\right)$ se ordena según una respectiva permutación aleatoria uniforme π_k sobre $[n]$. Una alternativa similar a esta última supone una elección determinística de la misma permutación en cada ciclo, llamado método de gradiente estocástico con permutación fija y presentado en el Algoritmo 4.

Algoritmo 3: $\mathcal{A}_{\text{SGD}}$: Método de Gradiente Estocástico (SGD) incremental con permutación aleatoria.

Input: Muestra $S = (\xi_1, ..., \xi_n) \in \mathcal{Z}^n$, número de rondas K, paso η, permutaciones uniformes $\boldsymbol{\pi}_1, ..., \boldsymbol{\pi}_K$ sobre $[n]$, $x_1 \in \mathcal{X}$ inicial ;

for $k = 1...K$ **do**
  for $i = 1...n$ **do**
    $x_{(k-1)n+i+1} := \Pi_{\mathcal{X}}\left(x_{(k-1)n+i} - \eta \nabla f(x_{(k-1)n+i}, \xi_{\boldsymbol{\pi}_k(i)})\right)$;
  end
end
return $\bar{x}_K := \frac{1}{K}\sum_{k\in[K]} x_{kn+1}$

Se hace notar que el promedio de los iterados se hace entre el iterado final de cada permutación, a diferencia del Algoritmo 2. Esto mantiene la idea de considerar para el promedio los iterados después de cada recorrido por toda la muestra. Se presenta un resultado de optimización para el caso no-suave con K pasadas en el Lema 2.5.

Algoritmo 4: $\mathcal{A}_{\text{SGD}}$: Método de Gradiente Estocástico (SGD) incremental con permutación fija.

Input: Muestra $S = (\xi_1, ..., \xi_n) \in \mathcal{Z}^n$, número de rondas K, paso η, permutación π sobre $[n]$, $x_1 \in \mathcal{X}$ inicial ;

for $k = 1...K$ **do**
  for $i = 1...n$ **do**
    $x_{(k-1)n+i+1} := \Pi_{\mathcal{X}}\left(x_{(k-1)n+i} - \eta \nabla f(x_{(k-1)n+i}, \xi_{\pi(i)})\right)$;
  end
end
return $\bar{x}_K := \frac{1}{K}\sum_{k\in[K]} x_{kn+1}$

Lema 2.5. *(Nedic y Bertsekas, 2001) Sea una pérdida $f(\cdot, \xi) \in \mathcal{F}_M(\mathcal{X})$ para todo $\xi \in \mathcal{Z}$ (Supuesto 1a). Luego, el Algoritmo 3 con paso $\eta > 0$ garantiza, para todo $S \in \mathcal{Z}^n$ y toda secuencia de permutaciones $\{\pi_k\}_{k\in[K]}$ (No necesariamente aleatoria):*

$$\varepsilon_{opt}(\mathcal{A}_{SGD}, S) \leq \frac{||x_1 - x_S^*||^2}{2\eta T} + \frac{M^2\eta(n+2)}{2}.$$

Este resultado permite convergencia sin necesidad de restringir la reutilización de datos, como es el caso del método SGD con *Model Averaging* de Polyak y Juditsky (1992).

A continuación, se presenta el análisis del error de estabilidad para los casos no-suave y suave de SGD incremental, donde restringir a $\mathcal{F}_{M,L}(\mathcal{X})$ sí produce una mejora en la garantía.

Lema 2.6. *(Bassily et al., 2020) Sea una pérdida $f(\cdot,\xi) \in \mathcal{F}_M(\mathcal{X})$ para todo $\xi \in \mathcal{Z}$ (Supuesto 1a). El Algoritmo 3 cumple, para toda secuencia de permutaciones $\{\pi_k\}_{k\in[K]}$ (No necesariamente aleatoria)*

$$\sup_{\mathbf{S}\simeq\mathbf{S}'} \delta(\mathcal{A}_{SGD}, \mathbf{S}, \mathbf{S}') \leq 2\eta M\sqrt{T} + \frac{4\eta MT}{n}.$$

El Lema 2.6 muestra que en el caso no-suave, SGD incremental tiene UAS idéntica al método de subgradiente. Sin embargo, en cada iteración (2.10) se realiza solamente una evaluación de gradiente, mientras que la iteración de tipo (2.9) realiza n distintas. Una mejora para la estabilidad algorítmica de SGD se presenta en el Lema 2.7 para el caso suave.

Lema 2.7. *(Hardt et al., 2016) Sea una pérdida $f(\cdot,\xi) \in \mathcal{F}_{M,L}(\mathcal{X})$ para todo $\xi \in \mathcal{Z}$ (Supuesto 1b). El Algoritmo 3 cumple, para toda secuencia de permutaciones $\{\pi_k\}_{k\in[K]}$ (No necesariamente aleatoria)*

$$\sup_{\mathbf{S}\simeq\mathbf{S}'} \delta(\mathcal{A}_{SGD}, \mathbf{S}, \mathbf{S}') \leq \frac{2\eta MT}{n}.$$

Pese a la mejoría mostrada por el Lema 2.7 para la Uniforme estabilidad de argumentos, el término $O(\eta n)$ del error de optimización no permite mejorar el exceso de riesgo utilizando la suavidad y la descomposición de riesgo (2.5). Realizando $Kn = n^2$ iteraciones y utilizando un paso $\eta = \Theta(n^{-3/2})$ se puede garantizar un exceso de riesgo $O(n^{-1/2})$ tanto para pérdidas Lipschitz como pérdidas suaves. Aún así, se tiene una mejoría respecto al método *Batch* no-suave, pues para ambos casos la realización de llamadas al oráculo de primer orden es de un orden menor que en el método de subgradiente.

Por último, se nota que las garantías mostradas para el Algoritmo 3 funcionan para una elección no necesariamente uniforme de las permutaciones. Más aún, la elección de estos ordenamientos no tiene por que ser aleatoria, resultando que tales garantías son válidas para el caso del Algoritmo 4.

2.3. El método de punto proximal

Una alternativa a las iteraciones de tipo gradiente es el método de punto proximal (PPM), cuya regla de actualización (2.11) está dada por una aplicación del operador proximal.

$$x_{t+1} = \mathbf{prox}_{\eta F_S}(x_t) \tag{2.11}$$

Esta regla de actualización puede ser representada por su forma equivalente (2.12), que a diferencia de la regla de gradiente, necesita acceso a un oráculo de subgradiente en un punto distinto a x_t donde, en general, no es fácil calcular este subgradiente del iterado siguiente.

$$x_{t+1} = \Pi_{\mathcal{X}} \left[x_t - \eta \nabla F_S(x_{t+1}) \right] \tag{2.12}$$

La dificultad de obtener el subgradiente se compensa con una mejora en la convergencia del algoritmo, provista por la suavidad que se induce en el problema. Esto se debe a la equivalencia entre la regla proximal (2.11) sobre una función no necesariamente suave y la regla de gradiente (2.9) sobre su Envoltura de Moreau con parámetro de regularización adecuado (Beck, 2017). Se define el Algoritmo 5 en base a esta nueva regla de actualización.

Algoritmo 5: $\mathcal{A}_{\text{Prox}}$: Método de punto Proximal.

Input: Muestra $S = (\xi_1, ..., \xi_n) \in \mathcal{Z}^n$, número de iteraciones T, paso de largo η, $x_1 \in \mathcal{X}$ inicial ;

for $t = 1...T$ **do**

  Elige $x_{t+1} \in \arg\min_{z \in \mathcal{X}} \left\{ F_S(z) + \frac{1}{2\eta} ||z - x_t||^2 \right\}$;

end

return x_{T+1}

El Lema 2.8 presenta una cota para el error de optimización igual a la presentada en el Lema 2.3 para el método de gradiente, lo que es consistente con la interpretación de este nuevo método como una aplicación del método de gradiente en el caso suave.

Lema 2.8. *(Güler, 1991) Sea una pérdida $f(\cdot, \xi) \in \mathcal{F}_M(\mathcal{X})$ para todo $\xi \in \mathcal{Z}$ (Supuesto 1a). Luego, el método de punto proximal con paso $\eta > 0$ garantiza*

$$F_S(x_{T+1}) - F_S(x_S^*) \leq \frac{||x_1 - x_S^*||^2}{2\eta T}.$$

Se nota que suponer suavidad de las pérdidas, por sí misma, no garantiza una mejora de la convergencia. Para una función con pérdidas Lipschitz, se considera una elección de largo de paso $\eta = \Theta\left(\frac{\sqrt{n}}{T}\right)$ que garantiza un error de optimización $O(n^{-1/2})$. Nótese que independiente de la elección de η y T para aplicar el método, el Lema 2.1 entrega una convergencia del error de aproximación de orden $\tilde{O}(n^{-1/2})$. Por lo mismo, una elección de parámetros que alcance error de optimización $o\left(\frac{1}{\sqrt{n}}\right)$ no implicará una mejora en términos del exceso de riesgo.

Respecto a la cuantificación de la estabilidad, en cambio, cómo establecer el símil con el método de gradiente no es claro. Se deduce del trabajo de Shalev-Shwartz, Shamir, Srebro, y Sridharan (2010, Demostración del Teorema 2) que una iteración del método proximal es $\frac{4M^2\eta}{n}$-uniformemente estable. Sin embargo, la idea de su demostración no puede ser replicada a más de una iteración directamente. Por otra parte, Rockafellar (1976) demostró la no-expansividad del operador $\mathbf{prox}_{\eta f}$ para f cerrada, convexa y propia. La propiedad de no-expansividad se utiliza en la elaboración de garantías de uniforme estabilidad para actualizaciones de gradiente, en Hardt et al. (2016). Aún así, no es inmediato cómo emplear cualquiera de estas técnicas a la actualización proximal, por sí mismas. La dificultad de obtener una demostración de estabilidad para PPM radica en que el gradiente involucrado en (2.11) para muestras vecinas utiliza subgradientes de funciones distintas en puntos distintos, limitando ambos acercamientos.

2.4. El método *IncrementalProx*

El símil estocástico del método proximal, es el método de punto proximal incremental o *IncrementalProx*, definido por la regla de actualización (2.13).

$$x_{t+1} \in \arg\min_{z\in\mathcal{X}} \left\{ f(z, \xi_{t \bmod n}) + \frac{1}{2\eta}||z - x_t||^2 \right\} \tag{2.13}$$

Este método elige un dato ξ en cada iteración y realiza un paso proximal con la pérdida $f(\cdot, \xi)$. Nuevamente, la actualización proximal de (2.13) puede reescribirse como una iteración implícita, resultando (2.14).

$$x_{t+1} = \Pi\left[x_t - \eta\nabla f(x_{t+1}, \xi_{t \bmod n})\right] \tag{2.14}$$

En base a la regla iterativa anterior, se formaliza este método de descenso en el Algoritmo 6.

Algoritmo 6: $\mathcal{A}_{\text{IP}}$: Método *IncrementalProx* con permutación aleatoria.

Input: Muestra $S = (\xi_1, ..., \xi_n) \in \mathcal{Z}^n$, número de rondas K, paso η, permutaciones uniformes $\boldsymbol{\pi}_1, ..., \boldsymbol{\pi}_k$ sobre $[n]$, iterado inicial $x_1 \in \mathcal{X}$;

for $k = 1...K$ **do**
  for $i = 1...n$ **do**
    Elige $x_{(k-1)n+i+1} \in \arg\min_{z\in\mathcal{X}} \left\{ f(z, \xi_{\boldsymbol{\pi}_k(i)}) + \frac{1}{2\eta}||z - x_{(k-1)n+i}||^2 \right\}$;
  end
end
return $\bar{x}_K := \frac{1}{K}\sum_{k\in[K]} x_{(kn+1)}$

Algoritmo 7: $\mathcal{A}_{\text{f.IP}}$: Método *IncrementalProx* con permutación fija.

Input: Muestra $S = (\xi_1, ..., \xi_n) \in \mathcal{Z}^n$, número de rondas K, paso η, permutacion π sobre $[n]$, iterado inicial $x_1 \in \mathcal{X}$;

for $k = 1...K$ **do**
  for $i = 1...n$ **do**
    Elige $x_{(k-1)n+i+1} \in \arg\min_{z\in\mathcal{X}} \left\{ f(z, \xi_{\pi_k(i)}) + \frac{1}{2\eta}||z - x_{(k-1)n+i}||^2 \right\}$;
  end
end
return $\bar{x}_K := \frac{1}{K}\sum_{k\in[K]} x_{(kn+1)}$

Respecto a la convergencia de este método, Bertsekas (2011) acota el error de optimización alcanzado por métodos de gradiente proximal incremental para el caso de una función compuesta. Este desarrollo puede ser refinado en el caso no compuesto para *IncrementalProx*, caso especial del método de gradiente proximal incremental, presentado en la Proposición 2.1.

Lema 2.9. *(Bertsekas, 2011) Sea $\mathcal{X} \subset \mathbb{R}^d$ convexo, cerrado, no-vacío y $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\} \in \Gamma_0(\mathbb{R}^d)$ tal que ri($\mathcal{X}$) $\cap$ ri(dom(g)) $\neq \emptyset$. Sean $z_t \in \mathbb{R}^d$ y $\eta > 0$, luego la iteración proximal $z_{t+1} = \mathbf{prox}_{\eta g}(z_t)$ cumple, para todo $z \in \mathcal{X}$, la desigualdad*

$$||z_{t+1} - z||^2 \leq ||z_t - z||^2 - 2\eta \left(g(z_{t+1}) - g(z)\right).$$

Antes se presenta el Lema 2.9, que permite la demostración de la convergencia del Algoritmo 6. Este lema permite controlar a través de las distancias de iterados consecutivos a un punto arbitrario los valores de la muestra que los conecta, como se utiliza en la demostración de la Proposición 2.1 con los iterados resultantes de cada aplicación del operador $\mathbf{prox}_{\eta f(\cdot, \xi_{\pi_k(i)})}$. Se nota que los supuestos de interior relativo son cumplidos en el caso no-suave, ya que el interior relativo de un convexo no-vacío es no-vacío, $\mathcal{X} \subset \mathbb{R}^d$ y $dom(g) = \mathbb{R}^d$ a partir de la condición de Lipschitz.

Proposición 2.1. *Sea una pérdida $f(\cdot, \xi) \in \mathcal{F}_M(\mathcal{X})$ para todo $\xi \in \mathcal{Z}$ (Supuesto 1a). Luego, el Algoritmo 6 con paso $\eta > 0$ garantiza*

$$F_S(\bar{x}_K) - F_S(x_S^*) \leq \frac{||x_1 - x_S^*||^2}{2\eta K n} + \frac{M^2 \eta n}{2}.$$

DEMOSTRACIÓN. *Primero, se demuestra que el gap de optimalidad de la respuesta al final de cada ronda puede acotarse superiormente por cantidades fijas. Sea $y \in \mathcal{X}$ y $r_{kn+i} := ||x_{kn+i} - y||$. Por el Lema 2.9 para una iteración del Algoritmo 6 se tiene:*

$$\begin{aligned} r_{(k-1)n+i+1}^2 - r_{(k-1)n+i}^2 &\leq -2\eta \left[f(x_{(k-1)n+i+1}, \xi_{\pi_k(i)}) - f(y, \xi_{\pi_k(i)})\right] \\ &= -2\eta \left[f(x_{(k-1)n+i+1}, \xi_{\pi_k(i)}) - f(y, \xi_{\pi_k(i)}) \pm f(x_{kn+1}, \xi_{\pi_k(t)})\right]. \end{aligned}$$

Sumando sobre el k-ésimo recorrido sobre la muestra:

$$\begin{aligned} r_{kn+1}^2 - r_{(k-1)n+1}^2 \leq & -2\eta \sum_{i=1}^{n} \left[f(x_{(k-1)n+i+1}, \xi_{\pi_k(i)}) - f(y, \xi_{\pi_k(i)}) \pm f(x_{kn+1}, \xi_{\pi_k(i)}) \right] \\ = & 2\eta \sum_{i=1}^{n} \left[f(x_{kn+1}, \xi_{\pi_k(i)}) - f(x_{(k-1)n+i+1}, \xi_{\pi_k(i)}) \right] \\ & - 2\eta n \left[F_S(x_{kn+1}) - F_S(y) \right]. \end{aligned}$$

Entonces, utilizando la M-Lipschitz continuidad de las pérdidas y (2.14):

$$\begin{aligned} \leq & 2\eta M \sum_{i=1}^{n} ||x_{kn+1} - x_{(k-1)n+i+1}|| - 2\eta n \left[F_S(x_{kn+1}) - F_S(y) \right] \\ \leq & 2\eta^2 M^2 \sum_{i=1}^{n} (n-i) - 2\eta n \left[F_S(x_{kn+1}) - F_S(y) \right]. \end{aligned}$$

Acotando superiormente la suma:

$$\leq \eta^2 M^2 n^2 - 2\eta n \left[F_S(x_{kn+1}) - F_S(y) \right].$$

Reordenando, para $y = x_S^*$:

$$\Rightarrow F_S(x_{kn+1}) - F_S(x_S^*) \leq \frac{1}{2\eta n} \left[\eta^2 M^2 n^2 + r_{(k-1)n+1}^2 - r_{kn+1}^2 \right].$$

Luego, sumando en el número de rondas:

$$\begin{aligned} \sum_{k=1}^{K} \left[F_S(x_{kn+1}) - F_S(x_S^*) \right] \leq & \frac{1}{2\eta n} \left[\eta^2 M^2 n^2 K + r_1^2 - r_{Kn+1}^2 \right] \\ \leq & \frac{1}{2\eta n} \left[\eta^2 M^2 n^2 K + r_1^2 \right]. \end{aligned}$$

Finalmente, por convexidad en $\bar{x}_K$:

$$\begin{aligned} F_S(\bar{x}_K) - F_S(x_S^*) \leq & \frac{1}{K}\sum_{k=1}^{K}\left[F_S(x_{kn+1}) - F_S(x_S^*)\right] \\ \leq & \frac{||x_1 - x_S^*||^2}{2\eta K n} + \frac{M^2\eta n}{2}. \end{aligned}$$

□

A partir del resultado anterior, se deduce que la elección de paso $\eta \in \Theta\left(\frac{1}{n\sqrt{K}}\right)$ entrega un error de optimización $O\left(\frac{1}{\sqrt{K}}\right)$ luego de Kn iteraciones. Suponer $K = n$ pasadas por cada dato garantiza una complejidad muestral $O\left(\frac{1}{\varepsilon^2}\right)$ para este tipo de error luego de $O\left(\frac{1}{\varepsilon^4}\right)$ iteraciones. Esto se suma a una convergencia del mismo orden del error de aproximación según el Lema 2.1. No obstante, este caso es similar a SGD para pérdidas no-suaves, donde ambos tipos de error pueden alcanzar esta tasa de convergencia por separado. En el caso de gradiente estocástico, la limitante de una mejor convergencia del exceso de riesgo es sólo el error de generalización.

La estabilidad algorítmica de este método puede ser derivada con las herramientas usuales de estabilidad en el caso Lipschitz y del operador proximal.

Proposición 2.2. *Sea una pérdida $f(\cdot, \xi) \in \mathcal{F}_M(\mathcal{X})$ para todo $\xi \in \mathcal{Z}$ (Supuesto 1a). Luego, el Algoritmo 6 después de $T = Kn$ iteraciones alcanza $2\eta MK$-Uniforme estabilidad de argumentos.*

DEMOSTRACIÓN. *Se acota superiormente la distancia entre los iterados de ambas trayectorias. Separamos dos casos:*

Supongamos $\pi_k(t \bmod n) = 1$, esto es, aquel caso donde ambas trayectorias utilizan la muestra en que difieren. Notar que, sin pérdida de generalidad, se puede suponer que la diferencia de las muestras se tiene en la primera muestra. Para $t = (k-1)n + i$:

$$\begin{aligned}\delta_{t+1} =& ||\Pi_{\mathcal{X}}[x_t - \eta\nabla f(x_{t+1}, \xi_1)] - \Pi_{\mathcal{X}}[y_t - \eta\nabla f(y_{t+1}, \xi_1')]|| \\ \leq& ||x_t - y_t|| + \eta||\nabla f(x_{t+1}, \xi_1) - \nabla f(y_{t+1}, \xi_1')|| \\ \leq& \delta_t + 2\eta M,\end{aligned}$$

donde para la primera desigualdad se utiliza la no-expansividad del operador proyección y desigualdad triangular, mientras que en la segunda se usa la M-Lipschitz continuidad de $f(\cdot, \xi)$. Consideramos ahora el caso que ambas muestras son iguales:

$$\begin{aligned}\delta_{t+1} =& ||Prox_{\eta f(\cdot, \xi_{\pi_k(i)})}(x_t) - Prox_{\eta f(\cdot, \xi'_{\pi_k(i)})}(y_t)|| \\ =& ||Prox_{\eta f(\cdot, \xi_{\pi_k(i)})}(x_t) - Prox_{\eta f(\cdot, \xi_{\pi_k(i)})}(y_t)||.\end{aligned}$$

Por la no-expansividad del operador proximal:

$$\leq ||x_t - y_t||.$$

Finalmente, como consideramos el método incremental con permutaciones, después de $K = \frac{T}{n}$ utilizaciones de cada una de las n muestras, se obtiene

$$\delta(\mathcal{A}_{IP}, S, S') = \delta_{T+1} \leq 2\eta MK.$$

□

Se nota que las Proposiciones 2.1 y 2.2 no dependen de una elección uniforme de las permutaciones. En particular, funcionan para órdenes arbitrarios, mientras se mantenga una utilización de cada dato una vez en cada ronda. Por lo tanto, ambos resultados resultan válidos para el caso de permutación fija del Algoritmo 7.

La Estabilidad uniforme alcanzada es de orden $O(\eta K)$. Luego, basta tomar las elecciones de K y η anteriores para alcanzar una convergencia de tasa $O\left(\frac{1}{\sqrt{n}}\right)$ del error de optimización en esperanza. En consecuencia, se tiene una tasa del mismo orden para el exceso de riesgo esperado. Nuevamente, recurriendo al Lema 2.1 y Teorema 2.1, es posible obtener una tasa de convergencia $\tilde{O}\left(\frac{1}{\sqrt{n}}\right)$ con alta probabilidad.

2.5. El método de gradiente estocástico con varianza reducida

El método SVRG es una versión acelerada de SGD, definido por la regla de actualización (2.15). Este método se encuentra sujeto a una elección de punto $\tilde{x}$ en cada una de K rondas, que mantiene la información de gradientes en la ronda anterior. La aceleración se basa en el cálculo del gradiente del riesgo empírico en $\tilde{x}$ cada m iteraciones, utilizando la diferencia entre el gradiente completo y una elección estocástica de una de sus componentes convenientemente. Así, se genera una reducción de la varianza que no sesga la aproximación estocástica de la pérdida empírica en SGD con muestreo-con-reemplazo, que permite una convergencia más rápida que SGD para el iterado final.

$$x_{t+1} = x_t - \nabla f(x_t, \xi_{i_t}) + \nabla f(\tilde{x}, \xi_{i_t}) - \nabla F_S(\tilde{x}) \tag{2.15}$$

Este método realiza m iteraciones en cada una de K rondas, calculando al principio de cada pasada un gradiente del riesgo empírico sobre un punto $\tilde{x}$ que depende de la implementación, con información de la ronda anterior. En particular, Johnson y Zhang (2013) describen dos posibles alternativas, la elección de $\tilde{x}$ determinística como el último iterado de la ronda pasada o la elección aleatorizada y uniformemente distribuida en los iterados donde se realiza cada actualización de la ronda anterior. En el Algoritmo 8 se formaliza la descripción anterior, para esta última elección de $\tilde{x}$.

Johnson y Zhang (2013) garantizan convergencia lineal del error de optimización para el Algoritmo 8 en el caso fuertemente convexo, como se expresa en el Lema 2.10. Los autores comentan que la evidencia computacional sugiere la elección de parámetros $m = O(n)$ y $K \approx \frac{1}{\varepsilon}$, que alcanza complejidades de iteración y cálculo de gradientes

Algoritmo 8: $\mathcal{A}_{\text{SVRG}}$: Método SVRG con elección de último iterado estocástica.

Input: Muestra $S = (\xi_1, ..., \xi_n) \in \mathcal{Z}^n$, número de rondas K, pasos del ciclo interno m, *Learning rate* η, iterado inicial $\tilde{x}_1 \in \mathcal{X}$;

for $k = 1...K$ **do**
  Define $\mu_k := \nabla F_S(\tilde{x}_k)$;
  Define $x_1^k := \tilde{x}_k$;
  for $j = 1...m$ **do**
    Muestrea $\mathbf{i}_{k,j} \sim \text{Unif}[n]$;
    Define $x_{j+1}^k := x_j^k - \eta \nabla f(x_j^k, \xi_{\mathbf{i}_{k,j}}) + \eta \nabla f(\tilde{x}_k, \xi_{\mathbf{i}_{k,j}}) - \eta \mu_k$;
  end
  Muestrea $\mathbf{j} \sim \text{Unif}[m]$;
  Define $\tilde{x}_{k+1} := x_{\mathbf{j}}^k$;
end
return $\tilde{x}_{K+1}$

del mismo orden que otros métodos acelerados basados en la misma idea de reducción de varianza. Además, discuten la extensión al *setting* suave y sin supuestos de convexidad fuerte (Supuesto 1b) mediante regularización, alcanzando una tasa de convergencia $O\left(\frac{1}{T}\right)$ para este tipo de error.

Lema 2.10. *(Johnson y Zhang, 2013) Sea una pérdida $f(\cdot, \xi) \in \mathcal{S}_{M,L}^{\mu}(\mathbb{R}^d)$ para todo $\xi \in \mathcal{Z}$ (Supuesto 1c) y sea m suficientemente grande para cumplir*

$$\beta = \frac{1}{\mu \eta m (1 - 2L\eta)} + \frac{2L\eta}{1 - 2L\eta} < 1,$$

el Algoritmo 8 cumple

$$\mathop{\mathbb{E}}_{\mathcal{A}} \left[F_{\mathcal{S}}(\tilde{x}_{K+1}) - F_{\mathcal{S}}(x_S^*) \right] \leq \beta^K \left[F_{\mathcal{S}}(\tilde{x}_1) - F_{\mathcal{S}}(x_S^*) \right].$$

Tener un resultado para el error de optimización del caso fuertemente convexo permite derivar un análisis para el caso suave mediante el uso de un regularizador. Así, se analiza el problema (1.3) sobre una pérdida μ-regularizada f_μ como fue definida anteriormente, eligiendo el regularizador cuadrático

$$\psi(x) = \frac{1}{2} ||x - \tilde{x}_1||^2.$$

Un cálculo simple muestra que agregar tal regularizador entrega un término adicional $O(\mu)$. Sea $\tilde{x}_1$ un iterado inicial tal que $||\tilde{x}_1 - x_S^*|| \leq R$:

$$\mathop{\mathbb{E}}_{\mathcal{A}}[F_S(\tilde{x}_{K+1}) - F_S(x_S^*)] \leq \mathop{\mathbb{E}}_{\mathcal{A}}[F_S^\mu(\tilde{x}_{K+1}) - F_S^\mu(x_S^*)] + \frac{\mu R^2}{2}. \tag{2.16}$$

Estos son los llamados *métodos indirectos* para el caso suave de (1.3). Allen-Zhu y Yuan (2016) discuten que la desventaja de este tipo de métodos radica en que en la práctica el control de parámetros propios de la regularización como μ debe hacerse *a priori*, fijando los valores de tales cantidades en términos de otros parámetros del problema no regularizado para garantizar convergencia. Más aún, (2.16) nos muestra que en el caso expuesto la elección de un parámetro $\mu \in O\left(\frac{1}{T}\right)$ debe imponerse para lograr la tasa de convergencia para error de optimización del caso suave de Johnson y Zhang (2013) por este medio.

Asimismo, el caso suave del problema (1.1) puede ser abordado mediante regularización, obteniendo una cota superior para el exceso de riesgo original que depende del exceso de riesgo regularizado y un término adicional $O(\mu)$. Específicamente, definiendo $y^* = \arg\min_{x\in\mathcal{X}} F^\mu(x)$ y $y_S^* = \arg\min_{x\in\mathcal{X}} F_S^\mu(x)$ se tiene la siguiente descomposición en valor esperado similar a (2.8):

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathcal{A},S}[F(\mathcal{A}(S)) - F(x^*)] &\leq \mathop{\mathbb{E}}_{\mathcal{A},S}[F^\mu(\mathcal{A}(S)) - F^\mu(x^*)] + \frac{\mu}{2}||\tilde{x}_1 - x^*||^2 \\
&\leq \mathop{\mathbb{E}}_{\mathcal{A},S}[F^\mu(\mathcal{A}(S)) - F^\mu(y^*)] + \frac{\mu}{2}||\tilde{x}_1 - x^*||^2 \\
&\leq \mathop{\mathbb{E}}_{\mathcal{A},S}[F_S^\mu(\mathcal{A}(S)) - F_S^\mu(y_s^*)] + \mathop{\mathbb{E}}_{\mathcal{A},S}[F_S^\mu(\mathcal{A}(S)) - F^\mu(\mathcal{A}(S))] \\
&\quad + \frac{\mu}{2}||\tilde{x}_1 - x^*||^2.
\end{aligned} \tag{2.17}$$

Es posible ver que para utilizar la descomposición anterior y generar una garantía de convergencia no trivial, se necesita un resultado de estabilidad para pérdidas fuertemente convexas que funcione para valores de μ que converjan a cero cuando el número de iteraciones o el tamaño de la muestra crezcan. Se nota la dependencia del último término con respecto a la distancia al óptimo de (1.1), que bajo el Supuesto 2 puede ser acotada por r.

En el Capítulo 4 se presenta un resultado de estabilidad para el Algoritmo 8, con pérdidas fuertemente convexas, y sus consecuencias en cuanto al exceso de riesgo para el caso suave utilizando regularización. Además, en el Anexo A se formaliza la versión alternativa de Johnson y Zhang (2013) y un análisis de estabilidad para ésta.

2.6. El problema de estimación de rendimiento

Se considera un problema de optimización convexa determinístico. Sea una clase de funciones $\mathcal{F}$ sobre $\mathbb{R}^d$ y $\mathcal{M}$ un método de T iteraciones con acceso a un oráculo de primer orden $\mathcal{O}_f$. Este oráculo entrega valores y subgradientes $\mathcal{O}_f(x) = \{f(x), \nabla f(x)\}$. Sea el iterado inicial x_1 que cumple $||x_1 - x^*|| \leq R$. Sea $\mathcal{P}$ una métrica de rendimiento de un problema de optimización. El problema de obtener el peor caso de un método de descenso que se actualiza utilizando $\mathcal{M}$ sobre una función $f \in \mathcal{F}$ se conoce como el problema de estimación de rendimiento (2.18). Taylor et al. (2017b) presentan la siguiente versión dependiente de la dimensión d.

$$\begin{aligned} w(\mathcal{F}, R, M, N, \mathcal{P}) = \sup_{f, x_1, ..., x_{T+1}, x^*} \quad & \mathcal{P}(\mathcal{O}_f, x_1, ..., x_{T+1}, x^*) \\ \text{s.a} \quad & f \in \mathcal{F}, \\ & x^* \text{ óptimo de } f, \\ & x_2, ..., x_{T+1} \text{ generados desde } x_1 \text{ por} \\ & \mathcal{M} \text{ con oráculo } \mathcal{O}_f, \\ & ||x_0 - x^*|| \leq R. \end{aligned} \tag{2.18}$$

Se nota que la métrica de rendimiento tiene dependencia en el oráculo, el óptimo del problema y los iterados del método utilizado. Los autores notan que esto permite incorporar al problema las métricas comunes en el análisis de convergencia, tales como el *gap* de optimalidad, la norma del gradiente y la distancia en norma entre la respuesta y el óptimo. Sin embargo, (2.18) corresponde a un problema infinito-dimensional, dada la elección de $f \in \mathcal{F}$.

Con el propósito de obtener una versión finito-dimensional de (2.18), se hace notar que los iterados generados por una aplicación de $\mathcal{M}$ pueden ser representados por la ecuación implícita (2.19).

$$x_{t+1} = \mathcal{M}(x_1, \mathcal{O}_f(x_1), x_2, \mathcal{O}_f(x_2), ..., x_{t+1}, \mathcal{O}_f(x_{t+1})) \tag{2.19}$$

La representación de (2.19) muestra que en el proceso de optimización están involucrados una cantidad finita y conocida de puntos, cuyas instanciaciones se desconocen. Para lograr obtener un problema finito-dimensional, se necesita que la elección de función $f \in \mathcal{F}$ sea representable por aquellos valores y subgradientes que la definen en una colección finita de puntos.

2.6.1. El problema de interpolación

Sean $\{(x_i, g_i, f_i)\}_{i\in\mathcal{I}}$ un conjunto de puntos x_i, junto a una propuesta de subgradientes g_i y valores de función f_i. Dada una clase de funciones $\mathcal{F}$ sobre $\mathcal{X} \supset conv(\{x_i\}_{i\in\mathcal{I}})$, se dice que el conjunto de tripletas es $\mathcal{F}$-interpolable si existe $f \in \mathcal{F}$ tal que la elección de puntos, subgradientes y valores de la función es consistente con f. Esto es, existe $f \in \mathcal{F}$ tal que

$$\begin{cases} f_i = f(x_i), \\ g_i \in \partial f(x_i) \end{cases} \qquad (\forall i \in \mathcal{I})$$

Idealmente, la $\mathcal{F}$-interpolabilidad de un conjunto de tuplas puede ser reducida a la decisión de si una propiedad Q sobre el conjunto $\{(x_i, g_i, f_i)\}_{i\in\mathcal{I}}$ es cierta o no. En tal caso, la reducción implica que Q describe condiciones necesarias y suficientes sobre la interpolabilidad de la clase $\mathcal{F}$ y se dice que Q es un conjunto de condiciones de interpolación de $\mathcal{F}$.

Taylor et al. (2017a) presentan condiciones de interpolación para las clases $\mathcal{F}_M(\mathbb{R}^d)$ y $\mathcal{F}_{M,L}(\mathbb{R}^d)$, basadas en dualidad de Fenchel; presentadas en el Teorema 2.2. Pese a que el resultado se presenta con condiciones para la subclase suave, los autores argumentan que

puede ser extendido sin problemas a la clase $\mathcal{F}_M(\mathbb{R}^d)$ utilizando el caso límite $L = \infty$ con la convención $1/\infty = 0$.

Teorema 2.2. *El conjunto* $\{(x_i, g_i, f_i)\}_{i\in\mathcal{I}}$ *es* $\mathcal{F}_{M,L}(\mathbb{R}^d)$ *interpolable si y solo si, para todo* $i, j \in \mathcal{I}$*, se tiene*

$$\begin{cases} f_i - f_j - \langle g_j, x_i - x_j\rangle \geq \frac{1}{2L}||g_i - g_j||^2, \\ ||g_i|| \leq M. \end{cases} \tag{2.20}$$

Para esta representación finita de una función, es necesario tener en cuenta que dada una elección de iterados, x^*, gradientes y valores; la elección de función cumpliendo las condiciones de interpolación del problema es arbitraria e indistinguible en términos del PEP. En otras palabras, aunque muchas funciones de la clase $\mathcal{F}$ interpolen el conjunto de iterados y el óptimo x^* (utilizando $x_* := x^*$), esto no importa en términos del PEP, pues no requiere la unicidad de esta interpolación.

2.6.2. Versión semidefinida del problema de estimación de rendimiento

Sea la matriz de Gram $X = P^\top P$, Taylor et al. (2017a, 2017b) eligen las columnas de P de forma apropiada en (2.21) para representar la dinámica de descenso y las restricciones del problema en términos de restricciones SDP equivalentes, llegando finalmente al problema semidefinido positivo SDP-PEP (2.22).

$$P = [g_1\ g_2\ ...\ g_{T+1}\ x_1] \tag{2.21}$$

Esto permite reescribir la dinámica del problema como parte de la estructura del problema SDP-PEP incluso cuando existe una iteración implícita, como es el caso de los métodos proximales. Se hace notar que las entradas de X consisten de productos internos entre columnas de P, descartando la dimensión del problema original, mientras sea suficientemente alta. Sea el conjunto de índices $\mathcal{I} = \{1, ..., T+1, *\}$, se fijan:

$$\begin{cases} x_i = Ph_i \quad (\forall i \in \mathcal{I}) \\ g_i = Pu_i \quad (\forall i \in \mathcal{I}) \end{cases}$$

con vectores h_i y u_i apropiados, determinados para el método utilizado. Sean $i, j \in \mathcal{I}$, la primera condición de $\mathcal{F}_{ML}(\mathbb{R}^d)$-interpolación de (2.2) se reescribe de la forma

$$\begin{aligned} & f_i - f_j - \langle g_j, x_i - x_j \rangle - \frac{1}{2L}||g_i - g_j||^2 && \geq 0 \\ \iff & f_i - f_j - \langle Pu_j, P(h_i - h_j) \rangle - \frac{1}{2L}||P(u_i - u_j)||^2 && \geq 0 \\ \iff & f_i - f_j - \text{Tr}(A_{ij}X) && \geq 0 \end{aligned}$$

con $A_{ij} = \frac{1}{2}u_j(h_i - h_j)^\top + \frac{1}{2}(h_i - h_j)u_j^\top + \frac{1}{2L}(u_i - u_j)(u_i - u_j)^\top$ para todo par de índices en $\mathcal{I}$. Igualmente, las condiciones de interpolación Lipschitz continua y la condición de radio inicial se pueden reformular matricialmente, como

$$\begin{cases} \text{Tr}(A_{M_i}X) - M^2 \leq 0 \quad (\forall i \in I), \\ \text{Tr}(A_R X) - R^2 \leq 0; \end{cases}$$

donde A_{M_i} y A_R son matrices dependientes de $\{u_i\}_{i \in \mathcal{I}}$. Reformular estas cantidades en términos matriciales permite convertir el problema PEP de obtener el peor caso de optimizar una función con dominio d dimensional en un problema generalizado que no depende de tal dimensión. Por lo tanto, la versión de optimización semidefinida de PEP (2.22), donde se asume que la métrica de rendimiento $\mathcal{P}$ puede ser caracterizada por un vector $b \in \mathbb{R}^{T+1}$ y una matriz $C \in \mathbb{S}^{T+2}$, captura la idea del problema de minimización infinito-dimensional (2.18), llevándolo a una versión finito-dimensional y ampliando la búsqueda a la clase de funciones $\mathcal{F}_{M,L}$.

$$\begin{aligned} \sup_{f \in \mathbb{R}^{T+1}, X \in \mathbb{S}^{T+2}} \quad & b^\top f + \text{Tr}(CX) \\ \text{s.a} \quad & \text{Tr}(A_{ij}X) + f_j - f_i \leq 0, \quad (\forall i, j \in I) \\ & \text{Tr}(A_{M_i}X) - M^2 \leq 0, \qquad (\forall i \in I) \\ & \text{Tr}(A_R X) - R^2 \leq 0, \\ & X \succeq O. \end{aligned} \tag{2.22}$$

Se nota que soluciones factibles consisten de una asignación de valores para las pérdidas f en los distintos iterados y una matriz X con la información de los subgradientes e iterados. La interpolación garantiza la existencia de funciones que cumplan con esta asignación, teniéndose minorantes, no necesariamente ajustadas, para el óptimo buscado en (2.22). Igualmente, para el problema dual se tienen mayorantes para el peor caso de la métrica de rendimiento. Más aún, la asignación de las variables duales y ponderación con las respectivas restricciones primales permite recuperar una combinación de desigualdades que puede reinterpretarse como una demostración de cota superior para la función objetivo del primal. Sin embargo, este proceso de recuperar demostraciones no siempre es claro.

Realizando este proceso para la métrica de *gap* de optimalidad, Taylor et al. (2017a) demuestran que es posible obtener una cota superior para el método de punto proximal que es más ajustada que el resultado clásico de (2.8).

Lema 2.11. *(Taylor et al., 2017a) Sea $f(\cdot, \xi) \in \mathcal{F}_M(\mathcal{X})$ para todo $\xi \in \mathcal{Z}$ (Supuesto 1a). Luego, el método de punto proximal (Algoritmo 5) con paso $\eta > 0$ garantiza*

$$F_S(x_{T+1}) - F_S(x_S^*) \leq \frac{||x_1 - x_S^*||^2}{4\eta T}.$$

3. ESTABILIDAD DE MÉTODOS PROXIMALES

Continuando con lo expuesto en la Sección 2.3, la obtención de cotas de generalización para el método de punto proximal está limitada por la falta de un resultado de uniforme estabilidad aplicable a múltiples iteraciones. En caso de tener tal resultado, es posible obtener cotas en esperanza y de alta probabilidad para el riesgo empírico de la misma manera que se expuso en las Secciones 2.1, 2.2 y 2.4 para otros métodos, utilizando las descomposiciones de riesgo (2.8) y (2.5), respectivamente.

Más aún, Bassily et al. (2020) plantean una técnica general para la iteración de tipo gradiente que permite su aplicación a distintos métodos tales como SGD Incremental, el método de subgradiente y otros no descritos en este trabajo, incluyendo aquellos de subgradiente *minibatch*. En este sentido, se busca aprovechar las propiedades de la iteración proximal para obtener resultados de UAS para trayectorias definidas por dos sucesiones potencialmente distintas de funciones, de manera que el algoritmo general sea aplicable a los casos particulares de interés para este trabajo (PPM e *IncrementalProx*) y algoritmos *minibatch*-prox, que son de interés en la literatura porque permiten un acercamiento en arquitecturas distribuidas a los métodos proximales (Wang, Wang, y Srebro, 2017).

Formalmente, se consideran dos trayectorias generales $\{x_t\}_{t\in[T+1]}$ y $\{y_t\}_{t\in[T+1]}$, generadas por las iteraciones

$$\begin{cases} x_{t+1} = \mathbf{prox}_{\eta_t f_t}(x_t) \\ y_{t+1} = \mathbf{prox}_{\eta_t f_t'}(y_t), \end{cases}$$

utilizando una secuencia de *learning rates* $\{\eta_t\}_{t\in[T]}$. Se nota que (2.11) y (2.13) ambos pueden ser representados por este tipo de iteración, eligiendo un único *learning rate* η y las funciones empleadas en cada paso, apropiadamente. Estas se explicitan más adelante.

Con esto, se propone un resultado basado en las distancias máximas entre los gradientes de las funciones elegidas por ambas trayectorias, dada por $\sup_{x\in\mathcal{X}} ||\nabla f_t(x) - \nabla f_t'(x)||$ en cada una de las iteraciones del algoritmo. Este se presenta en la Proposición 3.1 y se demuestra posteriormente.

Proposición 3.1. *Sean $\{x_t\}_{t\in[T+1]}$ y $\{y_t\}_{t\in[T+1]}$ las trayectorias generadas a partir de un punto $x_1 = y_1$ por las iteraciones proximales definidas sobre el conjunto convexo, cerrado y no-vacío $\mathcal{X}$ y las funciones $f_t, f_t' \in \Gamma_0(\mathbb{R}^d)$ tales que dom(f_t), dom$(f_t') \supseteq \mathcal{X}$ para todo $t \in [T]$. Esto es,*

$$\begin{cases} x_{t+1} = x_t - \eta_t \nabla(f_t + \mathbb{1}_{\mathcal{X}})(x_{t+1}) \\ y_{t+1} = y_t - \eta_t \nabla(f_t' + \mathbb{1}_{\mathcal{X}})(y_{t+1}) \end{cases}$$

para todo $t \in [T]$. Suponiendo para todo $t \in [T]$, $\sup_{x\in\mathcal{X}} ||\nabla f_t(x) - \nabla f_t'(x)|| \leq a_t$. Luego, para $t_0 = \text{ínf}\{t : f_t \neq f_t'\}$, se cumple

$$||x_{T+1} - y_{T+1}|| \leq \sum_{t=t_0}^{T} \eta_t a_t.$$

DEMOSTRACIÓN. *Siguiendo la idea de los resultados de estabilidad previos, se acota la distancia entre los iterados generados por ambas trayectorias en función de iterados anteriores y cantidades conocidas:*

$$\begin{aligned} \delta_{t+1}^2 =& ||x_{t+1} - y_{t+1}||^2 \\ =& \langle [x_t - \eta\nabla(f_t + \mathbb{1}_{\mathcal{X}})(x_{t+1})] - [y_t - \eta\nabla(f_t' + \mathbb{1}_{\mathcal{X}})(y_{t+1})], x_{t+1} - y_{t+1}\rangle \end{aligned}$$

Por la convexidad del no-vacío $\mathcal{X} \subset \mathbb{R}^d$ y el hecho que los dominios de las funciones cubren $\mathcal{X}$,

$$ri(\mathcal{X}) \cap ri(dom(f_t)) \cap ri(dom(f_t')) = ri(\mathcal{X}) \neq \emptyset.$$

Luego, el subgradiente de $f_t + \mathbb{1}_{\mathcal{X}}$ corresponde a la suma de una elección de subgradientes $\nabla f_t \in \partial f_t$, $\nabla \mathbb{1}_{\mathcal{X}} \in \partial \mathbb{1}_{\mathcal{X}}$. El argumento es idéntico para el caso de f_t'. Separando ambas sumas:

$$\begin{aligned} =& \langle x_t - y_t, x_{t+1} - y_{t+1}\rangle - \eta_t \langle \nabla f_t(x_{t+1}) - \nabla f_t'(y_{t+1}), x_{t+1} - y_{t+1}\rangle \\ & - \eta_t \langle \nabla \mathbb{1}_{\mathcal{X}}(x_{t+1}) - \nabla \mathbb{1}_{\mathcal{X}}(y_{t+1}), x_{t+1} - y_{t+1}\rangle . \end{aligned}$$

Utilizando la monotonía de los subgradientes de funciones convexas y posteriormente la desigualdad de Cauchy-Schwarz sobre los términos restantes:

$$\begin{aligned}
&= \langle x_t - y_t, x_{t+1} - y_{t+1} \rangle - \eta_t \langle \nabla f_t(y_{t+1}) - \nabla f_t'(y_{t+1}), x_{t+1} - y_{t+1} \rangle \\
&\quad - \eta_t \underbrace{\langle \nabla (f_t + \mathbb{1}_{\mathcal{X}})(x_{t+1}) - \nabla (f_t + \mathbb{1}_{\mathcal{X}})(y_{t+1}), x_{t+1} - y_{t+1} \rangle}_{\geq 0} \\
&\leq \langle x_t - y_t, x_{t+1} - y_{t+1} \rangle - \eta_t \langle \nabla f_t(y_{t+1}) - \nabla f_t'(y_{t+1}), x_{t+1} - y_{t+1} \rangle \\
&\leq \delta_t \delta_{t+1} + \eta_t a_t \delta_{t+1}.
\end{aligned}$$

Se concluye que

$$\Rightarrow \delta_{t+1} \leq \delta_t + \eta_t a_t.$$

Finalmente, sumando para $t_0 \leq t \leq T$, se obtiene el resultado enunciado. □

Se considera el caso particular del método de punto proximal, caracterizado por la iteración (2.11). Se nota que en una trayectoria generada por el PPM (Algoritmo 5), cada actualización corresponde a una aplicación del operador proximal que está asociado a una misma funcion, siendo ésta el riesgo empírico $F_{\mathbf{S}}$ resultante de utilizar una cierta muestra aleatoria $\mathbf{S}$. Esto permite derivar de manera simple el siguiente resultado de estabilidad para el PPM.

Corolario 3.1. *Sea una pérdida $f(\cdot, \xi) \in \mathcal{F}_M(\mathcal{X})$ para todo $\xi \in \mathcal{Z}$ (Supuesto 1a). Luego, el Algoritmo 5 después de T iteraciones alcanza $\frac{2\eta MT}{n}$-uniforme estabilidad de argumentos.*

DEMOSTRACIÓN. *Se nota que dadas dos muestras $\mathbf{S} \simeq \mathbf{S}'$, basta fijar s.p.g. $f_t \equiv F_{\mathbf{S}}$, $f_t' \equiv F_{\mathbf{S}}'$ y $\eta_t = \eta$ para todo $t \in [T]$, pues la M-Lipschitz continuidad de las pérdidas convexas es suficiente para asegurar que el riesgo empírico pertenezca a la clase Γ_0. Además, por el uso del batch completo en cada iteración, $t_0 = 1$ y*

$$\sup_{x \in \mathcal{X}} ||\nabla F_{\mathbf{S}}(x) - \nabla F_{\mathbf{S}'}(x)|| \leq \frac{2\eta M}{n}.$$

Dado que el Algoritmo 5 es determinista y $\mathbf{S}, \mathbf{S}'$ *son elegidas arbitrariamente, se cumple el resultado enunciado.* □

El Corolario 3.1 permite acotar el exceso de riesgo de este método de la forma ya planteada en la Sección 2.3. Sin embargo, este análisis se puede mejorar utilizando el resultado de convergencia para ERM del Lema 2.11, que mejora por un factor constante la cota superior presentada en el Lema 2.8. Un cálculo del largo de paso óptimo muestra que, en términos de n y T se debe elegir $\eta \in \Theta\left(\frac{\sqrt{n}}{T}\right)$. Después de fijar $T = n$ iteraciones del algoritmo, tanto los errores de estabilidad y de optimización alcanzan una tasa de convergencia $O\left(\frac{1}{\sqrt{n}}\right)$.

Utilizando ambos resultados de descomposición del riesgo, se concluye que en esperanza el exceso de riesgo converge con tasa $O\left(\frac{1}{\sqrt{n}}\right)$ y con alta probabilidad, con tasa $\tilde{O}\left(\frac{1}{\sqrt{n}}\right)$. Sin embargo, se nota que la elección $T = \sqrt{n}$ implica una complejidad de iteraciones y de gradientes proximales mejorada $O\left(\frac{1}{\varepsilon}\right)$, manteniendo la complejidad muestral alcanzada por la elección $T = n$. Si bien esto significa una mejora en cuanto a órdenes de convergencia, introduce problemas en la ejecución de la metodología PEP. A diferencia del enfoque asintótico, el enfoque asistido por computador modela de manera exacta una cantidad entera de iteraciones del problema, lo que restringiría el análisis en la práctica a una pequeña cantidad de instancias. Estas son aquellos valores de n que son cuadrados perfectos y son suficientemente pequeños para ser computados por un PEP en tiempo razonable, considerando la gran cantidad de ejecuciones que requiere la metodología utilizada. Por este motivo, se considera para efectos de este trabajo la elección $T = n$.

Por otra parte, se destaca que la Proposición 3.1 puede ser empleada para el análisis del método *IncrementalProx*, como se planteó originalmente. Para lograr esto basta tomar $T = Kn$ iteraciones, un par de muestras aleatorias arbitrarias $\mathbf{S} \simeq \mathbf{S}'$ con componentes $\xi_1, ..., \xi_n, \xi_1'$ que, sin pérdida de generalidad difieren en la primera componente. Además, se fijan las funciones $f_t \equiv f(\cdot, \xi_{\boldsymbol{\pi}_k(i)}), f_t' \equiv f(\cdot, \xi'_{\boldsymbol{\pi}_k(i)})$ en cada iteración t parametrizable por $t = (k-1)n + i$ con $i \in [n], k \in [K]$. Luego, se nota que $a_t = 2\eta M$ en el caso que los datos de cada muestra difieren y a_t se anula si no, que ocurre exactamente una

vez en cada ronda, debido al uso de permutaciones. Por último, se toma valor esperado en la aleatoriedad del algoritmo y supremo en la elección de las muestras vecinas anteriores, obteniendo un resultado idéntico al planteado en la Proposición 2.2.

Finalmente, se presentan dos tablas que resumen las complejidades para el exceso de riesgo presentadas en los Capítulos 2 y 3, alcanzadas utilizando las cotas superiores para el riesgo empírico y la uniforme estabilidad ,y la descomposición en valor esperado de (2.5). Primero, la Tabla 3.1 resume las cotas superiores para la complejidad del exceso de riesgo alcanzado por métodos de tipo gradiente en términos del número de muestras, de iteraciones y de cálculos de subgradiente totales requeridos para alcanzar la ε-suboptimalidad. Además, se incluye una elección de *learning rate* η como función de la cantidad total de iteraciones que garantiza las complejidades expuestas.

TABLA 3.1. Cotas superiores de complejidad muestral, de iteraciones y de subgradientes en valor esperado para el exceso de riesgo alcanzado por los distintos algoritmos basados en actualizaciones de subgradiente.

Algoritmo	*Learning rate*	Muestras	Iteraciones	Subgradientes
GD suave	$O\left(\frac{1}{\sqrt{T}}\right)$	$O\left(\frac{1}{\varepsilon^2}\right)$	$O\left(\frac{1}{\varepsilon^2}\right)$	$O\left(\frac{1}{\varepsilon^4}\right)$
GD no-suave	$O\left(T^{-\frac{1}{4}}\right)$	$O\left(\frac{1}{\varepsilon^2}\right)$	$O\left(\frac{1}{\varepsilon^4}\right)$	$O\left(\frac{1}{\varepsilon^6}\right)$
Incremental SGD	$O\left(T^{-\frac{1}{4}}\right)$	$O\left(\frac{1}{\varepsilon^2}\right)$	$O\left(\frac{1}{\varepsilon^4}\right)$	$O\left(\frac{1}{\varepsilon^4}\right)$

En segundo lugar, se presenta la Tabla 3.2, que resume las cotas superiores para la complejidad del exceso de riesgo esperado alcanzado por métodos proximales, incluyendo una elección de largo de paso, tamaño del *batch* y las complejidades de gradientes proximales, iteraciones y muestras derivadas de tal elección. Tal análisis no considera el costo computacional de calcular el paso proximal exacto.

COMENTARIO 3.1. *El método de punto proximal evalúa una operación proximal sobre la suma de las n pérdidas inducidas en cada iteración. Esto se cuenta como una sola evaluación, pero en la práctica el cálculo del operador no es fácil incluso si se asume acceso inmediato a las operaciones proximales sobre las componentes. Por lo mismo, una alternativa común en la literatura aplicable a ambos métodos proximales es calcular una*

TABLA 3.2. Cotas superiores de complejidad de iteraciones y de cálculos de gradientes proximales en valor esperado para los algoritmos proximales presentados.

Algoritmo	*Learning rate*	Muestras	Iteraciones	Grad. Proximales	*batch size*
PPM	$O\left(\frac{1}{\sqrt{T}}\right)$	$O\left(\frac{1}{\varepsilon^2}\right)$	$O\left(\frac{1}{\varepsilon^2}\right)$	$O\left(\frac{1}{\varepsilon^2}\right)$	n
IncrementalProx	$O\left(T^{-\frac{1}{4}}\right)$	$O\left(\frac{1}{\varepsilon^2}\right)$	$O\left(\frac{1}{\varepsilon^4}\right)$	$O\left(\frac{1}{\varepsilon^4}\right)$	1

Δ-*aproximación en norma al operador proximal exacto presente en cada iteración con una subrutina de* $\lceil \frac{8M^2}{\Delta^2} \rceil$ *iteraciones de GD (Bassily et al., 2019) aplicada sobre la función involucrada en la actualización.*

4. ESTABILIDAD DE SVRG

En la Sección 2.5 se planteó una manera de acotar el exceso de riesgo para el problema (1.1) en el caso suave mediante la utilización de un regularizador fuertemente convexo. Siguiendo tal fin, se considera primero el caso de utilizar SVRG sobre pérdidas fuertemente convexas, presentando una garantía de cota superior para la estabilidad de SVRG con elección estocástica (Algoritmo 8) en tal *setting*. Luego, se aplica tal resultado al caso suave más regularización.

Proposición 4.1. *Sea una pérdida* $f(\cdot, \xi) \in \mathcal{S}^{\mu}_{M,L}(\mathbb{R}^d)$ *para todo* $\xi \in \mathcal{Z}$ *(Supuesto 1c) y* $m \in \Theta(n)$. *Luego:*

1. *Para* $\eta \in \left(\frac{2}{3\mu}, \frac{1}{L}\right]$, *el Algoritmo 8 tiene uniforme estabilidad de argumentos* $O\left(\frac{\eta MK}{n}\right)$.
2. *Para* $\eta \leq \frac{2}{3\mu}$, *el Algoritmo 8 tiene uniforme estabilidad de argumentos* $O\left(\frac{M}{\mu n} \cdot \left(\frac{1}{\eta\mu}\right)^K\right)$.

Este resultado presenta dos casos en los que se tienen regímenes distintos de estabilidad, dependiendo del paso elegido. Ambas partes se demuestran a continuación:

DEMOSTRACIÓN. *Se analiza una actualización del método estocástico con varianza reducida. Por simplicidad, se consideran las funciones* f_i *y* f_i' *correspondientes a las instanciaciones de la pérdida para las muestras aleatorias* $\mathbf{S} \simeq \mathbf{S}'$ *que difieren, sin pérdida de generalidad, en el primer dato. Además, se consideran los iterados* x_i^k *y* y_i^k, $\tilde{x}_k$ *y* $\tilde{y}_k$ *respectivos.*

Por definición de la actualización (2.15) *y desigualdad triangular:*

$$\begin{aligned}\delta_{t+1}^k \leq & \left|\left|x_t^k - y_t^k - \eta\left(\nabla f_{i_t}(x_t^k) - \nabla f'_{i_t}(y_t^k)\right)\right|\right| \\ & + \eta\left|\left|\left(\nabla f_{i_t}(\tilde{x}_k) - \nabla f'_{i_t}(\tilde{y}_k\right) - \left(\nabla F_S(\tilde{x}_k) - \nabla F_{S'}(\tilde{y}_k)\right)\right|\right|.\end{aligned}$$

Sumando cero convenientemente se puede utilizar desigualdad triangular de nuevo, obteniendo tres términos propios de los análisis de SGD con muestreo con reemplazo y GD:

$$\begin{aligned}
&\leq \left|\left|x_t^k - y_t^k - \eta\left(\nabla f_{i_t}(x_t^k) - \nabla f'_{i_t}(y_t^k)\right)\right|\right| \\
&\quad + \left|\left|\tilde{x}_k - \tilde{y}_k - \eta\left(\nabla f_{i_t}(\tilde{x}_k) - \nabla f'_{i_t}(\tilde{y}_k\right)\right|\right| \\
&\quad + \left|\left|\tilde{x}_k - \tilde{y}_k - \eta\left(\nabla F_S(\tilde{x}_k) - \nabla F_{S'}(\tilde{y}_k)\right)\right|\right|.
\end{aligned}$$

Por la fuerte convexidad de las pérdidas, para $\eta \leq \frac{1}{L}$ se cumple la $(1-\eta\mu)$-expansividad de una actualización de gradiente. Aplicando esta propiedad a los términos anteriores, se tiene

$$\delta_{t+1}^k \leq (1-\eta\mu)\delta_t^k + 2\eta M \boldsymbol{r}_{\boldsymbol{i}_{k,t}} + 2(1-\eta\mu)\delta_1^k + 2\eta M \boldsymbol{r}_{\boldsymbol{i}_{k,t}} + \frac{2\eta M}{n},$$

donde $\boldsymbol{r}_{\boldsymbol{i}_{k,t}}$ es una variable aleatoria que toma valor 1 si la componente elegida en la iteración t de la $k-$ésima ronda difiere o cero en caso contrario. Luego, tomando valor esperado sobre la aleatoriedad del algoritmo se obtiene la siguiente recursión, para $t \geq 2$:

$$\mathbb{E}\delta_{t+1}^k \leq (1-\eta\mu)\mathbb{E}\delta_t^k + 2(1-\eta\mu)\mathbb{E}\delta_1^k + \frac{6\eta M}{n}.$$

Además, se nota que la primera iteración de cada ronda es equivalente a un paso del método de gradiente, cumpliendose la desigualdad

$$\mathbb{E}\delta_2^k \leq (1-\eta\mu)\mathbb{E}\delta_1^k + \frac{2\eta M}{n}.$$

Sumando las desigualdades sobre todo el ciclo interior del algoritmo, se obtiene una cota superior para cada iterado de la ronda en términos de $\mathbb{E}\delta_1^k$. En particular, para $i \in [m]$:

$$\begin{aligned}
\mathbb{E}\delta_{i+1}^k &\leq (1-\eta\mu)^i \mathbb{E}\delta_1^k + (1-\eta\mu)^{i-1}\frac{2\eta M}{n} + \left[2(1-\eta\mu)\mathbb{E}\delta_1^k + \frac{6\eta M}{n}\right]\sum_{j=0}^{i-2}(1-\eta\mu)^j \\
&\leq \left[\frac{2(1-\eta\mu)}{\eta\mu} - \frac{2-\eta\mu}{\eta\mu}(1-\eta\mu)^i\right]\mathbb{E}\delta_1^k + \frac{6\eta M}{n}\sum_{j=0}^{i-1}(1-\eta\mu)^j
\end{aligned}$$

Por lo tanto, dada la elección aleatoria uniforme del primer iterado para la ronda siguiente, se tiene:

$$\begin{aligned}
\mathbb{E}\delta_1^{k+1} &= \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\delta_i^k \\
&\leq \underbrace{\left[\frac{1}{m} + \frac{2(m-1)(1-\eta\mu)}{m\eta\mu} - \frac{(2-\eta\mu)(1-\eta\mu)(1-(1-\eta\mu)^{m-1})}{m\eta^2\mu^2}\right]}_{\phi_1}\mathbb{E}\delta_1^k \\
&\quad + \frac{6\eta M}{n}\cdot\frac{1}{m}\sum_{i=1}^{m}\sum_{j=0}^{i-2}(1-\eta\mu)^j. \\
&\leq \max\left\{\phi_1(m,\eta,\mu), \frac{2(1-\eta\mu)}{\eta\mu}\right\}\mathbb{E}\delta_1^k + \frac{6\eta M}{n}\cdot\frac{1}{m}\sum_{i=1}^{m}\sum_{j=0}^{i-2}(1-\eta\mu)^j.
\end{aligned}$$

Calculando la suma del segundo término:

$$\begin{aligned}
&= \max\left\{\phi_1(m,\eta,\mu), \frac{2(1-\eta\mu)}{\eta\mu}\right\}\mathbb{E}\delta_1^k + \frac{6\eta M}{mn}\sum_{i=1}^{m}\frac{1-(1-\eta\mu)^{i-1}}{\eta\mu} \\
&= \max\left\{\phi_1(m,\eta,\mu), \frac{2(1-\eta\mu)}{\eta\mu}\right\}\mathbb{E}\delta_1^k + \frac{6M}{mn\mu}\sum_{i=1}^{m}[1-(1-\eta\mu)^{i-1}] \\
&= \max\left\{\phi_1(m,\eta,\mu), \frac{2(1-\eta\mu)}{\eta\mu}\right\}\mathbb{E}\delta_1^k + \underbrace{\frac{6M}{mn\mu}\left(m - \frac{1-(1-\eta\mu)^m}{\eta\mu}\right)}_{\phi_2}
\end{aligned}$$

Como $m = \Theta(n)$, se nota que para valores crecientes de n el factor que acompaña a la esperanza del primer término converge por arriba hacia la cantidad $2(1-\eta\mu)/(\eta\mu)$, con η y μ fijos. Se separan dos casos:

I) Sea $\eta > \frac{2}{3\mu}$.Entonces, por la monotonía de la función $\frac{2(1-x)}{x}$, el factor que multiplica el valor esperado es menor a uno para valores grandes de n. Con esta nueva restricción, se puede calcular la estabilidad de la ejecución completa,

$$\mathbb{E}\delta_1^{k+1} \leq \phi_1\mathbb{E}\delta_1^k + \phi_2.$$

Luego, se concluye

$$\Rightarrow \mathbb{E}\delta_1^{K+1} \leq K \cdot \phi_2$$

Notando que $\phi_2 \in O\left(\frac{M}{\mu n}\right)$ *y el learning rate* $\eta \in \Omega\left(\frac{1}{\mu}\right)$*, se deduce una uniforme estabilidad de argumentos* $O\left(\frac{\eta M K}{n}\right)$.

II) Alternativamente, consideramos $\eta \leq \frac{2}{3\mu}$*. En tal caso, una valuación en el caso límite permite ver que el factor* ϕ_1 *que acompaña al valor esperado proveniente de la ronda anterior es mayor a uno. Consecuentemente, un cálculo directo muestra que esto implica crecimiento exponencial de la estabilidad. Se calcula la distancia de las trayectorias para la ejecución completa*

$$\mathbb{E}\delta_1^{k+1} \leq \phi_1 \mathbb{E}\delta_1^k + \phi_2.$$

Desenvolviendo la recursión,

$$\Rightarrow \mathbb{E}\delta_1^{K+1} \leq \phi_2 \sum_{k=0}^{K-1} \phi_1^k.$$

Finalmente, utilizando el hecho que $\phi_2 \in O\left(\frac{M}{\mu n}\right)$*, se obtiene una cota de orden* $O\left(\frac{M}{\mu n} \cdot \left(\frac{1}{\eta\mu}\right)^K\right)$ *para la estabilidad de argumentos.* □

Por lo tanto, se puede concluir de la Proposición 4.1 que la cota para la estabilidad es de un orden mejor si el paso tomado es suficientemente grande. Ambas cotas en el caso μ-fuertemente convexo (para μ fijo) proveen convergencia para una elección de largo de paso η independiente de n. Sin embargo, la elección de un producto $\eta\mu$ dependiente de n resulta en una dependencia exponencial del resultado de estabilidad. Estas consideraciones son necesarias al analizar el fenómeno de regularización, donde μ es una cantidad introducida en el problema que se desea hacer desvanecer para una cantidad creciente de iteraciones o muestras.

Además, ya que se quiere utilizar el resultado de estabilidad en la descomposición (2.17), se deben considerar las restricciones propias de la garantía de convergencia a utilizar. En particular, se utilizará el Lema 2.10, por lo que la elección de $m = \Theta(n)$ es apropiada según lo discutido en la Sección 2.5. Se recuerda el requisito para η y m presente en el Lema de convergencia 2.10:

$$\frac{1}{\mu\eta m(1-2L\eta)} + \frac{2L\eta}{1-2L\eta} < 1.$$

Un cálculo explícito del intervalo en que el *learning rate* cumple esta condición indica que esta cantidad debe tomar valores en $\left(\frac{1-\sqrt{1-\frac{16L}{\mu m}}}{8L}, \frac{1+\sqrt{1-\frac{16L}{\mu m}}}{8L}\right)$. La dependencia de L del intervalo anterior implica que el régimen de estabilidad en el que funciona el resultado para el error de optimización depende del condicionamiento del problema.

Asimismo, el problema (1.1) suave con regularización se restringe debido al comportamiento de la estabilidad para distintos números de condicionamiento. Por la demostración de estabilidad de la Proposición 4.1, la elección del largo de paso debe ser tal que $\eta\mu \leq \frac{\mu}{L} < 1$. Recordando la descomposición (2.17), una garantía que asegure la convergencia del exceso de riesgo requiere de $\mu \overset{n}{\to} 0$ para valores de n crecientes.

Para cualquier elección de μ anterior, considerando ambos regímenes de estabilidad provistos por la Proposición 4.1, se nota que la elección $\eta = \Omega\left(\frac{1}{\mu}\right)$ contradice el requisito $\eta \leq \frac{1}{L}$ del resultado de estabilidad. Esto es, cualquier elección de η en el primer caso $\eta \geq \frac{2}{3\mu}$; y en el segundo, elecciones de orden $\eta = \Theta\left(\frac{1}{\mu}\right)$. Por el contrario, una elección tal que $\eta\mu \to 0$ para el segundo caso implica que los términos de estabilidad y el término adicional de la regularización entregan cotas para los errores de orden

$$O\left(\frac{M}{\mu n} \cdot \left(\frac{1}{\eta\mu}\right)^K + \mu r^2\right),$$

donde cualquier elección de μ implica que la complejidad muestral alcanzada es peor, en términos de orden de convergencia, que la alcanzada en valor esperado por SGD, $O\left(\frac{1}{\sqrt{n}}\right)$.

5. UN MODELO PEP PARA ALGORITMOS BATCH

En este capítulo se presenta una formulación semidefinida de una versión del problema de estimación de rendimiento que modela los errores de optimización y estabilidad para el problema (1.3). Esta formulación es especifica para métodos *batch*, como los presentados en las Secciones 2.1 y 2.3.

La idea principal consiste en la formulación del problema (5.1) como aquel que busca el peor caso, en términos de una métrica de rendimiento que engloba los fenómenos de optimización y estabilidad, de dos trayectorias definidas por muestras vecinas $\mathbf{S} \simeq \mathbf{S}'$. Dado que cada muestra define n pérdidas aleatorias, donde $n-1$ de estas son compartidas por ambas trayectorias, es posible representar este fenómeno mediante tres funciones $f, f', F \in \mathcal{F}$, donde f y F representan una muestra $\mathbf{S}$ sobre la cual se minimiza y f' corresponde a la perturbación asociada a utilizar la muestra vecina $\mathbf{S}'$, cuantificando la peor estabilidad posible para el par y la dificultad de minimizar el riesgo asociado a f y F (según alguna métrica para funciones convexas, como en Taylor et al. (2017b, 2017a)). Sin pérdida de generalidad, se asume que estas muestras aleatorias difieren en el primer dato, cumpliéndose

$$\begin{cases} f & \equiv f(\cdot, \boldsymbol{\xi}_1), \\ f' & \equiv f(\cdot, \boldsymbol{\xi}'_1), \\ F & \equiv \frac{1}{n-1} \sum_{i=2}^{n} f(\cdot, \boldsymbol{\xi}_i). \end{cases}$$

Formalmente, $\frac{1}{n} f + \frac{n-1}{n} F$ es el funcional de riesgo empírico inducido por la muestra $\mathbf{S}$ y $\frac{1}{n} f' + \frac{n-1}{n} F$, por la muestra $\mathbf{S}'$. El problema (5.1) interpola las tres funciones anteriores separadamente, generando las trayectorias $\{x_i\}_{i \in I}$ y $\{y_i\}_{i \in I}$ sobre los funcionales de riesgo empírico antes mencionados a contar de un punto inicial x_1, potencialmente diferenciándose a contar de la primera actualización debido a la utilización de la muestra completa. Así, este programa entrega la peor elección (o en su defecto, el supremo) de iterados, valores y subgradientes de las tres funciones de manera de maximizar una métrica

de rendimiento $\mathcal{P}$, utilizando respuestas válidas del oráculo $\mathcal{O}_{(\cdot)}$, manteniendo la pertenencia de las pérdidas a una clase objetivo $\mathcal{F}$ y validando que corresponde a una aplicación del método $\mathcal{M}$.

$$\begin{aligned} w = \sup_{\mathcal{O}_f, \mathcal{O}_{f'}, \mathcal{O}_F, \{x_i\}_{i\in I}, \{y_i\}_{i\in I}} \quad & \mathcal{P}(\mathcal{O}_f, \mathcal{O}_{f'}, \mathcal{O}_F, \{x_i\}_{i\in I}, \{y_i\}_{i\in I}) \\ \text{s.a} \quad & f, f', F \in \mathcal{F} \\ & x_* \text{ óptimo de } \frac{1}{n} f + \frac{n-1}{n} F \\ & x_{i+1}, y_{i+1} \text{ generados por } \mathcal{M} \text{ desde } x_1 \quad (\forall i \in [T]) \\ & \mathcal{M} \text{ un método con acceso a un oráculo } \mathcal{O}_{(\cdot)} \\ & x_1 \text{ satisface la condición de inicialización } \mathcal{C}. \end{aligned} \tag{5.1}$$

5.1. Una Formulación Semidefinida para métodos *batch*

Tras formular el problema finito, se plantea un modelo semidefinido de manera que represente el mismo problema de estimación de rendimiento cuando la métrica, condiciones de interpolación, actualización del método utilizado y condición inicial son representables a través de funciones y restricciones propias de un problema SDP. Para esto se debe tener en cuenta que se desea elegir valores y subgradientes

$$\begin{cases} f_i = f(x_i), f_i' = f'(y_i), F_i = F(x_i), F_i' = F(y_i) & (\forall i \in \mathcal{I}) \\ g_i = \nabla f(x_i), g_i' = \nabla f'(y_i), G_i = \nabla F(x_i), G_i' = \nabla F(y_i) & (\forall i \in \mathcal{I}). \end{cases} \tag{5.2}$$

Taylor et al. (2017a) plantean esta idea de equivalencia entre el problema semidefinido y el problema abstracto de dimensión finita mediante la noción de Gram-representabilidad lineal para un problema de tipo PEP que considera una sola trayectoria. Se presenta una noción de Gram-representabilidad acorde al problema de optimización más estabilidad en el contexto de minimización de riesgo empírico, que extiende la noción original de Taylor et al. a emular dos trayectorias provenientes de muestras vecinas.

$$P = \left[g_1 \dots g_{T+1}\, G_1 \dots G_{T+1}\, g_1' \dots g_{T+1}'\, G_1' \dots G_{T+1}'\, \bar{g}\, \bar{G}\, \bar{g}'\, \bar{G}'\, g_*\, x_1\, y\right] \tag{5.3}$$

Siguiendo la idea presentada en la Sección 2.6, se plantea en (5.3) una matriz P cuyas columnas contienen la información de subgradientes e iterados involucrados en ambas trayectorias vecinas emuladas por el problema semidefinido, junto a un vector auxiliar y. Se define la matriz de Gram asociada a tales columnas de la forma $X = P^\top P$, representando la información anterior mediante los productos de vectores en cada una de sus entradas y eliminando la dependencia de la dimensión base d del problema convexo subyacente. Esto permite la búsqueda de un peor caso para las clases de funciones extendidas a dimensión finita presentadas en el Capítulo 2, descartando la variable d como una parámetro del PEP al igual que en (2.22), mientras la dimensión sea suficientemente alta. Además, se llama al vector v como aquel que tiene por coordenadas los valores de las funciones f, f' y F en los iterados relevantes para ambas trayectorias descritas por el problema.

$$v = \left[f_1 \dots f_{T+1}\, F_1 \dots F_{T+1}\, f_1' \dots f_{T+1}'\, F_1' \dots F_{T+1}'\, \bar{f}\, \bar{F}\, \bar{f}'\, \bar{F}'\, f_*\, F_*\right] \tag{5.4}$$

Las variables v y X; como elección de valores de funciones, iterados y subgradientes; corresponden a las variables del problema semidefinido en las que se realiza la optimización para buscar el peor caso. Se presenta en (5.5) un modelo semidefinido general asociado a esta elección de variables, que abarca la clase completa de problemas de estimación de rendimiento representables por SDP utilizando solamente las descripciones presentadas de v y X. En este, se representan directamente los valores tomados por las valuaciones de las funciones. No así los gradientes e iterados, donde el peor caso puede alcanzarse para cualquier orientación de estos vectores, teniéndose una invarianza bajo rotaciones. Por lo mismo, la información necesaria se puede definir a través de correlaciones entre estas cantidades vectoriales y la relación implícita que existe entre ellos de acuerdo a las actualizaciones del método.

$$\begin{aligned} \sup_{v\in\mathbb{R}^{4T+10}, X\in\mathbb{S}^{4T+11}} \quad & b^\top v + \text{Tr}(CX) \\ \text{s.a} \quad & a_j^\top v + \text{Tr}(M_j X) + c_j \leq 0, \quad (j \in J) \\ & X \succeq O. \end{aligned} \tag{5.5}$$

Si bien el problema puede ser definido para un espacio r-acotado $\mathcal{X}$ mediante una alteración de las restricciones del problema, la ejecución de los problemas semidefinidos primales y duales se hace notoriamente más lenta. Consecuentemente, por restricciones del tiempo de desarrollo de este trabajo, se utiliza la simplificación al problema en el *setting* irrestricto. Luego, se introduce la noción de *batch Stability-and-Optimization* linealmente Gram representable (abr. bSOLG-representable) y se propone un resultado de equivalencia entre los problemas (5.1) y (5.5) cuando el primero es bSOLG-representable, ejemplificando paralelamente que el problema para el método de subgradiente y para el método de punto proximal irrestrictos cumplen esta propiedad.

Definición 5.1. *Un método de primer orden es bSOLG-representable si y solo si el cómputo de sus iterados, definidos (posiblemente de manera implícita) por* (2.19)*, puede ser expresada utilizando una cantidad finita de restricciones lineales con dependencias solo de* $X \in \mathbb{S}^{4T+10}$ *y* $v \in \mathbb{R}^{4T+11}$.

Se nota que para el caso irrestricto la regla de actualización (2.9) puede ser representada utilizando la igualdad

$$x_{i+1} = x_i - \eta\left(\frac{1}{n}g_i + \frac{n-1}{n}G_i\right) \quad (\forall i \in [T]) \tag{5.6}$$

a través de las condiciones (5.2) deseadas para este problema. Sin embargo, no basta con obtener una recurrencia basada en la igualdad anterior, ya que las restricciones propias de un problema PEP semidefinido no cuentan con una representación explícita de las columnas de P. La representación de los iterados y subgradientes a través de la matriz de Gram X indica que estas equivalencias deben cumplirse al interior de productos de traza. Con este fin, se definen cantidades vectoriales $h_{(\cdot)}$ y $u_{(\cdot)}$ tales que representan puntos de interés

en el problema y el subgradiente de una elección de función en tales puntos.

$$\begin{cases} x_i = Ph_i & (\forall i \in [T+1]) \\ g_i = Pu_i & (\forall i \in [T+1]). \\ G_i = Pu_{j_i} & (\forall i \in [T+1]). \end{cases}$$

Lo anterior es un ejemplo para el exceso de riesgo F_S. De manera más general, se establece la necesidad de encontrar vectores $h_{(\cdot)}$ y $u_{(\cdot)}$ para ambas trayectorias x_i e y_i, los iterados promedio respectivos y toda instanciación de funciones en estos puntos. Además, se necesita la instanciación de f y F el punto x_*, incluido en las columnas de P y necesario para la representación de la métrica del error de optimización. A continuación, se explicitan la elección de $h_{(\cdot)}$ y $u_{(\cdot)}$.

Afirmación 5.1. *El método de gradiente y de subgradiente (Algoritmos 1 y 2) son bSOLG-representables.*

DEMOSTRACIÓN. *Se considera la secuencia de valores de las funciones en v. Se determina h_i de manera que la elección de $x \in \mathbb{R}^d$ sea consistente con el valor v_i. Esto es, para $y_1 = x_1$:*

$$Ph_i = \begin{cases} x_i & , i \in [T+1] \\ x_{i-T-1} & , i \in [T+2, 2T+2] \\ y_{i-2T-2} & , i \in [2T+3, 3T+3] \\ x_{i-3T-3} & , i \in [3T+4, 4T+4] \\ \bar{x} & , i \in \{4T+5, 4T+6\} \\ \bar{y} & , i \in \{4T+7, 4T+8\} \\ 0 & , i \in \{4T+9, 4T+10\} \end{cases}$$

Luego, utilizando (5.6) *se definen* h_i *de la forma*

$$h_i = h_{i+T+1} = e_{4T+10} - \eta \sum_{j=1}^{i-1} \left(\frac{1}{n} e_j + \frac{n-1}{n} e_{j+T+1} \right) \qquad (i \in [T+1])$$

$$h_i = h_{i+T+1} = e_{4T+10} - \eta \sum_{j=2T+3}^{i-1} \left(\frac{1}{n} e_j + \frac{n-1}{n} e_{j+T+1} \right) \quad (i \in [2T+3, 3T+3])$$

$$h_{4T+5} = h_{4T+6} = \sum_{i=2}^{T+1} h_i$$

$$h_{4T+7} = h_{4T+8} = \sum_{i=2}^{T+1} h_{i+2T+2}$$

$$h_{4T+9} = h_{4T+10} = 0,$$

donde, sin pérdida de generalidad se asume que $x_* = 0$*. Los vectores* $u_{(\cdot)}$ *se determinan de la misma manera, esta vez considerando los gradientes asociados a cada entrada de* v_i*. Se nota que para* $i \in [4T+9]$ *se puede elegir* $u_i = e_i$*. Debido a la condición de optimalidad de* F_S *en* x_**, se elige*

$$u_{4T+10} = -\frac{1}{n-1} e_{4T+9}.$$

Entonces, el método de gradiente es bSOLG-representable notando el hecho que $\mathbf{Tr}(ab^\top X) = (Pa)^\top (Pb)$ *permite la representación de las elecciones de* $u_{(\cdot)}$ *y* $h_{(\cdot)}$ *al interior de los productos de traza, incluyendo representaciones para el iterado final y promedio de iterados. Se concluye que esta representación semidefinida, que no depende cual sea la respuesta del método y representa ambos* x_{T+1} *y* $\bar{x}$ *, también es válida para el método de subgradiente.* □

Asimismo, para la actualización del método de punto proximal irrestricto (2.12) es equivalente a la resultante de la expresión

$$x_{i+1} = x_i - \eta \left(\frac{1}{n} g_{i+1} + \frac{n-1}{n} G_{i+1} \right) \quad (\forall i \in [T]), \tag{5.7}$$

y puede ser representada al interior de los productos de traza para una elección distinta de vectores multiplicadores h_i y u_i, debido a la elección distinta de gradientes en cada paso.

Afirmación 5.2. *El método de punto proximal (Algoritmo 5) es bSOLG-representable.*

DEMOSTRACIÓN. *Se considera nuevamente la secuencia de valores de las funciones en v. Se determina h_i de la misma manera que en la Afirmación 5.1, esta vez con la actualización de* (5.7)*:*

$$h_i = h_{i+T+1} = e_{4T+10} - \eta \sum_{j=1}^{i-1} \left(\frac{1}{n} e_{j+1} + \frac{n-1}{n} e_{j+T+2} \right) \qquad (i \in [T+1])$$

$$h_i = h_{i+T+1} = e_{4T+10} - \eta \sum_{j=2T+3}^{i-1} \left(\frac{1}{n} e_{j+1} + \frac{n-1}{n} e_{j+T+2} \right) \quad (i \in [2T+3, 3T+3])$$

$$h_{4T+5} = h_{4T+6} = \sum_{i=2}^{T+1} h_i$$

$$h_{4T+7} = h_{4T+8} = \sum_{i=2}^{T+1} h_{i+2T+2}$$

$$h_{4T+9} = h_{4T+10} = 0.$$

La elección de vectores $u_{(\cdot)}$ es idéntica al caso del método de subgradiente. Nuevamente, la elección de ambos $h_{(\cdot)}$ y $u_{(\cdot)}$ permite la representación del método al interior de los productos de traza. □

Teniendo una manera de representar los iterados y gradientes elegidos por el modelo, se puede considerar la representación semidefinida de las condiciones y métrica del problema. Partiendo por las condiciones asociadas a la interpolación, se nota que ésta debe realizarse entre todo par de puntos perteneciente a una misma función f, f' o F. La función a la cual pertenece cada índice se fija según el orden de las coordenadas del vector v, como se aprecia en (5.4), y la asignación para la interpolación de (5.2). Se sigue que los

índices que interpolan a la función f corresponden al conjunto

$$I_f = [T+1] \cup \{4T+5, 4T+9\},$$

aquellos que interpolan a la muestra perturbada f' son

$$I_{f'} = [2T+3, 3T+3] \cup \{4T+7\}$$

y los correspondientes a la función de referencia F son

$$I_F = [T+2, 2T+2] \cup [3T+4, 4T+4] \cup \{4T+6, 4T+8, 4T+10\}.$$

Contando con los conjuntos de índices anteriores, se define por $\mathcal{I}$ el conjunto de pares de interpolación de los índices pertenecientes a una misma función. Esto es,

$$\mathcal{I} = I_f \times I_f \cup I_F \times I_F \cup I_{f'} \times I_{f'}.$$

Consecuentemente, se puede utilizar el conjunto de pares $\mathcal{I}$ para representar el conjunto de restricciones de interpolación del problema (5.1) que dependen de un par de índices y por $I_M = [4T+10]$ se denota a los índices en los que se definen $u_{(\cdot)}$ y $h_{(\cdot)}$, aquellos utilizados para las restricciones de interpolación dependientes de un solo índice. A continuación, se define una noción de bSOLG-representabilidad para las restricciones del problema, mostrando que se puede hacer el símil de las restricciones de interpolación de (2.22) para la representación semidefinida del nuevo problema.

Definición 5.2. *Una clase de funciones es bSOLG-representable si y solo si sus condiciones de interpolación pueden ser reformuladas utilizando una cantidad finita de restricciones lineales dependiendo solo de $X \in \mathbb{S}^{4T+10}$ y $v \in \mathbb{R}^{4T+11}$.*

En particular, las clases F_M y $F_{M,L}$ son relevantes para la simplificación al caso irrestricto introducida previamente y son bSOLG-representables de acuerdo a la Afirmación 5.3, siguiendo el proceso introducido en la Subsección 2.6.2.

Afirmación 5.3. *Sea un método $\mathcal{M}$ bSOLG-representable a través de los iterados Ph_i y gradientes Pu_i para todo $i \in [4T + 10]$. Entonces, las clases F_M y $F_{M,L}$ son bSOLG-representables.*

DEMOSTRACIÓN. *Se demuestra que las restricciones de interpolación de f, f' y F pertenecientes a la clase $\mathcal{F}_{M,L}(\mathbb{R}^d)$ son representables de forma lineal en X y v para una elección arbitraria de d. Por el Teorema 2.2, la elección de coordenadas* (5.4) *y las representaciones de iterados Ph_i y Pu_i, se cumplen los siguientes conjuntos de restricciones respectivos a las tres funciones expresadas, denotados en función de la matriz de columnas P*

$$\begin{cases} v_i - v_j - \langle Pu_j, Ph_i - Ph_j \rangle \geq \frac{1}{2L}||Pu_i - Pu_j||^2 & (\forall i, j \in I_f) \\ ||Pu_i|| \leq M & (\forall i \in I_f) \end{cases}$$

$$\begin{cases} v_i - v_j - \langle Pu_j, Ph_i - Ph_j \rangle \geq \frac{1}{2L}||Pu_i - Pu_j||^2 & (\forall i, j \in I_{f'}) \\ ||Pu_i|| \leq M & (\forall i \in I_{f'}) \end{cases}$$

$$\begin{cases} v_i - v_j - \langle Pu_j, Ph_i - Ph_j \rangle \geq \frac{1}{2L}||Pu_i - Pu_j||^2 & (\forall i, j \in I_F) \\ ||Pu_i|| \leq M & (\forall i \in I_F). \end{cases}$$

Luego, el conjunto de restricciones de interpolación puede separarse en una condición de convexidad suave sobre los pares $\mathcal{I}$ y una condición de M-Lipschitz continuidad sobre los índices I_M

$$\begin{cases} v_i - v_j - \langle Pu_j, Ph_i - Ph_j \rangle \geq \frac{1}{2L}||Pu_i - Pu_j||^2 & (\forall (i, j) \in \mathcal{I}) \\ ||Pu_i|| \leq M & (\forall i \in I_M). \end{cases}$$

Además, por la definición de X como una matriz de Gram se puede representar la primera desigualdad por un producto de traza. Además, la segunda desigualdad puede reescribirse de manera equivalente por una restricción para el cuadrado de la norma, la que a la vez puede reescribirse por un producto de traza. Entonces, definiendo las matrices

$$A_{ij} = \frac{1}{2}u_j(h_i - h_j)^\top + \frac{1}{2}(h_i - h_j)u_j^\top + \frac{1}{2L}(u_i - u_j)(u_i - u_j)^\top$$
$$A_{M_i} = u_i u_i^\top,$$

se tiene la equivalencia con el conjunto de restriccciones lineales en X y v

$$\begin{cases} v_j - v_i + \mathbf{Tr}\,(A_{ij}X) \leq 0 & (\forall (i,j) \in \mathcal{I}) \\ \mathbf{Tr}(A_{M_i}X) - M^2 \leq 0 & (\forall i \in I_M), \end{cases}$$

terminando la demostración para funciones suaves, para cualquier valor de d. Para el caso no-suave, se considera la condición de interpolación convexa con el caso límite $L = +\infty$ para todos los pares. Esto presenta una simplificación de la matriz A_{ij} anterior, donde el término de normas de gradiente se anula. Por lo tanto, basta tomar la misma representación lineal, redefiniendo las matrices

$$A_{ij} = \frac{1}{2}u_j(h_i - h_j)^\top + \frac{1}{2}(h_i - h_j)u_j^\top.$$

□

COMENTARIO 5.1. *La Afirmación 5.3 indica que cualquier método bSOLG-representable a través de sólo el uso de vectores $h_{(\cdot)}$ y $u_{(\cdot)}$ permite obtener un conjunto de condiciones de interpolación bSOLG-representables para las clases $\mathcal{F}_M$ y $\mathcal{F}_{M,L}$, siendo este el caso para los métodos de punto proximal y de gradiente. En otras palabras, cualquier método batch (y sus condiciones de interpolación para pérdidas en alguna clase anterior) que utilice solamente información de los gradientes en la secuencia de iterados, el óptimo del riesgo empírico original y el promedio de los iterados puede ser representado en forma semidefinida.*

Adicionalmente, métodos que utilicen una respuesta distinta a las consideradas y se actualicen solo con la secuencia de iterados pueden representarse mediante una elección distinta de los vectores $h_{(\cdot)}$ correspondientes a $\bar{x}$ e $\bar{y}$.

COMENTARIO 5.2. *Para obtener condiciones de interpolación para el método proyectado a un conjunto r-acotado $\mathcal{X}$ se necesita agregar una cuarta función al problema 5.1, que corresponde a la indicatriz de $\mathcal{X}$. Esto significa una elección distinta de columnas de P de manera que el problema semidefinido guarde la información de la nueva función que se desea interpolar. A la vez, en las restricciones de interpolación se introduce una restricción de radio para los iterados, una modificación de $h_{(\cdot)}$ y $u_{(\cdot)}$ para incluir los gradientes proximales asociados a la proyección del método sobre la indicatriz convexa y una restricción en los valores que toman los iterados al interior de $\mathcal{X}$.*

Por otra parte, el problema semidefinido (5.5) necesita que una métrica $\mathcal{P}$ válida sea representable a través de solo una función. Se presenta la noción de bSOLG-representabilidad para una métrica de rendimiento de acuerdo con lo anterior:

Definición 5.3. *Una métrica de rendimiento es bSOLG-representable si y solo si puede ser expresada como una función lineal de $X \in \mathbb{S}^{4T+10}$ y $v \in \mathbb{R}^{4T+11}$.*

Una ventaja de definir la representación semidefinida a través de una única función lineal en X y v es que algunas métricas compuestas pueden ser definidas como la suma de al menos dos métricas simples. Considérese el caso de los fenómenos de optimización y estabilidad. Una forma de obtener una métrica de rendimiento válida consiste en sumar, con ponderadores apropiados, una métrica de rendimiento asociada al error de optimización y una asociada al peor caso de la estabilidad uniforme de argumentos del método (Esto es, la peor elección de muestras vecinas en términos de la norma).

En base a lo anterior, se obtienen métricas separadas para los fenómenos de optimización del riesgo empírico y estabilidad. Se consideran dos métricas distintas para el error de optimización.

Afirmación 5.4. *El gap de optimalidad para el último iterado $F_S(x_{T+1}) - F_S(x_*)$ es bSOLG-representable.*

La métrica solo depende de valores de las muestras, por lo que la representación semidefinida no depende del método elegido. Además, se puede representar mediante un producto interno entre v y el vector

$$b_{last} = \frac{1}{n}e_{T+1} + \frac{n-1}{n}e_{2T+2} - \frac{1}{n}e_{4T+9} - \frac{n-1}{n}e_{4T+10}. \tag{5.8}$$

Asimismo, la métrica del *gap* de optimalidad para el promedio de iterados también puede representarse en forma semidefinida mediante un producto interno. En esta instancia la elección de ponderadores está dada por

$$b_{av} = \frac{1}{n}e_{4T+5} + \frac{n-1}{n}e_{4T+6} - \frac{1}{n}e_{4T+9} - \frac{n-1}{n}e_{4T+10}, \tag{5.9}$$

deduciéndose la garantía expuesta en la Afirmación 5.5.

Afirmación 5.5. *El gap de optimalidad para el iterado promedio $F_S(\bar{x}) - F_S(x_*)$ es bSOLG-representable.*

Por otra parte, se consideran distintas métricas para el error de estabilidad basadas en estabilidad de argumentos, cuyas representaciones son circunstanciales a la elección de método. Tales métricas dependen explícitamente de la respuesta del método, que puede ser el iterado final o el promedio simple de los iterados. En cualquier caso, la dinámica de elección de gradientes en las actualizaciones del método debe estar codificada en los coeficientes que acompañan a la variable X.

Afirmación 5.6. *La distancia entre últimos iterados $M||x_{T+1} - y_{T+1}||$ generada por el método de gradiente (Algoritmo 1) es una métrica de rendimiento bSOLG-representable.*

DEMOSTRACIÓN. *Se utiliza la variable auxiliar y para representar la norma como un supremo de producto interno. Aunque se puede utilizar como métrica el cuadrado de la*

norma, se prefiere que la métrica de estabilidad elegida sea comparable con la métrica de minimización de riesgo al sumarlas, a modo de poder analizar los fenómenos de estabilidad y optimización en conjunto. Se nota que es posible reescribir la norma como sigue, utilizando el supuesto $||y|| \leq 1$*:*

$$\sup_{\mathcal{O}_f,\mathcal{O}_{f'},\mathcal{O}_F,\{x_i\}_{i\in I},\{y_i\}_{i\in I}} ||x_{T+1} - y_{T+1}|| = \sup_{\mathcal{O}_f,\mathcal{O}_{f'},\mathcal{O}_F,\{x_i\}_{i\in I},\{y_i\}_{i\in I}} \sup_{||y||\leq 1} \langle x_{T+1} - y_{T+1}, y\rangle.$$

Por la regla de actualización (2.9)*, se pueden descomponer los iterados como*

$$\langle x_{T+1} - y_{T+1}, y\rangle = \frac{\eta}{n} \left\langle \sum_{i=1}^{T} g_i' + (n-1)G_i' - g_i - (n-1)G_i, y \right\rangle. \tag{5.10}$$

Luego, por la Afirmación 5.1 existe una representación para los gradientes de la forma $Pu_{(\cdot)}$*. Además, notando que el vector auxiliar es una columna de P se tiene* $y = Pe_{4T+11}$*. Por lo tanto, se puede representar cada producto entre* y *y alguna otra columna de* X *mediante un producto de traza entre* X *y una matriz con una sola entrada no-nula. Por linealidad, se obtiene una matriz* $C_{GD} \in \mathbb{M}^{4T+11}$ *simétrica tal que el producto interior de* (5.10) *es representado por una forma equivalente Tr*$(C_{GD}X)$*.*

Explícitamente, $C_{GD} = (C_{ij})_{i,j\in[4T+11]}$ *toma valores*

$$C_{ij} = \begin{cases} -\frac{M\eta}{2n} & (i = 4T+11, j \in [T]) \\ -\frac{M\eta}{2n} & (j = 4T+11, i \in [T]) \\ -\frac{M\eta(n-1)}{2n} & (i = 4T+11, j \in [T+2, 2T+1]) \\ -\frac{M\eta(n-1)}{2n} & (j = 4T+11, i \in [T+2, 2T+1]) \\ \frac{M\eta}{2n} & (i = 4T+11, j \in [2T+3, 3T+2]) \\ \frac{M\eta}{2n} & (j = 4T+11, i \in [2T+3, 3T+2]) \\ \frac{M\eta(n-1)}{2n} & (i = 4T+11, j \in [3T+4, 4T+3]) \\ \frac{M\eta(n-1)}{2n} & (j = 4T+11, i \in [3T+4, 4T+3]) \\ 0 & e.o.c. \end{cases} \tag{5.11}$$

□

COMENTARIO 5.3. *Se nota que cualquier restricción agregada debe ser representable como una restricción lineal dependiente de sólo* X *y* v *para poder lograr representar el problema PEP abstracto* (5.1) *como un modelo semidefinido de la forma* (5.5)*. Este es el caso de* $||y|| \leq 1$*, como se demuestra más adelante, considerándola una condición de inicialización del problema PEP.*

El método de gradiente utiliza como respuesta el último iterado, pero la estabilidad del método de subgradiente se cuantifica según el iterado promedio. Esto, no obstante, no afecta los gradientes utilizados en las actualizaciones, sino como se ponderan estos. Luego, se introduce un segundo resultado de bSOLG-representabilidad para actualizaciones de tipo subgradiente, esta vez, aplicable al caso no-suave.

Afirmación 5.7. *La distancia entre respuestas del método de subgradiente (Algoritmo 2)* $M||\bar{x} - \bar{y}||$ *es una métrica de rendimiento bSOLG-representable.*

DEMOSTRACIÓN. *Nuevamente se utiliza la variable auxiliar y para reescribir la norma. Por el supuesto de la restricción $||y|| \leq 1$:*

$$\sup_{\mathcal{O}_f, \mathcal{O}_{f'}, \mathcal{O}_F, \{x_i\}_{i\in I}, \{y_i\}_{i\in I}} ||\bar{x} - \bar{y}|| = \sup_{\mathcal{O}_f, \mathcal{O}_{f'}, \mathcal{O}_F, \{x_i\}_{i\in I}, \{y_i\}_{i\in I}} \sup_{||y||\leq 1} \langle \bar{x} - \bar{y}, y\rangle.$$

Por la regla de actualización (2.9)*, se descomponen los iterados promedio como*

$$\begin{aligned}
\langle \bar{x} - \bar{y}, y\rangle &= \frac{1}{T}\sum_{t=1}^{T}\langle x_{t+1} - y_{t+1}, y\rangle \\
&= \frac{\eta}{nT}\sum_{t=1}^{T}\sum_{i=1}^{t}\langle g_i' + (n-1)G_i' - g_i - (n-1)G_i, y\rangle \\
&= \frac{\eta}{n}\sum_{i=1}^{T}\frac{T-i+1}{T}\langle g_i' + (n-1)G_i' - g_i - (n-1)G_i, y\rangle .
\end{aligned}$$

Luego, por la Afirmación 5.1, existe una representación para los gradientes de la forma $Pu_{(\cdot)}$. Además, la representación matricial $y = Pe_{4T+11}$ del vector auxiliar, se puede representar cada producto entre un gradiente e y mediante un producto de traza. Por linealidad, se obtiene otra matriz $C_{avGD} \in \mathbb{M}^{4T+11}$ simétrica. Esta vez, la matriz posee valores

$$= \mathit{Tr}(C_{avGD}X),$$

donde $C_{avGD} = (C_{ij})_{i,j\in[4T+11]}$ *toma valores*

$$C_{ij} = \begin{cases} -\frac{M\eta}{2n} \cdot \frac{T-j+1}{T} & (i = 4T+11, j \in [T]) \\ -\frac{M\eta}{2n} \cdot \frac{T-i+1}{T} & (j = 4T+11, i \in [T]) \\ -\frac{M\eta(n-1)}{2n} \cdot \frac{2T-j+2}{T} & (i = 4T+11, j \in [T+2, 2T+1]) \\ -\frac{M\eta(n-1)}{2n} \cdot \frac{2T-i+2}{T} & (j = 4T+11, i \in [T+2, 2T+1]) \\ \frac{M\eta}{2n} \cdot \frac{3T-j+3}{T} & (i = 4T+11, j \in [2T+3, 3T+2]) \\ \frac{M\eta}{2n} \cdot \frac{3T-i+3}{T} & (j = 4T+11, i \in [2T+3, 3T+2]) \\ \frac{M\eta(n-1)}{2n} \cdot \frac{4T-j+4}{T} & (i = 4T+11, j \in [3T+4, 4T+3]) \\ \frac{M\eta(n-1)}{2n} \cdot \frac{4T-i+4}{T} & (j = 4T+11, i \in [3T+4, 4T+3]) \\ 0 & e.o.c. \end{cases} \tag{5.12}$$

□

En tercer lugar, se presenta un resultado para el método de punto proximal (Algoritmo 5) que entrega por respuesta el iterado resultante después de T iteraciones.

Afirmación 5.8. *La distancia entre últimos iterados* $M||x_{T+1} - y_{T+1}||$ *generada por el método de punto proximal (Algoritmo 5) es una métrica de rendimiento bSOLG-representable.*

DEMOSTRACIÓN. *Nuevamente se utiliza la variable auxiliar* y *para reescribir la norma. Por el supuesto de la restricción* $||y|| \leq 1$*:*

$$\sup_{\mathcal{O}_f, \mathcal{O}_{f'}, \mathcal{O}_F, \{x_i\}_{i\in I}, \{y_i\}_{i\in I}} ||x_{T+1} - y_{T+1}|| = \sup_{\mathcal{O}_f, \mathcal{O}_{f'}, \mathcal{O}_F, \{x_i\}_{i\in I}, \{y_i\}_{i\in I}} \sup_{||y||\leq 1} \langle x_{T+1} - y_{T+1}, y\rangle.$$

Entonces, por la regla de actualización (2.12)*, se pueden descomponer los iterados como*

$$\langle x_{T+1} - y_{T+1}, y\rangle = \frac{\eta}{n} \left\langle \sum_{i=1}^{T} g'_{i+1} + (n-1)G'_{i+1} - g_{i+1} - (n-1)G_{i+1}, y \right\rangle.$$

Luego, por la Afirmación 5.2 existe una representación para los gradientes de la forma $Pu_{(\cdot)}$*. Además, nutilizando la representación* $y = Pe_{4T+11}$ *para el vector auxiliar, se puede reescribir cada producto entre un gradiente e* y *mediante un producto de traza. Por linealidad, se obtiene una matriz* $C_{PPM} \in \mathbb{M}^{4T+11}$ *simétrica tal que*

$$= \mathit{Tr}(C_{PPM}X).$$

Explícitamente, $C_{PPM} = (C_{ij})_{i,j\in[4T+11]}$ *toma valores*

$$C_{ij} = \begin{cases} \frac{M\eta}{2n} & (i = 4T+11, j \in [2, T+1]) \\ \frac{M\eta}{2n} & (j = 4T+11, i \in [2, T+1]) \\ \frac{M\eta(n-1)}{2n} & (i = 4T+11, j \in [T+3, 2T+2]) \\ \frac{M\eta(n-1)}{2n} & (j = 4T+11, i \in [T+3, 2T+2]) \\ -\frac{M\eta}{2n} & (i = 4T+11, j \in [2T+4, 3T+3]) \\ -\frac{M\eta}{2n} & (j = 4T+11, i \in [2T+4, 3T+3]) \\ -\frac{M\eta(n-1)}{2n} & (i = 4T+11, j \in [3T+5, 4T+4]) \\ -\frac{M\eta(n-1)}{2n} & (j = 4T+11, i \in [3T+5, 4T+4]) \\ 0 & e.o.c. \end{cases} \tag{5.13}$$

□

En definitiva, las afirmaciones anteriores permiten elaborar métricas de rendimiento sumando las métricas de optimización y estabilidad correspondientes a la actualización y respuesta del método. Así, métricas de la forma

$$F_{\mathbf{S}}(\mathcal{A}(\mathbf{S})) - F_{\mathbf{S}}(x_*) + M||\mathcal{A}(\mathbf{S}) - \mathcal{A}(\mathbf{S}')||$$

pueden ser bSOLG-representadas mediante la inclusión de una condición de inicialización adicional, teniéndose para cada método elecciones de ponderadores para X y v. En

particular, se tiene para el método de gradiente la representación $b_{last}^{\top}v + \text{Tr}(C_{GD}X)$, para el método de subgradiente $b_{av}^{\top}v + \text{Tr}(C_{avGD}X)$ y para el método de punto proximal $b_{last}^{\top}v + \text{Tr}(C_{PPM}X)$.

A continuación, se define el concepto de restricción bSOLG-representable para las condiciones de inicialización y se prueba que el conjunto de restricciones conveniente para la representación de métricas de estabilidad es bSOLG-representable.

Definición 5.4. *Un conjunto de condiciones de inicialización es bSOLG-representable si y solo si puede ser reformulado utilizando una cantidad finita de restricciones lineales dependiendo solo de $X \in \mathbb{S}^{4T+10}$ y $v \in \mathbb{R}^{4T+11}$.*

Afirmación 5.9. *El conjunto de condiciones iniciales*

$$\begin{cases} ||y|| & \leq 1 \\ ||x_1 - x_*|| & \leq R \end{cases}$$

es bSOLG-representable.

DEMOSTRACIÓN. *Por la elección de columnas de P, se tiene $x_1 = Pe_{4T+10}$ e $y = Pe_{4T+11}$. Dado que ambas restricciones pueden ser reescritas como una raiz cuadrada de un producto interno, se hace la representación semidefinida en base a los cuadrados de las restricciones, utilizando productos de traza y notando que $x_* = 0$. Para $A_R = e_{4T+10}e_{4T+10}^{\top}$ y $A_y = e_{4T+11}e_{4T+11}^{\top}$:*

$$\begin{cases} \textit{Tr}(A_y X) - 1 & \leq 0 \\ \textit{Tr}(A_R X) - R^2 & \leq 0. \end{cases}$$

□

Utilizando las definiciones anteriores de bSOLG-representabilidad, se puede probar que cualquier problema definido de la forma (5.1) con componentes bSOLG-representables puede ser llevado a la forma semidefinida general (5.5) que utiliza la información de solo X y v elegidos. Más aún, de la Proposición 5.1, presentada a continuación, se deduce que las cantidades que se demostraron bSOLG-representables todas tienen la forma (5.14) para la elección de parámetros correspondiente al método elegido.

Proposición 5.1. *Sea un método de primer orden $\mathcal{M}$, una clase de funciones $\mathcal{F}$, una métrica de rendimiento $\mathcal{P}$ y un conjunto de condiciones de inicialización $\mathcal{C}$ bSOLG-representables. Luego, el problema* (5.1) *de representar de manera exacta el peor caso de $\mathcal{P}$ alcanzado por ejecuciones de $\mathcal{M}$ sobre una pérdida perteneciente a la clase $\mathcal{F}$, con condiciones de inicialización $\mathcal{C}$ puede ser reformulado como un problema semidefinido de la forma* (5.5) *para dimensiones $d \geq 4T + 11$.*

DEMOSTRACIÓN. *La bSOLG-representabilidad de las partes del problema está dada por las Definiciones 5.1-5.4 y mostrada en las afirmaciones previas, donde se nota que todas las restricciones y la función objetivo pueden ser reformuladas como expresiones lineales de X y v, que son las únicas variables del problema semidefinido* (5.5).

Para mostrar la implicancia contraria, se prueba que la equivalencia se cumple solo para dimensiones altas. Se nota que la matriz de Gram X admite una descomposición que utiliza la matriz P de dimensiones $d \times (4T + 11)$. Sea una dimensión $d < 4T + 11$, una elección arbitraria sobre el cono semidefinido de tamaño $4T + 11$ puede incurrir en respuestas factibles de X cuya descomposición $P^\top P$ contenga una matriz P con rango mayor a d, lo que se contrapone a la elección inicial de una matriz de dimensiones $d \times (4T + 11)$.

COMENTARIO 5.4. *Para la elección de clase de funciones $\mathcal{F}_{M,L}$ o $\mathcal{F}_M$ se fija la dimensión mínima interpolable $d_0 = 4T + 11$.*

$$
\begin{aligned}
\sup_{v\in\mathbb{R}^{4T+10},X\in\mathbb{S}^{4T+11}} \quad & b^\top v + \mathrm{Tr}(CX) \\
\text{s.a} \quad & \mathrm{Tr}(A_{ij}X) + v_j - v_i \leq 0, \quad (\forall i,j \in \mathcal{I}) \\
& \mathrm{Tr}(A_{M_i}X) - M^2 \leq 0, \qquad (\forall i \in I_M) \\
& \mathrm{Tr}(A_R X) - R^2 \leq 0, \\
& \mathrm{Tr}(A_y X) - 1 \leq 0.
\end{aligned} \tag{5.14}
$$

Un cálculo del problema dual de (5.14) muestra la forma de este como el problema semidefinido (5.15). La prueba del par primal-dual se expone en el Anexo B.1.

$$
\begin{aligned}
\inf_{\lambda_{ij},\mu_i,\tau,\kappa} \quad & \tau R^2 + \sum_{i\in I_M} \mu_i M^2 + \kappa \\
\text{s.a} \quad & \sum_{(i,j)\in\mathcal{I}} \lambda_{ij} A_{ij} + \sum_{i\in I_M} \mu_i A_{M_i} + \tau A_R + \kappa A_y - C \succeq O \\
& \sum_{(i,j)\in\mathcal{I}} \lambda_{ij}(e_j - e_i) = b \\
& \lambda_{ij}, \mu_i \geq 0 \quad ((i,j) \in \mathcal{I}) \\
& \tau, \kappa \geq 0.
\end{aligned} \tag{5.15}
$$

Contando con formas semidefinidas explícitas para el par de problemas, una pregunta natural es si para todos los métodos y métricas empleados se puede garantizar que no exista un *gap* de dualidad entre ambos. Probar que el valor óptimo de ambos problemas es igual cobra importancia dentro de la metodología PEP ya que, a la vez que el problema primal representa una búsqueda de peor caso en los posibles certificados (iterados, gradientes y valores de funciones) de valores convexos, las variables duales representan ponderadores aplicados a las restricciones primales para elaborar una demostración (Taylor et al., 2017a), como se ejemplifica para el método proximal en la Subsección 5.2.3. Así, una demostración de cota superior válida, recuperada con la metodología PEP, será ajustada al peor caso cuando el *gap* de dualidad es nulo. La Proposición 5.2 presenta la dualidad fuerte para el método de gradiente.

Proposición 5.2. *Sea* (5.14) *bSOLG-representación del problema de peor caso conjunto de optimización y estabilidad de una ejecución de T pasos del método de gradiente. Sea* (5.15) *su dual semidefinido, donde ambos problemas son factibles. Entonces, ambos poseen un valor óptimo idéntico y existe un punto primal-factible que alcanza tal valor.*

DEMOSTRACIÓN. *Se demuestra la dualidad fuerte mediante un certificado para la condición de Slater en el problema dual. Se nota que la matriz*

$$S = \tau A_R + \kappa A_y + \sum_{i \in I_M} \mu_i A_{M_i}$$

es una matriz diagonal. Se buscan valores apropiados de las variables duales de manera que la condición semidefinida del problema (5.15) *se cumpla de manera estricta, esto es, el lado izquierdo de la desigualdad sea una matriz positiva definida y adicionalmente se cumpla la igualdad vectorial. En pos de esto, se calculan los valores propios de C_{GD} notando que un cálculo simple entrega los máximos valores propios en valor absoluto*

$$\pm \frac{M\eta}{2n}\sqrt{2T(n^2 - 2n + 2)}.$$

Además, se acota la norma de las matrices asociadas a la interpolación convexa suave. Sea $(i, j) \in \mathcal{I}$:

$$\begin{aligned} ||A_{ij}|| &= \sup_{||z||=1} z^\top A_{ij} z \\ &= \sup_{||z||=1} \left[\langle u_j, z\rangle \langle h_i - h_j, z\rangle + \frac{1}{2L}\langle u_i - u_j, z\rangle^2 \right] \\ &\leq ||h_i - h_j|| + \frac{1}{L}. \end{aligned}$$

Luego, fijando los valores

$$\varepsilon_{ij} = \begin{cases} \frac{1}{|\mathcal{I}|L}\left(||h_i - h_j|| + \frac{1}{L}\right)^{-1} & (i \neq j) \\ \frac{1}{|\mathcal{I}|L} & (i = j), \end{cases}$$

$$\lambda_{ij} = \begin{cases} \frac{1}{n} + \varepsilon_{ij} & (i = 4T+9, j = T+1) \\ \frac{n-1}{n} + \varepsilon_{ij} & (i = 4T+10, j = 2T+2) \\ \varepsilon_{ij} & \textit{e.o.c.}, \end{cases}$$

se cumple la restricción de igualdad vectorial a b_{last}*, definido en* (5.8)*. Resta demostrar el cumplimento estricto de la restricción semidefinida, por lo que se acota superiormente la norma, utilizando desigualdad triangular y notando los hechos* $||A_{4T+9,T+1}||, ||A_{4T+10,2T+2}|| \leq ||h_{T+1}||$ *se tiene*

$$\begin{aligned} \left|\left|\sum_{(i,j)\in\mathcal{I}} \lambda_{ij}A_{ij} - C_{GD}\right|\right| &\leq ||h_{T+1}|| + \sum_{(i,j)\in\mathcal{I}, i\neq j} \varepsilon_{ij}||A_{ij}|| + ||C_{GD}|| \\ &< ||h_{T+1}|| + \frac{2}{L} + \frac{M\eta}{2n}\sqrt{2T(n^2-2n+2)} \\ &= \underbrace{\sqrt{2\eta^2 T\frac{n^2-2n+2}{n^2}+1} + \frac{2}{L} + \frac{M\eta}{2n}\sqrt{2T(n^2-2n+2)}}_{B}, \end{aligned}$$

donde en la desigualdad estricta utiliza el cálculo del valor propio máximo de C *y los valores fijados para las variables lambda. Finalmente, se considera* B *como la cota superior estricta a la norma, que corresponde a una función solo de los parámetros del problema. Basta fijar el resto de las variables duales con valor* B *para concluir la prueba con la desigualdad estricta*

$$S + \sum_{(i,j)\in\mathcal{I}} \lambda_{ij} - C_{GD} \succ O.$$

□

COMENTARIO 5.5. *Se nota que la prueba anterior puede extenderse fácilmente a los casos de tomar una métrica de solo gap de optimalidad o de solo estabilidad. Para el primero, basta considerar la matriz nula $C = O$, por lo que la elección de B anterior para el nuevo caso es válida.*

En cambio, para la segunda se está en el caso $b = 0$. Esto altera la definición de las variables duales lambda, pudiendo fijarse en este caso los valores $\lambda_{ij} = \varepsilon_{ij}$ para todo par de interpolación (i, j), de manera de cumplirse la igualdad vectorial. Nuevamente, tomar la elección de B para el problema conjunto permite obtener un punto factible estricto, con una cota superior estricta válida que aún puede ser refinada.

El mismo resultado puede replicarse para los métodos de subgradiente y de punto proximal, ambos aplicados sobre pérdidas no necesariamente suaves. Se presentan las garantías respectivas en las Proposiciones 5.3 y 5.4, relegando sus pruebas al Anexo B.2.

Proposición 5.3. *Sea* (5.14) *bSOLG-representación del problema de peor caso conjunto de optimización y estabilidad de una ejecución de T pasos del método de subgradiente. Sea* (5.15) *su dual semidefinido, donde ambos problemas son factibles. Entonces, ambos poseen un valor óptimo idéntico y existe un punto primal-factible que alcanza tal valor.*

Proposición 5.4. *Sea* (5.14) *bSOLG-representación del problema de peor caso conjunto de optimización y estabilidad de una ejecución de T pasos del método de punto proximal. Sea* (5.15) *su dual semidefinido, donde ambos problemas son factibles. Entonces, ambos poseen un valor óptimo idéntico y existe un punto primal-factible que alcanza tal valor.*

Se concluye que todos los PEP semidefinidos *batch* de interés, correspondientes a representaciones de métricas de rendimiento útiles para medir el error de estabilidad y/o error de optimización en el peor caso para alguno de los métodos *batch* presentados, poseen garantías de dualidad fuerte. Más aún, la condición de Slater en el problema dual nos

dice que los problemas primales respectivos admiten una solución óptima, interpretándose como un la existencia de un certificado óptimo para el problema primal.

Sin embargo, no se tiene claridad respecto a la existencia de un certificado de Slater para el problema primal, lo que indicaría la existencia de un conjunto de variables duales óptimas, como es el caso de optimización convexa de Taylor et al. (2017b, Teorema 6). Pese a no tener tal garantía, la búsqueda de las variables duales óptimas (o en su defecto, cercanas al óptimo) puede realizarse apoyada en información de una implementación computacional.

5.2. Implementación de PEP *batch*

Siguiendo la metodología expuesta, se implementan inicialmente los modelos semidefinidos *batch* desarrollados utilizando el *solver* de optimización MOSEK. Esto corresponde a la implementación del problema primal para cada uno de los métodos *batch* expuestos, analizando el problema dual para algunos casos de mayor interés. Se presentan los resultados obtenidos para cada método, de manera secuencial.

5.2.1. Método de gradiente

Se analizan los resultados obtenidos de la implementación del problema primal para el método de gradiente en el *setting* suave (Supuesto 1b), tomando los problemas semidefinidos relacionados con el error de optimización, de generalización mediante estabilidad y el error conjunto basado en las descomposiciones del exceso de riesgo presentadas en el Capítulo 2.

Primero, se considera la representación de peor caso del error de optimización por sí solo. Dado que la respuesta entregada por el Algoritmo 1 es el último iterado, se utiliza el *gap* de optimalidad para tal respuesta y su cota respectiva provista por el Lema 2.3:

$$F_{\mathbf{S}}(x_{T+1}) - F_{\mathbf{S}}(x^*_{\mathbf{S}}) \leq \frac{R^2}{2\eta T}.$$

Como se discutió en la Sección 2.1, una elección de *learning rate* $\eta = \Theta\left(\frac{\sqrt{n}}{T}\right)$ garantiza un exceso de riesgo de orden $O\left(\frac{1}{\sqrt{n}}\right)$ para el problema conjunto de optimización y estabilidad. Aunque el problema de optimización presente una convergencia más rápida para largo de paso constante, se consideran $T = n$ iteraciones del algoritmo y una elección óptima para la cota superior conjunta $\eta = \frac{2}{\sqrt{n}}$, según lo explicado en el Capítulo 3. Así, en la Figura 5.1 se muestra una comparación entre el peor caso computado por el modelo PEP *batch* semidefinido y la cota superior provista por la teoría clásica, dependiendo ambas del tamaño de la muestra tomada.

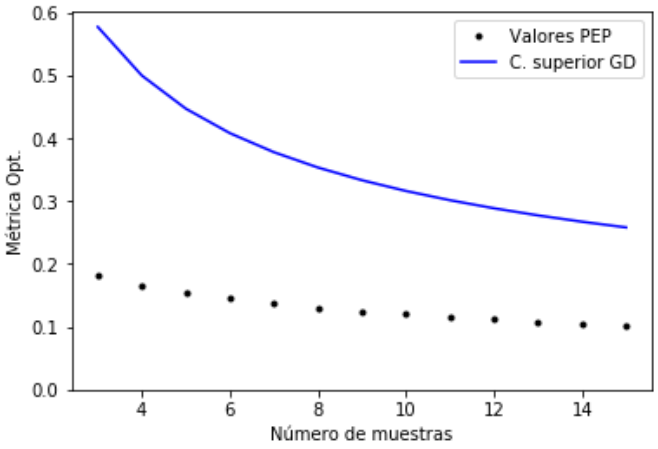

FIGURA 5.1. Comparación del peor caso computado del riesgo empírico con la cota superior teórica después de $T = n$ iteraciones del método de gradiente.

En la Figura 5.1 se aprecia una diferencia notoria entre los valores del *gap* de optimalidad y la mayorante del Lema 2.3. Esta diferencia es de un orden comparable con el valor del problema y no puede atribuirse a imprecisiones en la optimización hecha por computador. Por lo tanto, existe una diferencia entre la cota teórica del error de optimización y la respuesta entregada por el PEP.

En segundo lugar, se considera el problema semidefinido que representa el peor caso del error de estabilidad solamente. Este se representa a través de la relación entre las nociones de uniforme estabilidad y uniforme estabilidad de argumentos, la que se puede

cuantificar de acuerdo a la Afirmación 2.1 con la desigualdad:

$$M||x_{T+1} - y_{T+1}|| \leq \frac{2M^2\eta T}{n}.$$

Utilizando las elecciones de cantidades T y η previas, la peor elección de muestras vecinas en cuanto a la estabilidad uniforme de argumentos alcanza la cota superior de Hardt et al., salvo un error despreciable y atribuible al cálculo del óptimo para el problema semidefinido basado en el método de punto interior, según según se expone en la Figura 5.2.

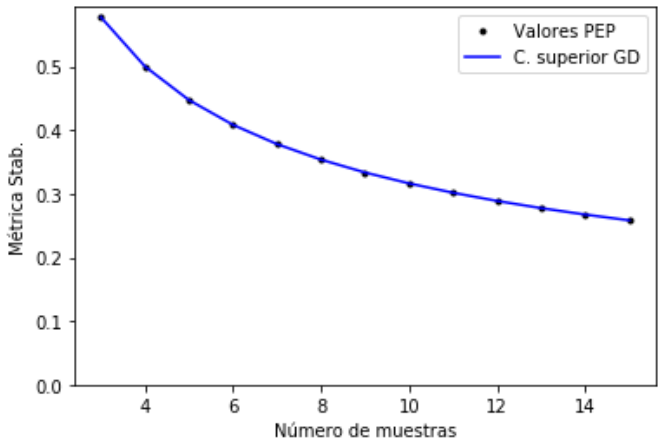

FIGURA 5.2. Comparación del peor caso computado de la uniforme estabilidad con la cota superior teórica después de $T = n$ iteraciones del método de gradiente.

El hecho que el problema primal sea ajustado a la cota, junto al resultado de dualidad fuerte de la Proposición 5.2 sugieren la existencia de una secuencia puntos dual-factibles cuyo límite se ajusta a la cota teórica ya expuesta. Un acercamiento heurístico mediante la introducción de restricciones extra al problema dual podría permitir encontrar soluciones simbólicas arbitrariamente cercanas al óptimo u óptimas (sin tener garantías de la existencia de las últimas). Nótese que tener una solución simbólica óptima puede derivar una demostración no necesariamente igual a la del resultado de estabilidad de la Afirmación 2.1.

Por último, se analiza el problema conjunto utilizando la combinación de las desigualdades provenientes del Lema 2.3 y la Afirmación 2.1:

$$F_{\mathbf{S}}(x_{T+1}) - F_{\mathbf{S}}(x^*_{\mathbf{S}}) + M||x_{T+1} - y_{T+1}|| \leq \frac{R^2}{2\eta T} + \frac{2M^2\eta T}{n}.$$

Como se presenta en la Figura 5.3, nuevamente existe una diferencia notoria entre la métrica y la cota superior conjunta, que no puede atribuirse solo a la tolerancia del método de optimización semidefinida. Esto es una consecuencia natural de utilizar la cota superior como la suma de las mayorantes a ambos fenómenos por separado. Más aún, una interrogante derivada de utilizar el problema conjunto es si la elección de cada peor caso por separado es estrictamente peor que la elección de peor caso hecha por el problema conjunto. Una mirada a los valores obtenidos (presentados en la Tabla 5.1) sugiere que tomar el supremo sobre la métrica conjunta no altera el valor del problema.

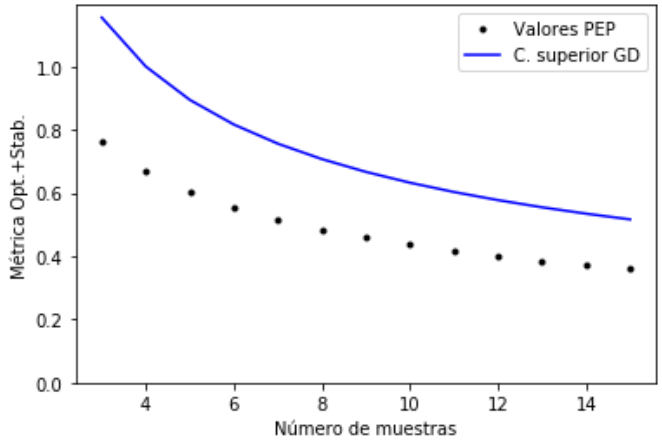

FIGURA 5.3. Comparación del peor caso computado de la métrica conjunta de riesgo empírico más estabilidad algorítmica con la cota superior teórica después de $T = n$ iteraciones del método de gradiente.

5.2.2. Método de subgradiente

Se analizan los resultados obtenidos de la implementación del problema primal para el método de subgradiente para el caso de pérdidas no-suaves (Supuesto 1a), tomando los problemas semidefinidos relacionados con el error de optimización, de generalización mediante estabilidad y el error conjunto basado en las descomposiciones del exceso de riesgo

TABLA 5.1. Diferencia entre la suma de los problemas de optimización y estabilidad por separado, y el problema conjunto (Δ) para el problema de peor caso del método de gradiente con distintos tamaños de muestra.

n	Δ
3	$2{,}29 \times 10^{-5}$
4	$2{,}66 \times 10^{-5}$
5	$3{,}53 \times 10^{-5}$
6	$1{,}66 \times 10^{-5}$
7	$2{,}94 \times 10^{-6}$
8	$4{,}16 \times 10^{-6}$
9	$3{,}45 \times 10^{-6}$
10	$5{,}03 \times 10^{-6}$
11	$4{,}37 \times 10^{-6}$
12	$4{,}38 \times 10^{-6}$
13	$1{,}87 \times 10^{-7}$
14	$1{,}12 \times 10^{-5}$
15	$1{,}04 \times 10^{-5}$

presentadas en el Capítulo 2. Debido a limitaciones computacionales y el gran tamaño del problema que se computa, los resultados presentados abarcan solo valores pequeños del número de muestras n.

En primer lugar, se considera la representación de peor caso del error de optimización por sí solo. Dado que la respuesta entregada por el Algoritmo 2 es el promedio de los iterados $\bar{x}$, esta vez se utiliza el *gap* de optimalidad para tal respuesta y la cota respectiva proveniente del Lema 2.2:

$$F_{\mathbf{S}}(\bar{x}) - F_{\mathbf{S}}(x^*_{\mathbf{S}}) \leq \frac{R^2}{2\eta T} + \frac{M^2\eta}{2}.$$

Según lo expuesto en la Sección 2.1, una elección de *learning rate* $\eta = \Theta\left(\frac{1}{T^{3/4}}\right)$ garantiza un exceso de riesgo de orden $O\left(\frac{1}{\sqrt{n}}\right)$ para el problema conjunto de optimización y estabilidad después de $T = n^2$ iteraciones. Se fija tal cantidad de iteraciones del algoritmo y una elección de largo de paso $\eta = \frac{1}{T^{3/4}}$, en vista de lo discutido en el Capítulo 2. Así, el gráfico de la Figura 5.4 muestra una comparación entre el peor caso computado por el modelo PEP *batch* semidefinido y la cota superior de la convergencia del Algoritmo 2 (Lema 2.2) con respecto al número de iteraciones, teniéndose solamente el cálculo cuando

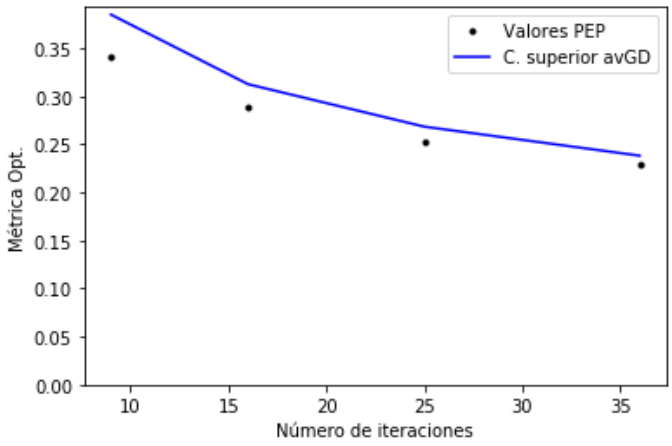

FIGURA 5.4. Comparación del peor caso computado del riesgo empírico con la cota superior teórica después de $T = n^2$ iteraciones del método de subgradiente.

T es un cuadrado perfecto. La figura muestra una diferencia relevante entre el peor caso real y la cota teórica.

En segundo lugar, se considera la representación de peor caso de la estabilidad de argumentos del algoritmo, manteniendo las elecciones de T y η anteriores. Por convexidad sobre los iterados y desigualdad triangular, se tiene que el iterado promedio cumple la

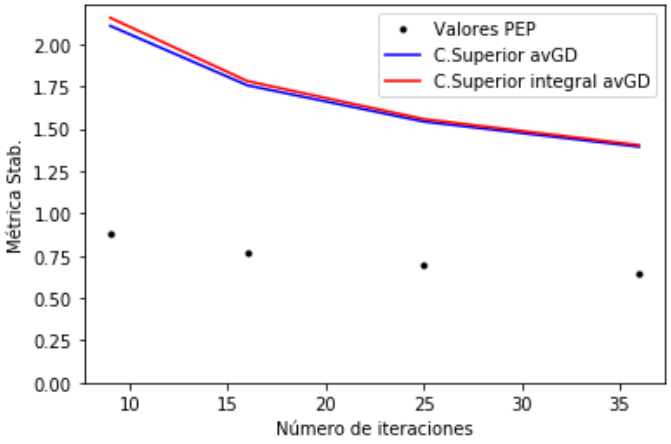

FIGURA 5.5. Comparación del peor caso computado de la uniforme estabilidad con las cotas superiores teóricas después de $T = n^2$ iteraciones del método de subgradiente.

siguiente desigualdad, siendo una consecuencia del Lema 2.4:

$$M||\bar{x} - \bar{y}|| \leq \frac{2M^2\eta(T+1)}{n} + \frac{2M^2\eta}{T} \cdot \sum_{t=1}^{T} \sqrt{T}.$$

A la vez, el lado derecho puede aproximarse por una expresión integral, tras hacer el cálculo de la integral definida resulta la cota alternativa

$$M||\bar{x} - \bar{y}|| \leq \frac{2M^2\eta(T+1)}{n} + \frac{4M^2\eta}{3T}\left[(T+1)^{3/2} - 1\right],$$

que contiene una expresión más simple, pero menos ajustada para el lado derecho. Se grafican ambas cotas superiores y el valor real de peor caso en la Figura 5.5, mostrando nuevamente una diferencia notoria entre las mayorantes y el resultado del PEP. El hecho de utilizar como respuesta el iterado promedio introduce una nueva dificultad, la falta de ajuste real de la cota puede deberse tanto a la cota para el último iterado del Lema 2.4, como a la derivación del resultado para el iterado promedio mediante convexidad.

Luego, se analiza el peor caso de los errores de optimización y estabilidad en conjunto. Esta vez se consideran dos cotas para este problema derivadas de combinar el resultado del Lema 2.2 con ambas expresiones presentadas para el análisis del problema de estabilidad. De acuerdo con la Figura 5.6, ambas cotas superiores presentan un desajuste esperado del análisis para el problema de estabilidad de la Figura 5.5.

Además, un análisis de los valores obtenidos por los distintos problemas PEP relativos a este método muestra la existencia de una diferencia notoria entre los peores casos separados de optimización y estabilidad, y el problema conjunto. En la Tabla 5.2 se resumen tales resultados, mostrando una diferencia pequeña, pero no despreciable. Esto se puede interpretar como que el peor caso de ambas métricas por separado es notoriamente peor que el de la métrica conjunta, lo que podría indicar que los certificados de peor caso representados por los PEP provienen de la construcción de funciones estructuralmente distintas. Consecuentemente, una función construida para ambos fenómenos de optimización y estabilidad no puede alcanzar el valor de los peores casos separados, en términos de su desempeño en la métrica conjunta.

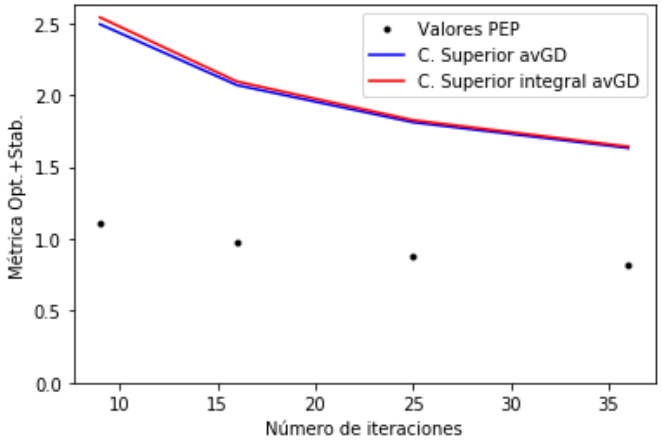

FIGURA 5.6. Comparación del peor caso computado de la métrica conjunta de riesgo empírico más estabilidad algorítmica con la cotas superiores teóricas después de $T = n^2$ iteraciones del método de subgradiente.

TABLA 5.2. Diferencia entre la suma de los problemas de optimización y estabilidad por separado, y el problema conjunto (Δ) para el problema de peor caso del método de subgradiente con distinto número de iteraciones $T = n^2$.

n	T	Δ
3	9	0,10860789752275424
4	16	0,08232297517611487
5	25	0,0648048062750256
6	36	0,05217494645806475

Un análisis de las variables duales en el óptimo podría ayudar a identificar la fuente del desajuste y desarrollar una cota más ajustada. Sin embargo, encontrar expresiones simbólicas para las variables duales requiere interpretación y validación del crecimiento de los datos con respecto a los distintos parámetros, lo que sumado al tamaño del problema semidefinido $\Theta(T^2)$ incurre en un crecimiento muy rápido de los tiempos de ejecución debido a la elección $T = n^2$.

5.2.3. Método de punto proximal

En lo que sigue, se presentan los resultados obtenidos de la implementación de las distintas métricas de rendimiento para el método de punto proximal para pérdidas Lipschitzianas (Supuesto 1a). El Algoritmo 5 entrega por respuesta el último iterado, por lo

que se considera la misma métrica conjunta para el último iterado, esta vez representada de manera semidefinida por b_{last} y la matriz C_{PPM}. Además, se acota la métrica de acuerdo al resultado de convergencia para el error de optimización del Lema 2.11 y del resultado de estabilidad del Corolario 3.1:

$$F_{\mathbf{S}}(x_{T+1}) - F_{\mathbf{S}}(x_{\mathbf{S}}^*) + M||x_{T+1} - y_{T+1}|| \leq \frac{R^2}{4\eta T} + \frac{2M^2\eta T}{n}.$$

De acuerdo con lo expuesto en la Sección 2.3 y el Capítulo 3, se fijan $T = n$ iteraciones y un *learning rate* de orden $\Theta\left(\frac{1}{\sqrt{n}}\right)$. En particular, se elige $\eta = \frac{1}{2\sqrt{2n}}$ que minimiza la cota superior antes mostrada.

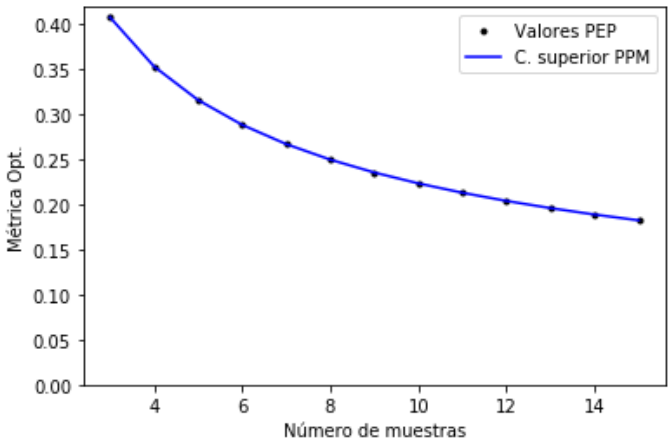

FIGURA 5.7. Comparación del peor caso computado del riesgo empírico con la cota superior teórica después de $T = n$ iteraciones del método de punto proximal.

Primero, se analiza el caso de sólo el error de optimización. La Figura 5.7 muestra un gráfico de los valores alcanzados en el peor caso, junto a la cota superior respectiva para este tipo de error. El PEP alcanza los valores de la cota superior proveniente del Lema 2.11, lo que indica que tal cota es ajustada, concordando con el resultado basado en PEP de Taylor et al.

Se estudia el problema dual (5.15), aplicando la metodología PEP para intentar extraer una demostración válida del caso solo optimización. Se nota que la clase de funciones M-Lipschitz es cerrada con respecto a la suma ponderada de funciones, resultando que este

```
mu 0: 1.2470141464479143e-08
mu 1: 6.146585112577228e-09
mu 2: 5.6877476395556445e-09
mu 3: 5.546030174942963e-09
mu 4: 2.2744169775807385e-08
mu 5: -1.123670426546038e-09
mu 6: -1.1211742750039691e-09
mu 7: -1.122663580113056e-09
mu 8: 1.4855994199491563e-09
mu 9: 4.57910398372921e-09
mu 10: 2.2847736068735464e-10
mu 11: 2.6578286883719484e-09
mu 12: 2.2744169769058233e-08
mu 13: 1.0491867713448307e-09
mu 14: -7.085093227499633e-10
mu 15: -1.5567902400901077e-09
mu 16: 7.416611088596891e-09
mu 17: 3.662378552493647e-08
mu 18: -1.2346493744883446e-09
mu 19: 5.203725821160701e-09
mu 20: -2.207289143008543e-09
mu 21: 0.0
tau: [0.40824816]
kappa: [-2.26263795e-09]
```

FIGURA 5.8. Recorte del *output* de la implementación dual del problema PEP de optimización previo a la introducción de restricciones de umbral, para $n = 3$.

problema sobre el riesgo empírico es equivalente al problema de minimizar mediante el PPM el peor caso de una única función en la clase generalizada.

Siguiendo la implementación del problema dual, se imponen restricciones a modo de simplificar el problema. La Figura 5.8 muestra parte del *output* de la implementación dual, explicitando los valores que toman las variables duales[1] en la respuesta del *solver*. En esta figura se aprecian variables duales nulas. También se nota que el *solver* entrega variables cercanas a cero que pueden descartarse tras ejecutar el problema nuevamente, agregando la restricción de umbral que anule tal variable. Este fenómeno de variables cercanas al cero se debe a la resolución de la implementación del PEP semidefinido con el método de punto interior.

Una vez se descartan las variables nulas, se puede analizar el crecimiento del resto de las variables duales con respecto a n, η y con respecto a los parámetros no explicitados del problema 5.1. Se infieren dependencias, que se prueban mediante ejecuciones con restricciones de umbral adicionales, hasta lograr una representación simbólica exacta de cada variable dual para el problema, en caso de ser posible. En este caso, se infieren los

[1]Se nota que la enumeración de las variables duales implementadas parte desde cero, a diferencia de I_M.

siguientes valores para las variables duales:

$$\begin{aligned}
\lambda_{i+1,i+2} &= \frac{i}{n(2n-i)} && (\forall i \in [n-1]) \\
\lambda_{n+2+i,n+3+i} &= \frac{i(n-1)}{n(2n-i)} && , (\forall i \in [n-1]) \\
\lambda_{4n+9,i+1} &= \frac{2}{(2n-i)(2n+1-i)} && , (\forall i \in [n]) \\
\lambda_{4n+10,n+2+i} &= \frac{2(n-1)}{(2n-i)(2n+1-i)} && , (\forall i \in [n]) \\
\lambda_{i,j} &= 0 && \text{e.o.c. } (i,j) \in \mathcal{I} \\
\mu_i &= 0 && (\forall i \in I_M) \\
\tau &= \frac{1}{4\eta n} \\
\kappa &= 0
\end{aligned} \tag{5.16}$$

Esto define un punto que debe probarse factible y que bajo tal garantía indica los ponderadores respectivos a las restricciones primales necesarios para recuperar una prueba del PEP. En particular, la Demostración 5.2.3 provee una prueba alternativa al Lema 2.11 para el caso especial $T = n$, que compete al caso planteado. La demostración de factibilidad dual, en cambio, se relega al Anexo B.3.

Lema 5.1. *Sea $f(\cdot, \xi) \in \mathcal{F}_M(\mathcal{X})$ para todo $\xi \in \mathcal{Z}$. Luego, el método de punto proximal (Algoritmo 5) con paso $\eta > 0$ garantiza*

$$F_S(x_{n+1}) - F_S(x_S^*) \leq \frac{||x_1 - x_S^*||^2}{4\eta n}.$$

DEMOSTRACIÓN. *Se nota que el conjunto de restricciones primales asociadas a cada variable corresponden a dos tipos de desigualdades. Estas son desigualdades de interpolación convexa de la forma*

$$v_j - v_i + \langle g_j, x_i - x_j \rangle \leq 0$$

y la condición de cercanía inicial al óptimo:

$$\langle x_1, x_1\rangle^2 - R^2 \leq 0.$$

Ponderando tales restricciones por sus variables duales respectivas se tiene la desigualdad:

$$\begin{aligned}
0 \geq & \sum_{i=1}^{n-1} \frac{i}{n(2n-i)} \left(v_{i+2} - v_{i+1} + \langle g_{i+2}, x_{i+1} - x_{i+2}\rangle\right) \\
& + \sum_{i=1}^{n-1} \frac{i(n-1)}{n(2n-i)} \left(v_{i+n+3} - v_{i+n+2} + \langle g_{i+n+3}, x_{i+n+2} - x_{i+n+3}\rangle\right) \\
& + \sum_{i=1}^{n} \frac{2}{(2n-i)(2n+1-i)} \left(v_{i+1} - v_{4n+9} + \langle g_{i+1}, x_{4n+9} - x_{i+1}\rangle\right) \\
& + \sum_{i=1}^{n} \frac{2(n-1)}{(2n-i)(2n+1-i)} \left(v_{i+n+2} - v_{4n+10} + \langle g_{i+n+2}, x_{4n+10} - x_{i+n+2}\rangle\right) \\
& + \frac{1}{4\eta n} \left(\langle x_1, x_1\rangle - R^2\right).
\end{aligned}$$

Además, tales sumas pueden reescribirse utilizando la desigualdad vectorial de (5.15) *y la propiedad de monotonicidad cíclica del gradiente:*

$$= \frac{1}{n}\left(v_{n+1} - v_{4n+9}\right) + \frac{n-1}{n}\left(v_{2n+2} - v_{4n+10}\right) + \frac{1}{4\eta n}\left(\langle x_1, x_1\rangle - R^2\right).$$

Finalmente, de la no-negatividad de la norma se obtiene:

$$\geq \frac{1}{n}\left(v_{n+1} - v_{4n+9}\right) + \frac{n-1}{n}\left(v_{2n+2} - v_{4n+10}\right) - \frac{R^2}{4\eta n}.$$

Se nota que v_{n+1}, v_{4n+9} *son los valores de las evaluaciones de la muestra* f*, mientras que* v_{2n+2}, v_{4n+10} *se interpretan como evaluaciones sobre la función de referencia* F*, que contiene el resto de las muestras asociadas al problema de minimización de riesgo empírico. Luego, por la interpolación convexa Lipschitziana de* (5.2)*, se tiene:*

$$F_{\mathbf{S}}(x_{T+1}) - F_{\mathbf{S}}(x^*) = (f+F)(x_{n+1}) - (f+F)(x^*) \leq \frac{R^2}{4\eta n}.$$

□

En segundo lugar, se analiza el problema de obtener el peor caso de estabilidad para el mismo método. La Figura 5.9 muestra que la distancia entre respuestas de muestras vecinas se ajusta a la cota superior derivada de la uniforme estabilidad de argumentos, alcanzando los valores de ésta con errores despreciables y que, al igual que el caso anterior, pueden atribuirse al *solver* utilizado en la resolución del problema semidefinido.

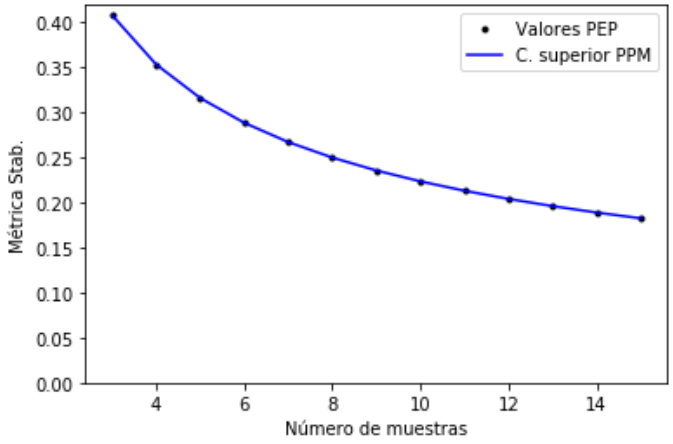

FIGURA 5.9. Comparación del peor caso computado de la uniforme estabilidad con la cota superior teórica después de $T = n$ iteraciones del método de punto proximal.

Se busca una demostración para este resultado a través de la metodología PEP. Nuevamente, se añaden restricciones de umbral al dual de la representación semidefinida del PEP *batch*, con el fin de encontrar una expresión simbólica para todas las variables duales. Se postulan los siguientes valores de variables duales, los que mantienen el valor óptimo

alcanzado por el problema computacional y se obtuvieron a contar de aplicar la metodología expuesta, agregando restricciones con postulados de expresiones simbólicas para cada variable tras analizar su crecimiento respecto a los parámetros del problema.

$$\begin{aligned}
\lambda_{i+n+2,i+3n+4} &= \frac{n-1}{2i} && (\forall i \in [n]) \\
\lambda_{i+3n+4,i+n+2} &= \frac{n-1}{2i} && (\forall i \in [n]) \\
\lambda_{i,j} &= 0 && \text{e.o.c. } (i,j) \in \mathcal{I} \\
\mu_i &= \frac{\eta}{2n} && (\forall i \in [2, n+1] \cup [2n+4, 3n+3]) \\
\mu_i &= 0 && \text{e.o.c. } i \in I_M \\
\tau &= 0 \\
\kappa &= M^2\eta
\end{aligned} \tag{5.17}$$

En base a esta elección de variables se plantea una conjetura proveniente de la aplicación de la metodología. Se provee una demostración parcial en el Anexo B.3, notando que esta idea se basa en la metodología de trabajo para los PEP. La Conjetura 5.1 concierne la factibilidad del punto encontrado, cuya prueba no se logró a cabalidad. Sin embargo, la evidencia computacional provista por la implementación del dual y una implementación complementaria para el cálculo de los valores propios de la matriz involucrada en la restricción semidefinida apuntan a que tal matriz es, en efecto, semidefinida positiva. A través de ambas implementaciones, se logró obtener expresiones para la mayoría de los valores propios una matriz que puede comprobarse su condición de ser semidefinida como una restricción equivalente a la desigualdad matricial original, dados los valores duales de (5.17).

Conjetura 5.1. *El punto dado por la asignación* (5.17) *es dual factible para el problema de peor caso de estabilidad luego de* $T = n$ *iteraciones del método de punto proximal.*

Suponiendo el cumplimiento de la conjetura anterior, una pregunta subsecuente es si será posible traducir este certificado dual en una demostración de cota superior para la

estabilidad. Se nota que las restricciones primales asociadas a cada variable dual no-nula son de alguna de las siguientes formas:

$$\begin{cases} v_j - v_i + \langle g_j, x_i - x_j \rangle \leq 0, \\ ||g_i||^2 - M^2 \leq 0, \\ ||y||^2 - 1 \leq 0. \end{cases}$$

Por lo tanto, la metodología PEP permite recuperar la siguiente combinación de restricciones:

$$\begin{aligned} 0 \geq & \sum_{i=1}^{n} \frac{n-1}{2i} \left(v_{i+3n+4} - v_{i+n+2} + \langle g_{i+3n+4}, x_{i+n+2} - x_{i+3n+4} \rangle \right) \\ & + \sum_{i=1}^{n} \frac{n-1}{2i} \left(v_{i+n+2} - v_{i+3n+4} + \langle g_{i+n+2}, x_{i+3n+4} - x_{i+n+2} \rangle \right) \\ & + \sum_{i=1}^{n} \frac{\eta}{2n} \left(||g_{i+1}||^2 - M^2 \right) + \sum_{i=1}^{n} \frac{\eta}{2n} \left(||g_{i+2n+3}||^2 - M^2 \right) + \eta M^2 \left(||y||^2 - 1 \right). \end{aligned}$$

Reordenando términos, se puede traducir en el contexto de SCO para modelos *batch* como la desigualdad

$$\begin{aligned} 2\eta M^2 \geq & - \sum_{i=1}^{n} \frac{n-1}{2i} \langle \nabla F(y_{i+1}) - F(x_{i+1}), y_{i+1} - x_{i+1} \rangle \\ & + \sum_{i=1}^{n} \frac{\eta}{2n} ||\nabla f(x_{i+1})||^2 + \sum_{i=1}^{n} \frac{\eta}{2n} ||\nabla f'(y_{i+1})||^2 + \eta M^2, \end{aligned}$$

de la que no se deduce fácilmente la demostración de estabilidad esperada. Pese a no tener una respuesta certera, los problemas fueron planteados para el caso $T = n$, que concierne a las garantías de exceso de riesgo planteadas en los Capítulos 2 y 3. El estudio de este PEP podría facilitarse estudiando el crecimiento de T y n independientemente. Preliminarmente, se encontró que para el caso $T \leq n$ el proceso de recuperar expresiones

simbólicas para las variables lleva a un candidato a certificado dual

$$\begin{aligned}
\lambda_{i+T+2,i+3T+4} &= \frac{n-1}{2i} && (\forall i \in [T]) \\
\lambda_{i+3T+4,i+T+2} &= \frac{n-1}{2i} && (\forall i \in [T]) \\
\lambda_{i,j} &= 0 && \text{e.o.c. } (i,j) \in \mathcal{I} \\
\mu_i &= \frac{\eta}{2n} && (\forall i \in [2, T+1] \cup [2T+4, 3T+3]) \\
\mu_i &= 0 && \text{e.o.c. } i \in I_M \\
\tau &= 0 \\
\kappa &= \frac{M^2 \eta T}{n},
\end{aligned} \tag{5.18}$$

cuya factibilidad debe ser probada de la misma manera que fue planteado para la Conjetura 5.1 en el caso $T = n$.

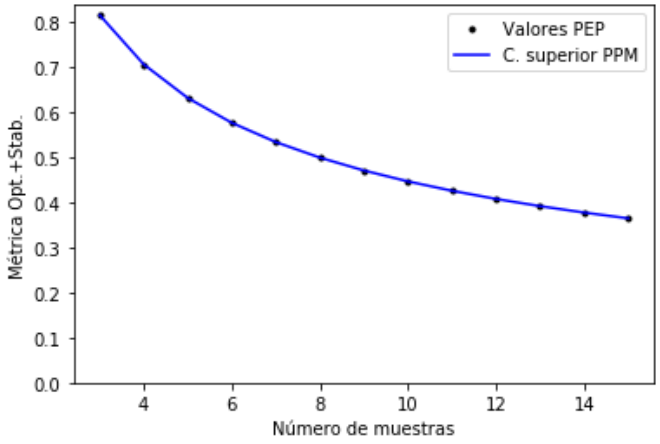

FIGURA 5.10. Comparación del peor caso computado de la métrica conjunta de riesgo empírico más estabilidad algorítmica con la cota superior teórica después de $T = n$ iteraciones del método de punto proximal.

Por último se estudia la implementación primal del problema conjunto de errores de estabilidad y optimización. Como se muestra en la Figura 5.10, los valores alcanzados por el problema conjunto se ajustan a la cota superior conjunta proveniente del Lema 2.11 y el Corolario 3.1, alcanzando una diferencia despreciable para la escala del problema y que

podría explicarse por la tolerancia del método de descenso utilizados sobre el problema semidefinido.

TABLA 5.3. Diferencia entre la suma de los problemas de optimización y estabilidad por separado, y el problema conjunto (Δ) para el problema de peor caso del método de gradiente con distintos tamaños de muestra.

n	Δ
3	$1{,}315 \times 10^{-7}$
4	$7{,}928 \times 10^{-8}$
5	$2{,}011 \times 10^{-7}$
6	$3{,}572 \times 10^{-8}$
7	$2{,}331 \times 10^{-7}$
8	$2{,}486 \times 10^{-7}$
9	$9{,}887 \times 10^{-8}$
10	$3{,}739 \times 10^{-7}$
11	$4{,}136 \times 10^{-7}$
12	$3{,}015 \times 10^{-7}$
13	$3{,}452 \times 10^{-7}$
14	$5{,}974 \times 10^{-8}$
15	$5{,}542 \times 10^{-7}$

Además, los valores expuestos en la Tabla 5.3 indican que la implementación alcanza una diferencia despreciable entre el problema conjunto y los problemas separados, concluyéndose que para este caso el acercamiento a ambos problemas por separado no constituye un deterioro notorio al valor del peor caso conjunto.

6. UN MODELO PEP PARA ALGORITMOS INCREMENTALES

En este capítulo se presenta una segunda formulación semidefinida del problema de estimación de rendimiento que modela los errores de optimización y estabilidad para el problema (1.3). Esta formulación permite la representación matricial del PEP de ambos tipos de error para una minimización hecha mediante actualizaciones de un método incremental con permutación fija, como los expuestos en las Secciones 2.2 y 2.4.

El problema finito dimensional (6.1) busca el peor caso, en términos de una métrica de rendimiento que engloba los fenómenos de optimización y estabilidad, de dos trayectorias definidas por muestras vecinas $\mathbf{S} \simeq \mathbf{S}'$. Esta vez, la utilización de los datos cambia en cada actualización, eligiendo un único dato secuencialmente. El uso de permutaciones implica el reúso de datos una cantidad K exacta de veces durante toda la minimización. Consecuentemente, el problema debe modelar la elección de cada función por separado, necesitándose un total de $n+1$ funciones interpoladas. Sin pérdida de generalidad, se considera nuevamente a $\mathbf{S}$ como aquella muestra sobre la cual se minimiza el riesgo empírico definido por pérdidas $f_1, ..., f_n$, y $\mathbf{S}'$ como su perturbación marginal, denotando la pérdida f_1 perturbada por f'. Se nota que salvo el par correspondiente a la perturbación, todas estas pérdidas son intercambiables entre sí, cobrando importancia solamente cuándo se actualizan las trayectorias utilizando las pérdidas que difieren.

$$
\begin{aligned}
w = \sup_{\{\mathcal{O}_{f_i}\}_{i\in[n]}, \mathcal{O}_{f'}, \{x_i\}_{i\in I}, \{y_i\}_{i\in I}} \quad & \mathcal{P}(\{\mathcal{O}_{f_i}\}_{i\in[n]}, \mathcal{O}_{f'}, \{x_i\}_{i\in I}, \{y_i\}_{i\in I}) \\
\text{s.a} \quad & f_1, ..., f_n, f' \in \mathcal{F}, \\
& x_* \text{ óptimo de } \frac{1}{n}\sum_{i\in[n]} f_i, \\
& x_{i+1}, y_{i+1} \text{ generados por } \mathcal{M} \text{ desde } x_1 \quad (\forall i \in [T]) \\
& \mathcal{M} \text{ un método incremental con acceso a } \mathcal{O}_{(\cdot)} \\
& x_1 \text{ satisface la condición de inicialización } \mathcal{C}.
\end{aligned}
\tag{6.1}
$$

Se quiere reescribir (6.1) mediante un problema semidefinido, para lo que debe fijarse el orden de aparición de f_1 en las permutaciones realizadas por un método incremental. De lo contrario, el modelo debe incluir el máximo entre n problemas, cada uno referente a las posibles posiciones de aparición de f_1 en el orden de utilización de los datos de $\mathbf{S}$, para cada pasada realizada sobre los datos.

Considérese la clase de funciones Lipschitz $\mathcal{F}_M(\mathbb{R}^d)$ que se utilizará durante este capítulo, notando la relación ya expuesta entre las nociones de estabilidad algorítmica y estabilidad algorítmica de argumentos. Sea una pérdida $g \in \mathcal{F}_M(\mathbb{R}^d)$ que alcanza una diferencia en norma

$$||\mathcal{A}(\mathbf{S}) - \mathcal{A}(\mathbf{S}')||$$

después de la ejecución de un algoritmo $\mathcal{A}$ de una pasada con permutación, para trayectorias vecinas definidas por las muestras vecinas $\mathbf{S}, \mathbf{S}'$ que difieren en la i-ésima actualización. Mientras $\mathcal{A}$ utilice una respuesta en términos de la ronda completa, es posible determinar muestras alternativas que entregan la misma respuesta modificando solamente el punto inicial del algoritmo. Esto es, existe un algoritmo $\mathcal{A}'$ que inicia desde x_i y muestras $\mathbf{S}_i = (\boldsymbol{\xi}_i, ..., \boldsymbol{\xi}_n, 0, ..., 0)$ y $\mathbf{S'}_i = (\boldsymbol{\xi}'_i, ..., \boldsymbol{\xi}_n, 0, ..., 0)$ tales que

$$||\mathcal{A}(\mathbf{S}) - \mathcal{A}(\mathbf{S}')|| = ||\mathcal{A}'(\mathbf{S}_i) - \mathcal{A}'(\mathbf{S}'_i)||,$$

donde los algoritmos utilizan los datos de las tuplas en orden de izquierda a derecha, sin pérdida de generalidad. Sin embargo, la extensión a múltiples pasadas, incluso con una única permutación, no es directa por dos razones. Primero, un algoritmo que responda el último iterado no necesariamente puede extenderse por cero en las últimas $i-1$ iteraciones sin modificar las diferencias entre trayectorias. Segundo, para una respuesta más general, manteniendo el requisito de utilizar solo información de las rondas completas, no puede establecerse una relación clara entre las normas de los iterados que finalizan cada ronda.

Añadir algún supuesto extra, como la suavidad de las pérdidas, permite utilizar características como la no-expansividad de actualizaciones de gradiente o de pasos proximales para garantizar que trasladar la perturbación hacia el inicio de cada ronda no empeora

la diferencia entre normas. Esto sugiere que en el caso no-suave podría encontrarse una diferencia en norma más grande cuando se traslada la perturbación al inicio de la ronda, mientras no se pueda asegurar la no-expansividad de las actualizaciones. Basado en tales ideas, se plantea la siguiente conjetura como una base para escribir el PEP a través de un único problema semidefinido.

Conjetura 6.1. *El problema de peor caso de la estabilidad se alcanza cuando la perturbación a las muestras ocurre en la primera actualización.*

Además, en este capítulo se utilizan las versiones de permutación fija de los algoritmos incrementales expuestos, manteniendo el orden en que se utiliza cada dato de las muestras. En vista de los resultados introducidos en el marco teórico, se nota que este supuesto aún lleva a algoritmos que hacen converger a cero la estabilidad y el riesgo empírico, cuando $n \to \infty$.

Suponiendo el cumplimiento de la conjetura anterior, se tiene la equivalencia entre el modelo finito ya expuesto y el problema de (6.2).

$$
\begin{aligned}
w = \sup_{\{\mathcal{O}_{f_i}\}_{i\in[n]},\mathcal{O}_{f'},\{x_i\}_{i\in I},\{y_i\}_{i\in I}} \quad & \mathcal{P}(\{\mathcal{O}_{f_i}\}_{i\in[n]}, \mathcal{O}_{f'}, \{x_i\}_{i\in I}, \{y_i\}_{i\in I}) \\
\text{s.a} \quad & f_1, ..., f_n, f' \in \mathcal{F} \\
& x_* \text{ óptimo de } \frac{1}{n}\sum_{i\in[n]} f_i \\
& x_{i+1}, y_{i+1} \text{ generados por } \mathcal{M} \text{ desde } x_1 \quad (\forall i \in [T]) \\
& f_1 \text{ es la primera muestra utilizada en cada ronda} \\
& f' \text{ es la perturbación de } f_1 \\
& \mathcal{M} \text{ un método incremental con acceso a } \mathcal{O}_{(\cdot)} \\
& x_1 \text{ satisface la condición de inicialización } \mathcal{C}.
\end{aligned}
\tag{6.2}
$$

Sin embargo, por motivos de costo computacional se considera la simplificación del modelo 6.2 al caso irrestricto, de la misma manera que para el PEP para métodos *batch*. Se nota que la extensión al caso proyectado sigue la misma idea del modelo *batch*, añadiendo una función indicatriz extra que debe ser interpolada y restricciones adicionales al problema, que acotan tanto los iterados como los valores de la función que toman. En cuanto a garantías teóricas, las cotas de estabilidad para métodos incrementales introducidas en el Capítulo 2 no poseen una dependencia del radio inicial R y pueden acotarse trivialmente por el radio r que encierra a las trayectorias (del Supuesto 2) en aquellos casos donde la elección de η no asegura una convergencia a cero de tal tipo de error. Sin embargo, esta nueva cota no influye si se puede asegurar dicha convergencia, concluyéndose que la Conjetura 6.1 es razonable en el caso proyectado inclusive.

6.1. Una formulación semidefinida para modelos incrementales

Tras formular el problema finito, se plantea un modelo semidefinido equivalente al último PEP a través de una noción de Gram-representabilidad lineal específica para métodos incrementales. En pos de tal objetivo, se presenta para cada método incremental un resultado que argumenta la equivalencia cuando los elementos que definen un este nuevo PEP son representables en una forma semidefinida, basándose en la información utilizada por la actualización de los métodos incrementales.

Además, debido a que la representación del riesgo empírico depende de n pérdidas, debe incluirse la representabilidad de la condición de optimalidad que es necesaria para la representación del iterado óptimo. En este caso, a diferencia del Capítulo 5, la dependencia de múltiples funciones dificulta la inclusión de esta condición en la estructura del problema. Se nota que para lograr esto se deben interpolar los valores y subgradientes, donde esta vez se plantea una elección de puntos, valores y gradientes de interpolación que dependerá de cada método. Es decir, ya que los métodos utilizan un conjunto distinto de subgradientes en la actualización del iterado, en este modelo se elegirá la correspondencia entre cada punto y las valuaciones de su valor y subgradientes necesarios de acuerdo a cada método, permitiendo una reducción en las dimensiones de la matriz de Gram. Se

requiere fijar las columnas

$$P = [g_1 \dots g_{Kn}\, g'_1 \dots g'_{Kn}\, \bar{g}_1 \dots \bar{g}_n\, \bar{g}'_1 \dots \bar{g}'_n\, g^*_1 \dots g^*_n\, x_1\, y] \tag{6.3}$$

y el vector de valores

$$v = \left[f_1 \dots f_{Kn}\, f'_1 \dots f'_{Kn}\, \bar{f}_1 \dots \bar{f}_n\, \bar{f}'_1 \dots \bar{f}'_n\, f^*_1 \dots f^*_n\right]^\top, \tag{6.4}$$

junto a una elección pertinente de puntos. En particular, para ambos métodos se eligen en las primeras Kn columnas los puntos que corresponden a los subgradientes empleados en las Kn iteraciones del método.

Para el método SGD incremental, esto implica una elección a contar de x_1, dado que la primera actualización utiliza un subgradiente del mismo iterado donde inicia. Extrapolando, para todo $i \in [n]$ y $k \in [K]$ se formaliza la interpolación:

$$\begin{cases} f_{(k-1)n+i} = f_i(x_{(k-1)n+i}), \\ g_{(k-1)n+i} = \nabla f_i(x_{(k-1)n+i}), \\ f'_{(k-1)n+i} = f_i(y_{(k-1)n+i}), \\ g'_{(k-1)n+i} = \nabla f_i(y_{(k-1)n+i}). \end{cases} \tag{6.5}$$

Por otra parte, el método *IncrementalProx* utiliza un paso proximal, lo que resulta en una elección implícita del subgradiente involucrado en la actualización como uno perteneciente al punto resultante del paso. Luego, la interpolación debe hacerse a contar de x_2 o y_2, dependiendo de la trayectoria tomada. Formalizando lo anterior, para todo $i \in [n]$ y $k \in [K]$:

$$\begin{cases} f_{(k-1)n+i} = f_i(x_{(k-1)n+i+1}), \\ g_{(k-1)n+i} = \nabla f_i(x_{(k-1)n+i+1}), \\ f'_{(k-1)n+i} = f_i(y_{(k-1)n+i+1}), \\ g'_{(k-1)n+i} = \nabla f_i(y_{(k-1)n+i+1}). \end{cases} \tag{6.6}$$

Luego, ambas actualizaciones irrestrictas (2.10, 2.13) pueden reescribirse de la forma

$$x_{t+1} = x_t - \eta g_t \quad (\forall t \in [Kn]), \tag{6.7}$$

siempre que se tenga la representación semidefinida de iterados $Ph_{(\cdot)}$ adecuada para el método elegido. Esto descarta la elección de subgradiente de la pérdida asociada en el punto que no necesita evaluarse, para cada actualización. Por ejemplo, para la regla de actualización de gradiente estocástico (2.10) no se necesita incorporar la elección de $\nabla f_t(x_{t+1})$, donde f_t es la pérdida elegida en la actualización t-ésima. Además, se considera para ambos casos la interpolación de la información común a ambos métodos:

$$\begin{cases} \bar{f}_i = f_i(\bar{x}_K) & (i \in [n]), \\ \bar{g}_i = \nabla f_i(\bar{x}_K) & (i \in [n]), \\ \bar{f}'_i = f_i(\bar{y}_K) & (i \in [n]), \\ \bar{g}'_i = \nabla f_i(\bar{y}_K) & (i \in [n]), \\ \bar{f}^*_i = f_i(x^*_{\mathbf{S}}), & \\ \bar{g}^*_i = \nabla f_i(x^*_{\mathbf{S}}). & \end{cases} \tag{6.8}$$

Luego, se debe definir una noción de Gram-representabilidad para llevar el problema finito (6.2) a la forma matricial (6.9), considerándose que se quiere representar el vector de valores v de (6.4) y la elección de columnas de (6.3) a través de la matriz de Gram $X = P^\top P$ y cumpliendo la interpolación adecuada a cada método. Tal propiedad se expresa como sigue, mediante las Definiciones 6.1-6.5 y se le llama *incremental Stability-and-Optimization* Gram-representabilidad lineal (abr. iSOLG-representable).

$$\begin{aligned} \sup_{v \in \mathbb{R}^{(2K+3)n}, X \in \mathbb{S}^{(2K+3)n+2}} \quad & b^\top v + \mathrm{Tr}(CX) \\ \text{s.a} \quad & a_j^\top v + \mathrm{Tr}(M_j X) + c_j \leq 0, \quad (j \in J) \end{aligned} \tag{6.9}$$

Definición 6.1. *Un método de primer orden es iSOLG-representable si y solo si el cómputo de sus iterados, definidos (posiblemente de manera implícita) por* (2.19)*, puede*

ser expresada utilizando una cantidad finita de restricciones lineales con dependencias solo de $X \in \mathbb{S}^{(2K+3)n+2}$ *y* $v \in \mathbb{R}^{(2K+3)n}$.

La Definición 6.1 alude al cómputo de los iterados de los métodos incrementales. Se muestra que tanto SGD incremental como *IncrementalProx* cumplen esta propiedad de representabilidad, basándose en (6.7).

Afirmación 6.1. *El método de gradiente estocástico incremental con permutación fija (Algoritmo 4) es iSOLG-representable.*

DEMOSTRACIÓN. *Utilizando el hecho que ambos métodos en su forma irrestricta pueden representarse por* (6.7)*, resta derivar vectores* $h_{(\cdot)}$ *y* $u_{(\cdot)}$ *de manera de tener la correspondencia entre los iterados utilizados y los gradientes y valores asociados a las entradas del vector* v *y a las columnas de* P*. Consecuentemente, se fijan* g_1, g_1' *de manera que interpolan a subgradientes en el iterado inicial en base a* (6.5) *y los valores de las últimas columnas de acuerdo a* (6.8)*, buscándose* $h_{(\cdot)}$ *de manera que, para* $x_1 = y_1$*:*

$$Ph_i = \begin{cases} x_i & (i \in [Kn]), \\ y_{i-Kn} & (i \in [Kn+1, 2Kn]), \\ \bar{x} & (i \in [2Kn+1, (2K+1)n]), \\ \bar{y} & (i \in [(2K+1)n+1, (2K+2)n]), \\ 0 & (i \in [(2K+2)n+1, (2K+3)n]). \end{cases}$$

Luego, utilizando (6.7) *se definen* h_i *de la forma*

$$
\begin{aligned}
h_i &= e_{(2K+3)n+1} - \eta \sum_{j=1}^{i-1} e_j && (i \in [Kn]) \\
h_{i+Kn} &= e_{(2K+3)n+1} - \eta \sum_{j=1}^{i-1} e_{j+Kn} && (i \in [Kn]) \\
h_{i+2Kn} &= e_{(2K+3)n+1} - \eta \sum_{k=1}^{K} \sum_{j=1}^{n} \frac{K+1-k}{K} e_{(k-1)n+j} && (i \in [n]) \\
h_{i+(2K+1)n} &= e_{(2K+3)n+1} - \eta \sum_{k=1}^{K} \sum_{j=1}^{n} \frac{K+1-k}{K} e_{(K+k-1)n+j} && (i \in [n]) \\
h_{i+(2K+2)n} &= 0 && (i \in [n]),
\end{aligned}
$$

donde sin pérdida de generalidad se asume que $x_* = 0$*. Los vectores* $u_{(.)}$ *se determinan de la misma manera, tal que calce el iterado con cada gradiente* Pu_i*, lo que lleva a la elección* $u_i = e_i$ *para todo* $i \in [(2K+3)n]$*. Se concluye que el método de gradiente estocástico incremental es iSOLG-representable, notando el hecho que Tr*$(ab^\top X) = (Pa)^\top (Pb)$ *permite la representación de las elecciones de* $u_{(.)}$ *y* $h_{(.)}$ *al interior de productos de traza.* □

Afirmación 6.2. *El método IncrementalProx con permutación fija (Algoritmo 7) es iSOLG-representable.*

DEMOSTRACIÓN. *Siguiendo la prueba de la afirmación anterior, se deduce de* (6.7) *que es necesario fijar* g_1, g_1' *como los subgradientes* $\nabla f_1(x_2)$ *y* $\nabla f'(y_2)$ *buscados por la primera actualización realizada por el método IncrementalProx. Asimismo, las primeras* $2Kn$ *columnas se fijan de manera de representar el iterado calculado implícitamente por la actualización, de acuerdo a la interpolación de* (6.6) *y el resto, utilizando* (6.8)*.*

Entonces, se buscan valores $h_{(\cdot)}$ tales que

$$Ph_i = \begin{cases} x_{i+1} & (i \in [Kn]), \\ y_{i+1-Kn} & (i \in [Kn+1, 2Kn]), \\ \bar{x} & (i \in [2Kn+1, (2K+1)n]), \\ \bar{y} & (i \in [(2K+1)n+1, (2K+2)n]), \\ 0 & (i \in [(2K+2)n+1, (2K+3)n]). \end{cases}$$

Luego, se definen h_i de la forma

$$\begin{aligned} h_i &= e_{(2K+3)n+1} - \eta \sum_{j=1}^{i-1} e_{j+1} && (i \in [Kn]) \\ h_{i+Kn} &= e_{(2K+3)n+1} - \eta \sum_{j=1}^{i-1} e_{j+Kn+1} && (i \in [Kn]) \\ h_{i+2Kn} &= e_{(2K+3)n+1} - \eta \sum_{k=1}^{K} \sum_{j=1}^{n} \frac{K+1-k}{K} e_{(k-1)n+j+1} && (i \in [n]) \\ h_{i+(2K+1)n} &= e_{(2K+3)n+1} - \eta \sum_{k=1}^{K} \sum_{j=1}^{n} \frac{K+1-k}{K} e_{(K+k-1)n+j+1} && (i \in [n]) \\ h_{i+(2K+2)n} &= 0 && (i \in [n]), \end{aligned}$$

donde sin pérdida de generalidad se asume que $x_* = 0$. Los vectores $u_{(\cdot)}$ se determinan de la misma manera que el método anterior, tal que correspondan el iterado con cada gradiente Pu_i, lo que lleva a la elección $u_i = e_i$ para todo $i \in [(2K+3)n]$. Se concluye que el método IncrementalProx también es iSOLG-representable. □

Luego, se define la representabilidad para una clase de funciones, mostrando que para ambos métodos se puede encontrar una representación lineal de las condiciones de interpolación asociada a $\mathcal{F}_M$ en términos de las nuevas variables, utilizando el producto de traza para la representación de iterados y subgradientes. Ambos casos se resumen en la

Afirmación 6.3, demostrándolo para el caso de la subclase $\mathcal{F}_{M,L}$, que se puede extender fácilmente a $\mathcal{F}_M$.

Definición 6.2. *Una clase de funciones es iSOLG-representable si y solo si sus condiciones de interpolación pueden ser reformuladas utilizando una cantidad finita de restricciones lineales dependiendo solo de $X \in \mathbb{S}^{(2K+3)n+2}$ y $v \in \mathbb{R}^{(2K+3)n}$.*

Afirmación 6.3. *Sea un método $\mathcal{M}$ iSOLG-representable a través de los iterados Ph_i y gradientes Pu_i para todo $i \in [(2K+3)n]$. Entonces, las clases $\mathcal{F}_M$ y $\mathcal{F}_{M,L}$ son iSOLG-representables.*

DEMOSTRACIÓN. *Se demuestra que las restricciones de interpolación de $\{f_i\}_{i\in[n]}$ pertenecientes a la clase $\mathcal{F}_{M,L}(\mathbb{R}^d)$ son representables de forma lineal en $X \in \mathbb{S}^{(2K+3)n+2}$ y $v \in \mathbb{R}^{(2K+3)n}$ para una elección arbitraria de d. Primero, se introducen los conjuntos de índices asociados a la interpolación de las $n+1$ pérdidas. El conjunto de índices relativos al primer dato está dado por*

$$I_{f_1} = \{(k-1)n+1 : k \in [K]\} \cup \{2Kn+1, (2K+2)n+1\}.$$

Además, se definen los índices correspondientes a la pérdida perturbada por

$$I_{f'} = \{(K+k-1)n+1 : k \in [K]\} \cup \{(2K+1)n+1\}.$$

Luego, se tiene que los indices relacionados con las pérdidas f_i, para $i > 1$, están relacionados por congruencia en módulo n:

$$I_{f_i} = \{(k-1)n+i : k \in [2K+3]\}.$$

Consecuentemente, se puede definir el conjunto de pares de interpolación convexa por

$$\mathcal{I} = \bigcup_{i\in[n]} [I_{f_i} \times I_{f_i}] \cup I_{f'} \times I_{f'}$$

y el conjunto de restricciones Lipschitz por $I_M = [(2K+3)n]$. Entonces, por las representaciones de iterados Ph_i y subgradientes Pu_i de las Afirmaciones 6.1 y 6.2, se tiene el

siguiente conjunto de restricciones de interpolación en función de la matriz de columnas P:

$$\begin{cases} v_i - v_j - \langle Pu_j, Ph_i - Ph_j \rangle \geq \frac{1}{2L}||Pu_i - Pu_j||^2 & (\forall (i,j) \in \mathcal{I}) \\ ||Pu_i|| \leq M & (\forall i \in I_M). \end{cases}$$

Luego, por la definición de X como una matriz de Gram, se puede representar la primera desigualdad por un producto de traza. Además, la segunda desigualdad puede reescribirse de manera equivalente por una restricción para el cuadrado de la norma, la que a la vez puede ser reescrita mediante un producto de traza. Luego, definiendo las matrices

$$\begin{aligned} A_{ij} &= \frac{1}{2}u_j(h_i - h_j)^\top + \frac{1}{2}(h_i - h_j)u_j^\top + \frac{1}{2L}(u_i - u_j)(u_i - u_j)^\top \\ A_{M_i} &= u_i u_i^\top \end{aligned}$$

se tiene la equivalencia con el conjunto de restriccciones lineales en X y v

$$\begin{cases} v_j - v_i + \mathbf{Tr}\left(A_{ij}X\right) \leq 0 & (\forall (i,j) \in \mathcal{I}) \\ \mathbf{Tr}(A_{M_i}X) - M^2 \leq 0 & (\forall i \in I_M), \end{cases}$$

terminando la demostración para funciones suaves, por la arbitrariedad de la elección de d. Para el caso no-suave, se considera la condición de interpolación convexa con el caso límite $L = +\infty$ para todos los pares. Esto presenta una simplificación de la matriz A_{ij} anterior, donde el término de normas de gradiente se anula. Por lo tanto, basta tomar la misma representación lineal, redefiniendo las matrices

$$A_{ij} = \frac{1}{2}u_j(h_i - h_j)^\top + \frac{1}{2}(h_i - h_j)u_j^\top.$$

□

Luego, la pertenencia a la subclase de pérdidas Lipschitz continuas puede representarse mediante un conjunto de restricciones lineales en X y v cuando se utiliza alguno de los métodos incrementales ya presentados. A continuación, se expresa la definición de

iSOLG-representabilidad para una métrica de rendimiento, considerando el *gap* de optimalidad y la diferencia entre respuestas de trayectorias generadas por muestras vecinas como medidas del error de optimización y estabilidad, respectivamente. Se nota que ambos algoritmos entregan por respuesta el promedio de los iterados finales de las rondas $\bar{x}_K$ e $\bar{y}_K$, cantidades en base a las que se definen ambas métricas.

Definición 6.3. *Una métrica de rendimiento es iSOLG-representable si y solo si puede ser expresada como una función lineal de* $X \in \mathbb{S}^{(2K+3)n+2}$ *y* $v \in \mathbb{R}^{(2K+3)n}$.

Afirmación 6.4. *El gap de optimalidad para el iterado promedio de las rondas* $F_S(\bar{x}_K) - F_S(x_*)$ *es iSOLG-representable.*

Esta métrica solamente depende de valores de las pérdidas y por lo tanto puede representarse linealmente mediante un producto entre las variables v y un vector b. Se nota que tal vector debe combinar linealmente las entradas de v a modo de obtener la siguiente suma de valores ponderados de las pérdidas,

$$\frac{1}{n}\sum_{i\in[n]}[f(\bar{x},\boldsymbol{\xi}_i) - f(x_*,\boldsymbol{\xi}_i)].$$

Entonces, se considera el vector $(2K+3)n$ dimensional

$$b = \frac{1}{n}\sum_{i\in[n]}\left[e_{2Kn+i} - e_{(2K+2)n+i}\right]. \tag{6.10}$$

Por otra parte, la métrica de estabilidad se representa mediante un producto de traza, como se muestra en la Afirmación 6.5.

Afirmación 6.5. *La distancia entre últimos iterados* $M||\bar{x}_K - \bar{y}_K||$ *generada por una respuesta del Algoritmo 4 o del Algoritmo 7 es una métrica de rendimiento iSOLG-representable.*

DEMOSTRACIÓN. *Se utiliza la variable auxiliar* y *para representar la norma como un supremo de producto interno. Se nota que es posible reescribir la norma como sigue*

para ambos algoritmos, en función de $h_{(\cdot)}$ y $u_{(\cdot)}$, utilizando el supuesto $||y|| \leq 1$:

$$\sup_{\mathcal{O}_{f_i},\mathcal{O}_{f'},\{x_i\}_{i\in I},\{y_i\}_{i\in I}} ||\bar{x}_K - \bar{y}_K|| = \sup_{\mathcal{O}_{f_i},\mathcal{O}_{f'},\{x_i\}_{i\in I},\{y_i\}_{i\in I}} \sup_{||y||\leq 1} \langle \bar{x}_K - \bar{y}_K, y\rangle.$$

Luego, se puede descomponer el promedio de iterados como

$$\langle \bar{x}_K - \bar{y}_K, y\rangle = \eta \left\langle \sum_{k=1}^{K}\sum_{i=1}^{n} \frac{K+1-k}{K}[g_{(K+k-1)n+i} - g_{(k-1)n+i}], y \right\rangle.$$

Debido a ser una suma ponderada de productos internos entre columnas de P, se puede expresar como un producto de traza entre X y una matriz de ponderadores C específica para la forma de actualización (6.7) *y la respuesta $\bar{x}_K$.*

$$= \textit{Tr}(CX).$$

Explícitamente, $C = (C_{ij})_{i,j\in[(2K+3)n+2]}$ toma valores

$$C_{ij} = \begin{cases} -\frac{M\eta(K+1-\lceil \frac{j}{n}\rceil)}{2K} & (i = (2K+3)n+2, j \in [Kn]) \\ -\frac{M\eta(K+1-\lceil \frac{i}{n}\rceil)}{2K} & (j = (2K+3)n+2, i \in [Kn]) \\ \frac{M\eta(K+1-\lceil \frac{j-K}{n}\rceil)}{2K} & (i = (2K+3)n+2, j \in [Kn+1, 2Kn]) \\ \frac{M\eta(K+1-\lceil \frac{i-K}{n}\rceil)}{2K} & (j = (2K+3)n+2, i \in [Kn+1, 2Kn]) \\ 0 & e.o.c. \end{cases} \tag{6.11}$$

□

Entonces, de las Afirmaciones 6.4 y 6.5 se comprueba la representabilidad de ambas métricas y subsecuentemente, la representabilidad de la suma de ambas por la expresión

$$b^\top v + \text{Tr}(CX).$$

Resta introducir una noción de iSOLG-representabilidad para dos tipos de restricciones del problema (6.2). Primero, se introduce esta idea para las condiciones de inicialización.

Se consideran tanto la condición inicial, como la restricción que permite representar la norma como un producto interno mediante el uso de la variable auxiliar y, garantizando que ambas pueden reescribirse como restricciones lineales de las variables semidefinidas en la Afirmación 6.6.

Definición 6.4. *Un conjunto de condiciones de inicialización es iSOLG-representable si y solo si puede ser reformulado utilizando una cantidad finita de restricciones lineales dependiendo solo de $X \in \mathbb{S}^{(2K+3)n+2}$ y $v \in \mathbb{R}^{(2K+3)n}$.*

Afirmación 6.6. *El conjunto de condiciones iniciales*

$$\begin{cases} ||y|| & \leq 1 \\ ||x_1 - x_*|| & \leq R \end{cases}$$

es iSOLG-representable.

DEMOSTRACIÓN. *Por la elección de columnas de P, se tiene $x_1 = Pe_{(2K+3)n+1}$ y $y = Pe_{(2K+3)n+2}$. Dado que ambas restricciones pueden ser reescritas como una raiz cuadrada de un producto interno, se hace la representación semidefinida en base a los cuadrados de las restricciones, utilizando productos de traza y notando que $x_* = 0$. Para las matrices $A_R = e_{(2K+3)n+1}e_{(2K+3)n+1}^\top$ y $A_y = e_{(2K+3)n+2}e_{(2K+3)n+2}^\top$, se tiene:*

$$\begin{cases} \mathit{Tr}(A_y X) - 1 & \leq 0 \\ \mathit{Tr}(A_R X) - R^2 & \leq 0. \end{cases}$$

□

Por último, se considera la condición de optimalidad, necesaria debido a la utilización del óptimo x_S^*.

Definición 6.5. *Una condición de optimalidad es iSOLG-representable si y solo si puede ser reformulada utilizando una cantidad finita de restricciones lineales dependiendo solo de $X \in \mathbb{S}^{(2K+3)n+2}$ y $v \in \mathbb{R}^{(2K+3)n}$.*

Afirmación 6.7. *La condición inicial $\nabla F_S(x_S^*) = 0$ es iSOLG-representable.*

DEMOSTRACIÓN. *Esta condición se cumple si y solo si*

$$\langle \nabla F_S(x_S^*), \nabla F_S(x_S^*) \rangle = 0.$$

Por la interpolación (6.8)*, se tiene la equivalencia entre elecciones de subgradientes de las pérdidas en el óptimo y los vectores g_i^*, para $i \in [n]$. Luego, se puede reescribir el producto como*

$$\langle \nabla F_S(x_S^*), \nabla F_S(x_S^*) \rangle = \left\langle \sum_{i=1}^n g_i^*, \sum_{i=1}^n g_i^* \right\rangle = \mathit{Tr}\left(A_* X\right),$$

para la matriz

$$A_* = \left(\sum_{i \in [n]} e_{(2K+2)n+i} \right) \left(\sum_{i \in [n]} e_{(2K+2)n+i} \right)^\top.$$

□

Reuniendo todas las definiciones anteriores, se reescriben todas las componentes del problema (6.2), llevándolas a la forma definida (6.12). Esto se resume en la Proposición 6.1, que garantiza que tal problema permite representar el peor caso del problema convexo subyacente por una forma semidefinida equivalente, para dimensión suficientemente alta.

Proposición 6.1. *Sea un método de primer orden $\mathcal{M}$, una clase de funciones $\mathcal{F}$, una métrica de rendimiento $\mathcal{P}$ y un conjunto de condiciones de inicialización $\mathcal{C}$ y una condición de optimalidad $\mathcal{J}$ iSOLG-representables. Luego, el problema* (6.2) *de representar de manera exacta el peor caso de $\mathcal{P}$ alcanzado por ejecuciones de $\mathcal{M}$ sobre una pérdida perteneciente a la clase $\mathcal{F}$, con condiciones de inicialización $\mathcal{C}$ puede ser reformulado como un problema semidefinido de la forma* (6.12) *para dimensiones $d \geq (2K+3)n + 2$.*

DEMOSTRACIÓN. *La iSOLG-representabilidad de las partes del problema está dada por las Definiciones 6.1-6.5 y por definición pueden ser reformuladas como expresiones lineales de X y v, únicas variables del problema semidefinido* (6.9)*.*

Para mostrar que la equivalencia se cumple solo para dimensiones altas, se nota que la matriz de Gram X admite una descomposición que utiliza la matriz P de dimensiones $d \times (2K+3)n+2$. Sea una dimensión $d < (2K+3)n+2$, una elección arbitraria sobre el cono semidefinido de tamaño $(2K+3)n+2$ puede incurrir en respuestas factibles de X cuya descomposición $P^\top P$ contenga una matriz P con rango mayor a d, lo que se contrapone la elección inicial de una matriz P de dimensiones $d \times (2K+3)n+2$. Luego, se debe limitar el problema a una elección en dimensión alta, restringiendo la clase $\mathcal{F}$ a una dimensión $d \geq (2K+3)n+2$.

COMENTARIO 6.1. *En particular, para las métricas y restricciones del problema de peor caso de optimización y/o estabilidad, se toma la clase generalizada $\mathcal{F}_M$ a contar de la dimensión $d_0 \geq (2K+3)n+2$, resultando el problema semidefinido* (6.12).

$$
\begin{aligned}
\sup_{v \in \mathbb{R}^{(2K+3)n}, X \in \mathbb{S}^{(2K+3)n+2}} \quad & b^\top v + \text{Tr}(CX) \\
\text{s.a} \quad & \text{Tr}(A_{ij}X) + v_j - v_i \leq 0 \quad (\forall i,j \in \mathcal{I}) \\
& \text{Tr}(A_{M_i}X) - M^2 \leq 0 \qquad (\forall i \in I_M) \\
& \text{Tr}(A_R X) - R^2 \leq 0 \\
& \text{Tr}(A_y X) - 1 \leq 0 \\
& \text{Tr}(A_* X) = 0.
\end{aligned}
\tag{6.12}
$$

Se nota que la estructura del problema primal es similar a la presentada en el Capítulo 5, agregándose una restricción de igualdad y teniéndose una elección distinta de parámetros que, sin embargo, representan los mismos tipos de restricción anteriores. El cálculo del problema dual es similar a aquel realizado para modelos *batch* y permite llegar al problema semidefinido (6.13). La prueba del nuevo par primal-dual se expone en el Anexo C.1.

$$
\begin{aligned}
\inf_{\lambda_{ij},\mu_i,\tau,\kappa,\varphi} \quad & \tau R^2 + \sum_{i\in I_M} \mu_i M^2 + \kappa \\
\text{s.a} \quad & \sum_{(i,j)\in\mathcal{I}} \lambda_{ij} A_{ij} + \sum_{i\in I_M} \mu_i A_{M_i} + \tau A_R + \kappa A_y + \varphi A_* - C \succeq O \\
& \sum_{(i,j)\in\mathcal{I}} \lambda_{ij}(e_j - e_i) = b \\
& \lambda_{ij}, \mu_i \geq 0 \quad ((i,j)\in\mathcal{I}) \\
& \tau, \kappa \geq 0 \\
& \varphi \in \mathbb{R}.
\end{aligned}
\tag{6.13}
$$

Este nuevo problema se utiliza para obtener un resultado de dualidad fuerte, basándose en un certificado para el cumplimento de la condición de Slater en el dual. Esto permite garantizar que ambos problemas alcanzan el mismo valor óptimo y que este se alcanza en alguna valuación de X y v primal-factible. Esto se presenta en la Proposición 6.2, siguiendo la línea de las demostraciones para modelos *batch* antes mostradas, por lo que se relega su prueba al Anexo C.2.

Proposición 6.2. *Sea* (5.14) *iSOLG-representación del problema de peor caso conjunto de optimización y estabilidad de una ejecución de K rondas del método de gradiente estocástico incremental o el método IncrementalProx con pérdidas cumpliendo el Supuesto 1a. Sea* (5.15) *su dual semidefinido, donde ambos son problemas factibles. Entonces, ambos poseen un valor óptimo idéntico y existe un punto primal-factible que alcanza tal valor.*

Sin embargo, el resultado ideal y del cual no se tiene claridad, es la existencia de un certificado de Slater primal. La dualidad fuerte para SDP indica que una consecuencia de tal certificado es que el valor dual se alcanza. Una eventual demostración de cota superior ajustada para cualquiera de los problemas que se plantean por este modelo PEP incremental depende de la existencia de un punto exacto donde las variables duales alcancen su óptimo, de acuerdo al trabajo que se realiza al emplear la metodología PEP.

6.2. Implementación

A continuación, se presentan los resultados derivados de implementar el PEP para métodos incrementales (6.12). Se utiliza el *solver* de optimización MOSEK, ejecutando la implementación de cada modelo primal resultante. Se analizan las métricas de solo optimización, solo estabilidad y la métrica conjunta para ambos métodos expuestos, realizando una comparación entre el análisis de la representación del peor caso por separado y en conjunto. Además, para cada algoritmo se expone un resultado de una implementación con distintos números de pasadas.

6.2.1. El método SGD incremental

Se analizan los resultados obtenidos para el método de gradiente estocástico, fijando una elección de $K = n$ pasadas por cada dato y un *learning rate* $\eta = \frac{1}{n\sqrt{2n}}$, que permiten garantizar los órdenes de complejidad discutidos en la Sección 2.2.

Primero, se presenta el resultado para el caso del error de optimización. Se recuerda que para este método se hace uso de la métrica de rendimiento del *gap* de optimalidad, notando que la respuesta entregada es el iterado promedio después de K rondas, $\bar{x}_K$. La comparación se realiza con su cota superior provista por el Lema 2.5:

$$F_{\mathbf{S}}(\bar{x}_K) - F_{\mathbf{S}}(x^*_{\mathbf{S}}) \leq \frac{R^2}{2\eta T} + \frac{M^2\eta(n+2)}{2}.$$

En la Figura 6.1 se aprecia un claro desajuste entre el peor caso real emulado por el PEP y la cota superior teórica del método de gradiente estocástico, que no puede explicarse por la tolerancia del método de optimización semidefinida.

En segundo lugar, se considera el caso del error de estabilidad. Este se mide usando la métrica de UAS para el iterado promedio entre las rondas, manteniendo las mismas cantidades que para el caso de optimización. Esta vez, se presenta en la Figura 6.2 una comparación entre la estabilidad de argumentos derivada del Lema 2.6 y el promedio de

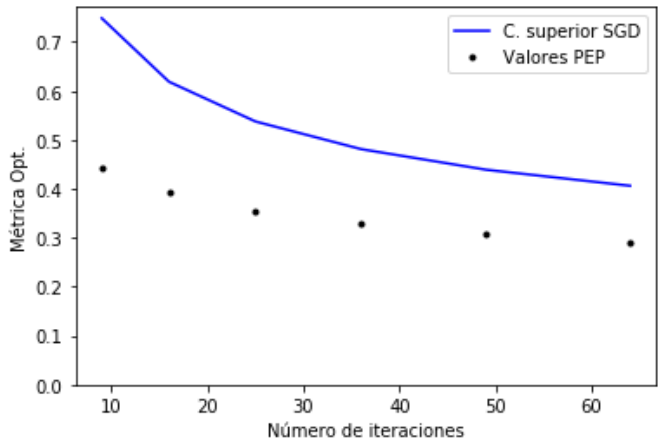

FIGURA 6.1. Comparación del peor caso computado del riesgo empírico con la cota superior teórica después de $K = n$ pasadas del método de SGD incremental.

las distancias entre iterados al final de cada ronda,

$$M||\bar{x}_K - \bar{y}_K|| \leq 2M^2\eta(K+1) + \frac{2M^2\eta}{K}\sum_{k=1}^{K}\sqrt{kn}.$$

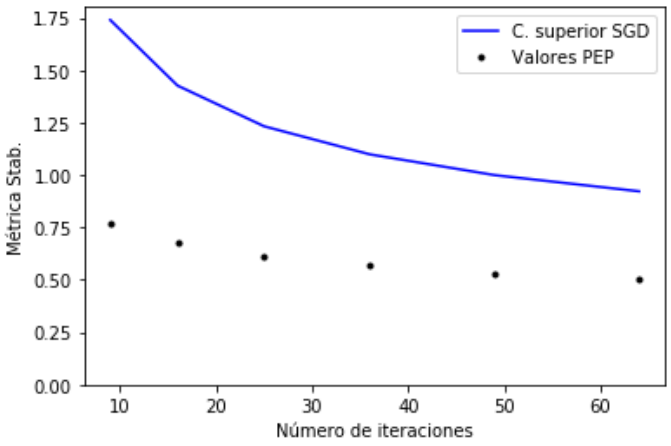

FIGURA 6.2. Comparación del peor caso computado de la métrica de estabilidad con la cota superior teórica después de $K = n$ pasadas del método de gradiente estocástico incremental.

La figura muestra que el error de estabilidad real no se ajusta a la cota teórica, produciendo una situación similar a la expuesta para el caso de optimización. Nuevamente, esta diferencia es suficientemente notoria para asegurar que existe un *gap* entre una respuesta exacta del PEP y la cota teórica.

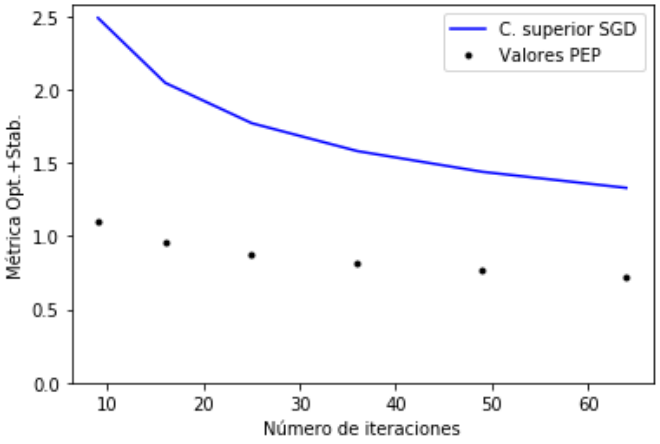

FIGURA 6.3. Comparación del peor caso computado de la métrica conjunta de optimización y estabilidad con la cota superior teórica después de $K = n$ pasadas del método de gradiente estocástico incremental.

Luego, se analiza en conjunto ambos tipos de error, medidos por la suma de ambas métricas anteriores y mayorados por la cota conjunta

$$F_{\mathbf{S}}(\bar{x}_K) - F_{\mathbf{S}}(x^*_{\mathbf{S}}) + M||\bar{x}_K - \bar{y}_K|| \leq \frac{R^2}{2\eta T} + \frac{M^2\eta(n+2)}{2} + 2M^2\eta(K+1) + \frac{2M^2\eta}{K}\sum_{k=1}^{K}\sqrt{kn}.$$

La Figura 6.3 muestra una diferencia esperable, considerando que la elección de un único supremo para el problema conjunto no puede ser estrictamente mayor a la suma de ambos problemas por separado. Más aún, tomar el problema de los errores conjuntos produce un resultado notoriamente mejor que sumar ambos problemas previos, como se puede apreciar en los valores de la Tabla 6.1.

TABLA 6.1. Diferencia entre la suma de los problemas de optimización y estabilidad por separado, y el problema conjunto (Δ) para el problema de peor caso del método de gradiente estocástico con permutación fija después de $K = n$ pasadas, para distintos tamaños de muestra.

n	T	Δ
3	9	0,11348985
4	16	0,10978327
5	25	0,09516638
6	36	0,08403393
7	49	0,07334893
8	64	0,06530311

Por último, se introduce una comparación de problemas de peor caso para distintos números de pasadas, entre 1 y n. Se analiza la métrica conjunta de optimización y generalización, tomando en cuentra cuatro casos. Se toman ambos casos límite, donde para una sola pasada no es posible garantizar convergencia del error de optimización y para n rondas se tiene la convergencia discutida en la Sección 2.2. Se añaden otros dos casos que cumplen las garantías de convergencia, de $\lceil \frac{n}{2} \rceil$ y $\lceil \frac{n}{4} \rceil$ rondas. Todos los casos mencionados utilizan la elección de *learning rate* $\eta = \frac{1}{\sqrt{Kn(K+n)}}$.

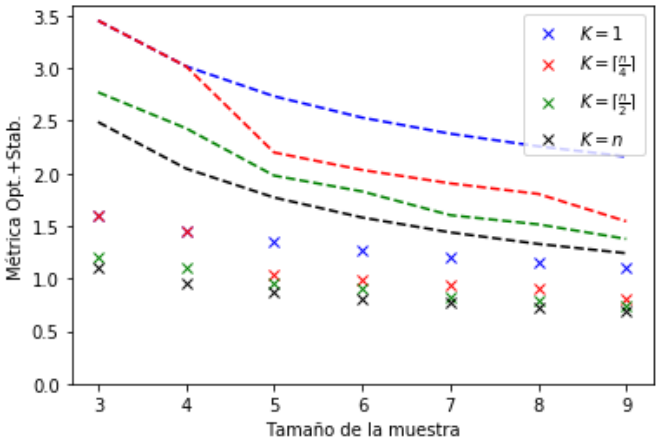

FIGURA 6.4. Peor caso computado de la métrica conjunta de optimización y estabilidad (cruces) con la cota superior teórica (líneas discontinuas) para SGD incremental con distintos números de pasadas K.

Si bien se quiso incorporar algún régimen intermedio, por ejemplo $\sqrt{n}$ pasadas, en la práctica esto no fue viable. El tamaño máximo de los problemas introducidos y el número de datos resultantes para cada K elegido son muy bajos para realizar una buena comparación, pudiendo implementarse tales problemas solo para PEP de tamaños pequeños. Sin embargo, en la Figura 6.4 se nota diferencias pequeñas a medida que se ejecuta una cantidad mayor de pasadas. Esto sugiere que la mejora obtenida de reutilizar los datos, en términos de la métrica conjunta, es cada vez más pequeña.

6.2.2. El método *IncrementalProx*

A continuación, se muestran los resultados obtenidos para el método proximal *IncrementalProx*, manteniendo la elección de n pasadas por cada dato y la elección de *learning*

rate $\eta = \frac{1}{n\sqrt{2n}}$ del método anterior. Esta elección es congruente con lo discutido en la Sección 2.4, entregando una complejidad muestral $O\left(\frac{1}{\varepsilon^2}\right)$, al igual que los resultados presentados para métodos anteriores.

En primer lugar, se hace el análisis de la implementación para el error de optimización. Se hace uso de la métrica de rendimiento del *gap* de optimalidad, notando que la respuesta del algoritmo es idéntica al otro método incremental mostrado. Esta vez, la cota superior comparable viene dada por el resultado de la Proposición 2.1, que es la desigualdad

$$F_{\mathbf{S}}(\bar{x}_K) - F_{\mathbf{S}}(x^*_{\mathbf{S}}) \leq \frac{R^2}{2\eta T} + \frac{M^2\eta n}{2}.$$

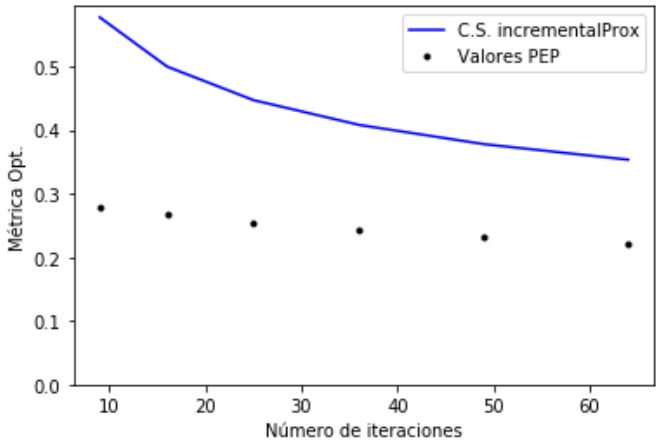

FIGURA 6.5. Comparación del peor caso computado del riesgo empírico con la cota superior teórica después de $K = n$ pasadas del método *IncrementalProx*.

La Figura 6.5 muestra la comparación del peor caso devuelto por la implementación computacional y la cota superior de la Proposición 2.1. Se puede ver una diferencia notoria entre ambos valores, concluyendo que esta cota no es ajustada.

Luego, se considera la métrica de estabilidad. De igual manera que casos anteriores, ésta utiliza la forma de la variable aleatoria proveniente de la estabilidad de argumentos. Esta vez se presenta en la Figura 6.2 una comparación entre la estabilidad de argumentos derivada de la Proposición 2.2 y el peor caso real de la estabilidad,

$$M||\bar{x}_K - \bar{y}_K|| \leq M^2\eta(K+1).$$

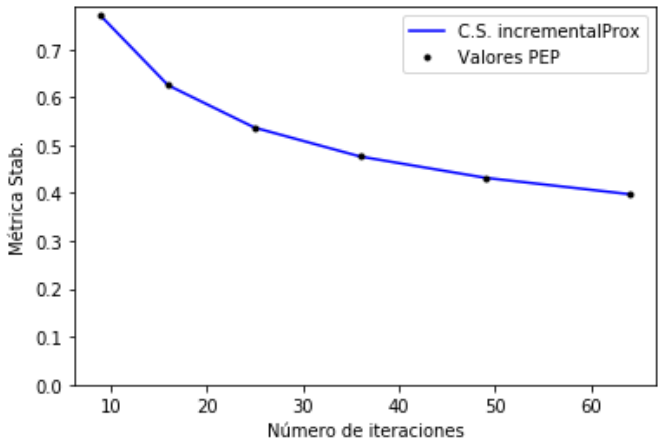

FIGURA 6.6. Comparación del peor caso computado de la métrica de estabilidad con la cota superior teórica después de $K = n$ pasadas del método *Incremental-Prox*.

Esta figura muestra que el peor caso se ajusta a la cota superior de manera similar al método proximal *batch*, como se mostró en el Capítulo 5. Las diferencias del gráfico presentado son de tamaño menor a 10^{-7}, lo que es atribuible a la tolerancia del método de punto interior que resuelve el programa semidefinido.

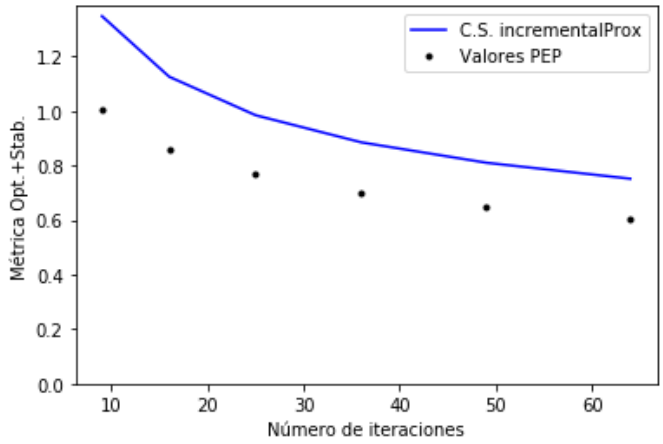

FIGURA 6.7. Comparación del peor caso computado de la métrica conjunta de optimización y estabilidad con la cota superior teórica después de $K = n$ pasadas del método *IncrementalProx*.

A continuación se expone el caso de los errores de optimización y estabilidad conjuntos, considerando la cota superior resultante de las dos proposiciones anteriores:

$$F_{\mathbf{S}}(\bar{x}_K) - F_{\mathbf{S}}(x^*_{\mathbf{S}}) + M||\bar{x}_K - \bar{y}_K|| \leq \frac{R^2}{2\eta T} + \frac{M^2\eta n}{2} + M^2\eta(K+1).$$

La Figura 6.7 muestra una diferencia esperada por el desajuste de la cota de optimización, pero existe una disminución relevante en los valores del problema conjunto con respecto a la suma de los problemas de peor caso real separados, como se puede ver en los valores de la Tabla 6.2.

TABLA 6.2. Diferencia entre la suma de los problemas de optimización y estabilidad por separado, y el problema conjunto (Δ) para el problema de peor caso del método *IncrementalProx* con permutación fija después de $K = n$ pasadas, para distintos tamaños de muestra.

n	T	Δ
3	9	0,04391967
4	16	0,03134486
5	25	0,02471281
6	36	0,02020384
7	49	0,01641539
8	64	0,01420947

Finalmente, se introduce la comparación de problemas de peor caso para distintos números de pasadas. Se analiza la métrica conjunta de optimización y generalización, tomando en cuenta la misma elección de parámetros de la sección anterior. Se recuerda que estos corresponden a una, $\lceil \frac{n}{4} \rceil, \lceil \frac{n}{2} \rceil$ y n pasadas respectivamente, además de la elección de largo de paso $\eta = \frac{1}{\sqrt{Kn(K+n)}}$ en cada caso. El análisis de las cotas superiores permite concluir convergencia solo para las elecciones de múltiples pasadas.

Sin embargo, existen elecciones distintas de K con garantías de convergencia, que de la misma manera que en el símil de este experimento para el método SGD incremental, no pudieron implementarse por las razones prácticas ya expuestas. La Figura 6.4 nuevamente indica diferencias pequeñas entre los valores de peor caso a medida que se ejecuta una cantidad mayor de pasadas, reafirmándose la hipótesis para este método también.

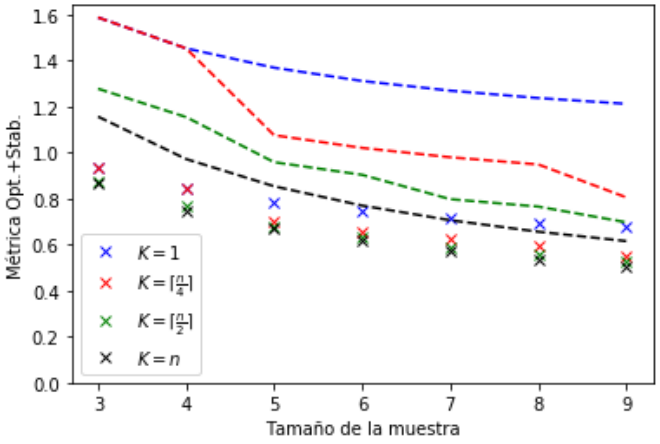

FIGURA 6.8. Peor caso computado de la métrica conjunta de optimización y estabilidad (cruces) con la cota superior teórica (líneas discontinuas) para distintos números de pasadas K del método *IncrementalProx*.

7. DISCUSIÓN

En esta sección se presenta una breve discusión de los resultados obtenidos, separándolos en tres áreas:

7.1. Análisis de SCO mediante modelos PEP

El modelo desarrollado para algoritmos *batch* permite la representación de peor caso real de convergencia del funcional de riesgo empírico y de la estabilidad de los métodos utilizados a través de un problema semidefinido, considerando distintos algoritmos con actualizaciones de gradiente o paso proximal. Para el modelo incremental, en cambio, se trabaja bajo la conjetura que el peor caso se obtiene con una discrepancia en la primera actualización, representando la diferencia entre el peor caso y las cotas superiores (citadas o desarrolladas, según corresponda) bajo esta restricción. La implementación de tales problemas permite diferenciar de manera clara entre casos donde la teoría entrega una cota superior ajustada o no. A continuación, se discuten los resultados teóricos desarrollados, datos obtenidos de la implementación y se introducen ciertas interrogantes que se desprenden de la aplicación de la metodología PEP al problema dual, para aquellos casos donde se realizó.

Primero, se cuestiona si será posible recuperar demostraciones de aquellos métodos *batch* donde la cota no es ajustada y si, en caso de ser cierto, será posible obtener una cota más ajustada. De acuerdo a lo expuesto, resta aplicar la metodología PEP completa a los métodos de gradiente y de subgradiente, los que muestran una diferencia significativa entre la cota superior provista por la teoría y el resultado de peor caso computado. En caso de no poder obtenerse cotas ajustadas por este medio, se debe notar que no existen garantías de que una elección óptima del *learning rate* de la cota superior sea también el óptimo para la elección de paso en el peor caso real. Esto es, dado que existe una diferencia entre la cota superior y la representación del peor caso del PEP utilizado, puede que la cota superior tenga distinto ajuste para elecciones distintas de parámetros, provocando una discrepancia entre elecciones de parámetros óptimos de la cota superior y del peor caso efectivo.

En cambio, para el caso del PPM se analizaron los problemas de optimización y estabilidad separados a través de la metodología PEP. En el primer caso, la aplicación de este proceso permitió recuperar la cota óptima en de la Demostración 5.2.3. Para la estabilidad, sin embargo, se llegó a un resultado parcial en base al que se plantea la Conjetura 5.1. La aplicación de la metodología completa se vio limitada por dos factores: La demostración de la factibilidad dual, donde el cumplimiento de la restricción semidefinida no es claro (pese a la evidencia computacional que la soporta) y por el proceso de reinterpretar la combinación de desigualdades, donde la utilización de la variable auxiliar y en relación a las variables y desigualdades recuperadas no es clara.

Se propone un acercamiento alternativo a este problema, que descarta el uso de la variable y, perdiendo la posibilidad de representar el problema conjunto. Esta alternativa utilizaría la métrica de estabilidad de argumentos al cuadrado,

$$M^2||\mathcal{A}(\mathbf{S}) - \mathcal{A}(\mathbf{S}')||^2,$$

cuya representación semidefinida se logra como una combinación solamente de subgradientes representados en X.

Además, se nota el experimento de las Figuras 6.4 y 6.8, que muestran diferencias más pequeñas en el peor caso a medida que aumenta el número de pasadas realizadas por los distintos métodos incrementales. No obstante, este análisis se realiza solo para tamaños de muestra pequeños debido a las limitaciones de la implementación de los modelos PEP desarrollados, no pudiendo realizarse un estudio más completo del crecimiento de distintos regímenes de K entre 1 y n. Queda la interrogante de si en la práctica es posible obtener garantías de convergencia de los errores de estabilidad, de optimización y, consecuentemente, del exceso de riesgo para una cantidad menor de pasadas a las n fijadas para el resto de los experimentos de métodos incrementales.

Finalmente, se reitera que todos los resultados fueron obtenidos simplificando al caso irrestricto, por motivos de tiempo de inicialización y de cómputo de la implementación. Sin embargo, en ambos casos la extensión al caso proyectado necesita representar una

función indicatriz adicional en el problema semidefinido, introduciendo una cantidad de columnas $\Theta(T)$ extra a la matriz de Gram, además de restricciones de radio para cada iterado y de interpolación adecuadas a la nueva función, siguiendo el trabajo de Taylor et al. (2017a).

7.2. Generalización de FOM proximales con *minibatch*

Respecto a los algoritmos proximales utilizados, se nota una diferencia en la convergencia del método de punto proximal y su variante incremental de acuerdo a la teoría de cotas asintóticas, como se aprecia en la Tabla 3.2. Estos métodos tienen, sin embargo, garantías de estabilidad del mismo orden de acuerdo al Corolario 3.1, siendo ambos casos límite de un algoritmo *minibatch*-prox (con b *minibatchs* fijos en cada ronda) cuyos peores casos reales se ajustan a cada cota superior respectiva. Una pregunta derivada es si los resultados pueden mejorarse utilizando una elección de *minibatch* distinta de los casos límite, especialmente en términos del peor caso real, ya que los métodos implementados no revelan información sobre los casos intermedios. Esta interrogante podría resolverse mediante la implementación de un modelo PEP adecuado para métodos *minibatch*.

¿Cómo podría representarse este modelo? La representabilidad de una suma de funciones se desprende de lo realizado para el método *batch*, mientras que la elección de los *minibatchs* puede realizarse siguiendo el modelo incremental, pudiendo representarse el peor caso mediante el máximo de n/b problemas que representan el peor caso de trayectorias con separación en alguna actualización específica. De acuerdo con esto, se podría plantear tal modelo como una extensión natural de los dos modelos desarrollados en este trabajo.

Sin embargo, la implementación de $n+1$ funciones que comparten múltiples puntos necesita una gran cantidad de recursos computacionales, tanto en memoria como tiempo de cómputo, considerando además la representación de n/b problemas distintos o necesitando probarse algún símil a la Conjetura 6.1. Lo primero implica que la representación

de un modelo *minibatch* generalizado sin el resultado es mucho más costosa que la representación de los casos límite, generando un problema práctico en la aplicación de tal metodología.

7.3. Gradiente estocástico con varianza reducida

Para el método SVRG, se introdujo un resultado de estabilidad que, dependiendo del condicionamiento del problema, puede determinar si se alcanza una estabilidad mejor o peor que SGD. Sin embargo, se discutió en el Capítulo 4 que esta dependencia del condicionamiento no permite su aplicación a la descomposición de riesgo para pérdidas suaves regularizadas. Luego, añadiendo la reducción de varianza y regularización *dummy* no se garantiza una mejoría con respecto a SGD en la aproximación a este caso de SCO.

Sin embargo, en este trabajo se realiza un análisis solo para las versiones originales del algoritmo de Johnson y Zhang. En la literatura existen modificaciones del método con promedio al interior de cada ronda, actualizaciones proximales y/o esquemas de aceleración mediante momento negativo (véase Allen-Zhu y Yuan (2016),Allen-Zhu (2017)); posiblemente mejorando la estabilidad alcanzada.

7.4. Trabajo futuro

Considerando las preguntas adicionales planteadas en la discusión anterior y derivadas de las preguntas de investigación, se proponen las siguientes lineas de trabajo futuro.

Primero, si será posible una representación semidefinida unificada del modelo incremental equivalente al problema de peor caso. Para esto se requiere demostrar la Conjetura 6.1 propuesta o desarrollar un modelo semidefinido manejable que permita la representación del peor caso sin la restricción de separar trayectorias en la primera iteración. Para ser aplicable en la práctica, este nuevo modelo debe ser distinto a la representación de los n pares de trayectorias de peor caso a través del máximo de n problemas semidefinidos distintos, debido a la dificultad de escalar el problema con el número de muestras.

En segundo lugar, la extensión natural de ambos modelos al *setting* proyectado. En particular, se destaca el interés por el caso donde $\mathcal{X}$ es un compacto de radio R, lo que permite trivializar el Supuesto 2 y obtener garantías de aproximación y generalización en términos de esta cantidad. Como se mencionó con anterioridad, incorporar la indicatriz de $\mathcal{X}$ incrementa notoriamente el tamaño del problema, complejizando su aplicación práctica y restringiéndola a rangos de parámetros más acotados, asumiendo los mismos recursos computacionales. Esto, a su vez, complica la aplicación de la metodología PEP, que depende de múltiples ejecuciones del problema semidefinido.

Por último, notando la diferencia encontrada en la práctica para los métodos incrementales se puede plantear la siguiente pregunta que escapa de las cotas asintóticas de convergencia conocidas. ¿Será necesario en la práctica $\Theta(n)$ utilizaciones de cada dato para lograr la tasa óptima de convergencia del riesgo empírico? Según lo discutido en la Sección 7.2, debe realizarse un estudio detenido de la influencia del número de pasadas con la capacidad de computar PEP de dimensiones más altas para probar distintos regímenes de crecimiento de K en función de la cantidad de datos.

8. CONCLUSIONES

En este trabajo se estudió el problema de optimización convexa estocástica (1.1) a través del problema de estimación de rendimiento, planteándose la hipótesis que el uso de esta representación de peor caso permite recuperar cotas superiores no asintóticas para los errores de optimización y estabilidad asociados a (1.1).

De acuerdo con el primer objetivo planteado para el estudio de esta, se crearon dos modelos semidefinidos diferentes, aprovechando las características de las actualizaciones de los distintos métodos. Se concluyó la equivalencia del problema semidefinido para algoritmos *batch* al peor caso real, lograda introduciendo una noción alternativa de Gram-representabilidad que incluye el fenómeno de estabilidad emulando dos trayectorias vecinas y está especializada para este tipo de algoritmos. Para el modelo incremental, en cambio, se logró demostrar la equivalencia de la misma forma, restringida bajo un supuesto sobre el peor caso que se conjetura es cierto. Habiendo desarrollado ambos métodos, se estudió su dualidad Lagrangiana en correspondencia con el segundo objetivo, alcanzando resultados de dualidad fuerte para ambos problemas y obteniendo representaciones explícitas para los duales asociados a ambos programas matriciales.

Después, se implementaron los problemas utilizando el *solver* de optimización MOSEK. Los resultados permitieron validar el modelo desarrollado, obteniendo diferencias entre el peor caso real del problema y las cotas superiores provenientes de la teoría. Se pudo diferenciar claramente en cada caso si el peor caso real es ajustado al resultado de la teoría o no, y si existe una diferencia entre el peor caso de las métricas de optimización y estabilidad por separado o en conjunto.

Luego, se utilizó la metodología PEP sobre los problemas separados de optimización y estabilidad para el método de punto proximal. Esto permitió evidenciar un caso donde la metodología funciona correctamente y un caso donde falla, pudiendo explorarse las limitaciones prácticas que ésta tiene. En particular, se concluye que la metodología se ve limitada por no siempre existir una relación fácil de establecer entre las distintas variables

del problema y por el proceso de demostrar la factibilidad del punto óptimo recuperado heurísticamente de los resultados computacionales restringidos.

Además, se complementa la teoría con una cota de estabilidad para métodos proximales generalizados al caso *minibatch* con permutación fija, idéntica para cualquier elección de tamaño *batch* y aplicable al cálculo de la estabilidad del PPM. También, se mostró una cota de estabilidad para la aplicación de SVRG al caso fuertemente convexo dependiente del condicionamiento propio del problema, concluyendo que no tiene utilidad en mejorar la convergencia del exceso de riesgo mediante el uso de un regularizador *dummy*.

En general, se concluye que la metodología puede ser aplicada al problema de convergencia del exceso de riesgo, pudiendo desarrollarse modelos para este *setting* más estructurado que el original de optimización convexa, pero posee limitantes prácticas que dificultan recuperar una demostración válida. Estas dificultades, sin embargo, podrían sobrellevarse mediante algún planteamiento alternativo como el que se discute para el caso de sólo estabilidad, generando un *trade-off* entre la aplicación del problema ideal (como resolver el PEP para los errores conjuntos) y una aplicación que sea efectiva en la práctica de los errores separados de optimización y estabilidad. De cualquier manera, este acercamiento alternativo constituye un apoyo para el desarrollo de garantías sobre los tipos de errores de aprendizaje y que se enfoca en casos más pequeños de los parámetros de cada algoritmo, complementando el tratamiento asintótico de los problemas que se enfoca en el comportamiento del número de muestras e iteraciones al límite.

BIBLIOGRAFIA

Allen-Zhu, Z. (2017). Katyusha: The first direct acceleration of stochastic gradient methods. *The Journal of Machine Learning Research*, *18*(1), 8194–8244.

Allen-Zhu, Z., y Yuan, Y. (2016). Improved svrg for non-strongly-convex or sum-of-non-convex objectives. En *International conference on machine learning* (pp. 1080–1089).

Amir, I., Koren, T., y Livni, R. (2021). Sgd generalizes better than gd (and regularization doesn't help). *arXiv preprint arXiv:2102.01117*.

ApS, M. (2021). MOSEK Fusion API for .NET 9.2.47 [Manual de software informático]. Descargado de `https://docs.mosek.com/9.2/dotnetfusion/index.html`

Bassily, R., Feldman, V., Guzmán, C., y Talwar, K. (2020). Stability of stochastic gradient descent on nonsmooth convex losses. *arXiv preprint arXiv:2006.06914*.

Bassily, R., Feldman, V., Talwar, K., y Thakurta, A. (2019). Private stochastic convex optimization with optimal rates. *arXiv preprint arXiv:1908.09970*.

Beck, A. (2017). *First-order methods in optimization*. SIAM.

Bertsekas, D. P. (2011). Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. *Optimization for Machine Learning*, *2010*(1-38), 3.

Bousquet, O., y Elisseeff, A. (2002). Stability and generalization. *The Journal of Machine Learning Research*, *2*, 499–526.

Bousquet, O., Klochkov, Y., y Zhivotovskiy, N. (2020). Sharper bounds for uniformly stable algorithms. En *Conference on learning theory* (pp. 610–626).

Drori, Y., y Teboulle, M. (2014). Performance of first-order methods for smooth convex minimization: a novel approach. *Mathematical Programming*, *145*(1), 451–482.

Güler, O. (1991). On the convergence of the proximal point algorithm for convex minimization. *SIAM Journal on Control and Optimization*, *29*(2), 403–419.

Hardt, M., Recht, B., y Singer, Y. (2016). Train faster, generalize better: Stability of stochastic gradient descent. En *International conference on machine learning* (pp. 1225–1234).

Johnson, R., y Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, *26*, 315–323.

Nedic, A., y Bertsekas, D. P. (2001). Incremental subgradient methods for nondifferentiable optimization. *SIAM Journal on Optimization*, *12*(1), 109–138.

Nemirovski, A. S., y Yudin, D. B. (1978). Cesari convergence of the gradient method of approximating saddle points of convex-concave functions. En *Doklady akademii nauk* (Vol. 239, pp. 1056–1059).

Polyak, B. T., y Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, *30*(4), 838–855.

Robbins, H., y Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, 400–407.

Rockafellar, R. T. (1976). Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, *14*(5), 877–898.

Shalev-Shwartz, S., y Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Shalev-Shwartz, S., Shamir, O., Srebro, N., y Sridharan, K. (2010). Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, *11*, 2635–2670.

Shor, N. Z. (2012). *Minimization methods for non-differentiable functions* (Vol. 3). Springer Science & Business Media.

Taylor, A., y Bach, F. (2019). Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions. En *Conference on learning theory* (pp. 2934–2992).

Taylor, A., Hendrickx, J., y Glineur, F. (2017a). Exact worst-case performance of first-order methods for composite convex optimization. *SIAM Journal on Optimization*, *27*(3), 1283–1313.

Taylor, A., Hendrickx, J., y Glineur, F. (2017b). Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, *161*(1-2), 307–345.

Wang, J., Wang, W., y Srebro, N. (2017). Memory and communication efficient distributed stochastic optimization with minibatch prox. En *Conference on learning theory* (pp. 1882–1919).

ANEXOS

ANEXO A. RESULTADOS COMPLEMENTARIOS PARA SVRG

En este anexo se presentan resultados complementarios a los expuesto en la Sección 2.5 y el Capítulo 4, sobre el método con varianza reducida. Las Proposiciones A.1 y A.2 presentan una cuantificación de la estabilidad de argumentos para el método SVRG con pérdidas suaves, basado en las dos versiones originales planteadas por Johnson y Zhang (2013) en el trabajo seminal sobre este método (Estas son los Algoritmos 8 y 9). Si bien el caso suave no tiene garantías de convergencia directa, estos análisis proveen una intuición de cómo se altera la estabilidad del método de gradiente estocástico por la introducción de una reducción de varianza.

Además, se presenta un análisis de la estabilidad de argumentos para el caso de pérdidas fuertemente convexas, utilizando la versión alternativa de SVRG con elección determinista al final de cada ronda (Johnson y Zhang, 2013). Ésta se formaliza en el Algoritmo 9. Igualmente, pese a que los autores no plantean una garantía para la convergencia del error de optimización, el análisis de este nuevo método corresponde a una simplificación del análisis para el Algoritmo 8, mostrando cómo afecta la estabilidad de argumentos la introducción de la elección estocástica al final de cada ronda. También la elección determinística es interesante debido a su uso en algoritmos proximales desarrollados posteriormente, basados en esta versión original de Johnson y Zhang. En particular, SVRG++ (Allen-Zhu y Yuan, 2016) y *Katyusha* (Allen-Zhu, 2017) fijan la evaluación del gradiente completo sobre iterados promedio de cada ronda, para distintas elecciones de ponderadores.

En primer lugar, se presenta el resultado restante para la versión ya presentada de SVRG:

Proposición A.1. *Sea una pérdida $f(\cdot, \xi) \in \mathcal{F}_{M,L}(\mathbb{R}^d)$ para todo $\xi \in \mathcal{Z}$ (Supuesto 1b). Luego, para $\eta \leq \frac{2}{L}$ el Algoritmo 8 tiene uniforme estabilidad de argumentos $O\left(\frac{\eta MT}{n} \cdot m^K\right)$.*

DEMOSTRACIÓN. *Se analiza una actualización de SVRG. Separando convenientemente la norma por desigualdad triangular:*

$$\begin{aligned}||x_{t+1}^k - y_{t+1}^k|| \leq & \left|\left|x_t^k - y_t^k - \eta\left(\nabla f_{i_t}(x_t^k) - \nabla f'_{i_t}(y_t^k)\right)\right|\right| \\ & + \eta\left|\left|\left(\nabla f_{i_t}(\tilde{x}_k) - \nabla f'_{i_t}(\tilde{y}_k\right) - \left(\nabla F_S(\tilde{x}_k) - \nabla F_{S'}(\tilde{y}_k)\right)\right|\right|.\end{aligned}$$

Más aún, sumando cero para después separar las diferencias de gradientes en el segundo término:

$$\begin{aligned}\leq & \left|\left|x_t^k - y_t^k - \eta\left(\nabla f_{i_t}(x_t^k) - \nabla f'_{i_t}(y_t^k)\right)\right|\right| \\ & + \left|\left|\tilde{x}_k - \tilde{y}_k - \eta\left(\nabla f_{i_t}(\tilde{x}_k) - \nabla f'_{i_t}(\tilde{y}_k\right)\right|\right| \\ & + ||\tilde{x}_k - \tilde{y}_k - \eta\left(\nabla F_S(\tilde{x}_k) - \nabla F_{S'}(\tilde{y}_k)\right)||.\end{aligned}$$

Se nota que todos los términos corresponden a actualizaciones del tipo (2.10) *para el método SGD con muestreo con reemplazo o actualizaciones del método de gradiente* (2.9)*. Luego, para $\eta \leq \frac{2}{L}$ se tiene la no-expansividad de las actualizaciones de tipo gradiente, aplicable a ambos casos anteriores. Utilizando también la M-Lipschitz continuidad, se concluye la desigualdad*

$$\delta_{t+1}^k \leq \delta_t^k + 2\eta M \boldsymbol{r}_{\boldsymbol{i}_{k,t}} + \delta_1^k + 2\eta M \boldsymbol{r}_{\boldsymbol{i}_{k,t}} + \delta_1^k + \frac{2\eta M}{n}.$$

Donde $\boldsymbol{r}_{\boldsymbol{i}_{k,t}} \sim$ Bern$(\frac{1}{n})$, debido a la muestra elegida por el algoritmo en la iteración t. Tomando valor esperado sobre la aleatoriedad de $\mathcal{A}_{SVRG}$,

$$\mathbb{E}\delta_{t+1}^k \leq \mathbb{E}\delta_t^k + 2\mathbb{E}\delta_1^k + \frac{6\eta M}{n}.$$

Además, se puede observar que la primera iteración al iniciar cada ronda corresponde a un paso de GD. Por lo tanto se cumple

$$\delta_2^k \leq \delta_1^k + \frac{2\eta M}{n}.$$

Sumando desigualdades sobre el ciclo interior del algoritmo, se obtiene la desigualdad

$$\mathbb{E}\delta_{i+1}^k \leq \mathbb{E}\delta_1^k + \frac{2\eta M}{n}\mathbf{1}_{\{i\geq 1\}} + (i-1)\left[2\mathbb{E}\delta_1^k + \frac{6\eta M}{n}\right]\mathbf{1}_{\{i\geq 2\}}.$$

Luego, el iterado inicial de la ronda siguiente es tal que

$$\begin{aligned}\mathbb{E}\delta_1^{k+1} &= \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\delta_i^k \\ &\leq \mathbb{E}\delta_1^k + \frac{2\eta M(m-1)}{mn} + \left[2\mathbb{E}\delta_1^k + \frac{6\eta M}{n}\right]\cdot\sum_{i=3}^{m}\frac{i-2}{m} \\ &= \left[\frac{(m-1)(m-2)}{m} + 1\right]\mathbb{E}\delta_1^k + \frac{\eta M(m-1)(6m-4)}{mn}.\end{aligned}$$

Finalmente, sumando para el ciclo exterior del algoritmo se cumple

$$\Rightarrow \mathbb{E}\delta_1^{K+1} \leq \frac{\eta M(m-1)(6m-4)}{mn}\sum_{i=0}^{K-1}\left[\frac{(m-1)(m-2)}{m} + 1\right]^i,$$

alcanzando el orden de estabilidad enunciado. □

Se puede apreciar que la garantía de estabilidad uniforme empeora notoriamente, teniéndose una dependencia del número de pasos por ronda m en el factor exponencial en K. Como se discutió para el caso bajo el Supuesto 1c, una buena elección de m (y la más usada en la práctica) es de orden $\Theta(n)$. A su vez, esta elección permite comparar el resultado de la Proposición 4.1 con esta nueva garantía. Se nota que a diferencia del nuevo resultado, la garantía del Capítulo 4 puede controlar el crecimiento de la cota de estabilidad con una elección de *learning rate* apropiada e incluso demostrar convergencia en algunos casos. Esto se debe a que en el *setting* fuertemente convexo se puede asegurar la contractividad de una actualización de (sub)gradiente, pudiendo deducirse aun en algunos casos la contractividad de la ronda completa.

A continuación, se presenta la versión alternativa al método SVRG descrito en el Capítulo 4. Ésta corresponde a la elección alternativa expuesta en la Sección 2.5, donde

al final de cada ronda se guarda la información del último iterado para su utilización en el cálculo del gradiente de la ronda siguiente. Este método se formaliza en el Algoritmo 9.

Algoritmo 9: $\mathcal{A}_{\text{SVRG2}}$: Método SVRG con elección determinista de último iterado.

Input: Muestra $S = (\xi_1, ..., \xi_n) \in \mathcal{Z}^n$, número de rondas K, pasos del ciclo interno m, *Learning rate* η, iterado inicial $\tilde{x}_1 \in \mathcal{X}$;

for $k = 1...K$ **do**
  Define $\mu_k := \nabla F_S(\tilde{x}_k)$;
  Define $x_1^k := \tilde{x}_k$;
  for $j = 1...m$ **do**
    Muestrea $\mathbf{i}_{k,j} \sim \text{Unif}[n]$;
    Define $x_{j+1}^k := x_j^k - \eta \nabla f(x_j^k, \xi_{\mathbf{i}_{k,j}}) + \eta \nabla f(\tilde{x}_k, \xi_{\mathbf{i}_{k,j}}) - \eta \mu_k$;
  end
  Define $\tilde{x}_{k+1} := x_{m+1}^k$;
end
return $\tilde{x}_{K+1}$

Se presentan dos resultados de uniforme estabilidad de argumentos para los casos suave y fuertemente convexo, respectivamente. Por simplicidad, se consideran las funciones f_i y f_i' correspondientes a las instanciaciones de la pérdida para las muestras $\mathbf{S} \simeq \mathbf{S}'$ que difieren, sin pérdida de generalidad, en el primer dato. Además, se consideran los iterados x_i^k y y_i^k; $\tilde{x}_k$ y $\tilde{y}_k$ respectivos.

Proposición A.2. *Sea una pérdida $f(\cdot, \xi) \in \mathcal{F}_{M,L}(\mathbb{R}^d)$ para todo $\xi \in \mathcal{Z}$ (Supuesto 1b). Luego, para $\eta \leq \frac{2}{L}$ el Algoritmo 9 tiene uniforme estabilidad de argumentos $O\left(\frac{\eta MT}{n} \cdot m^K\right)$.*

DEMOSTRACIÓN. *Se analiza una actualización de SVRG. Separando convenientemente la norma de la manera mostrada para el Algoritmo 8 en la proposición anterior, se*

llega a la expresión:

$$\begin{aligned}||x_{t+1}^k - y_{t+1}^k|| &\leq \left|\left|x_t^k - y_t^k - \eta\left(\nabla f_{i_t}(x_t^k) - \nabla f'_{i_t}(y_t^k)\right)\right|\right| \\ &\quad + \eta\left|\left|\left(\nabla f_{i_t}(\tilde{x}_k) - \nabla f'_{i_t}(\tilde{y}_k\right) - \left(\nabla F_S(\tilde{x}_k) - \nabla F_{S'}(\tilde{y}_k)\right)\right|\right| \\ &\leq \left|\left|x_t^k - y_t^k - \eta\left(\nabla f_{i_t}(x_t^k) - \nabla f'_{i_t}(y_t^k)\right)\right|\right| \\ &\quad + \left|\left|\tilde{x}_k - \tilde{y}_k - \eta\left(\nabla f_{i_t}(\tilde{x}_k) - \nabla f'_{i_t}(\tilde{y}_k\right)\right|\right| \\ &\quad + ||\tilde{x}_k - \tilde{y}_k - \eta\left(\nabla F_S(\tilde{x}_k) - \nabla F_{S'}(\tilde{y}_k)\right)||\,.\end{aligned}$$

Se nota que nuevamente los términos presentan normas asociadas a actualizaciones de (2.10) *con muestreo con reemplazo o actualizaciones del método de gradiente. Luego, para* $\eta \leq \frac{2}{L}$ *se tiene la no-expansividad de cada paso de gradiente. Además, por la* M*-Lipschitz continuidad, se concluye:*

$$\delta_{t+1}^k \leq \delta_t^k + 2\eta M \boldsymbol{r}_{\boldsymbol{i}_{k,t}} + \delta_1^k + 2\eta M \boldsymbol{r}_{\boldsymbol{i}_{k,t}} + \delta_1^k + \frac{2\eta M}{n},$$

donde $\boldsymbol{r}_{\boldsymbol{i}_{k,t}} \sim$ *Bern*$(\frac{1}{n})$ *debido a la muestra elegida por el algoritmo. Tomando valor esperado sobre la aleatoriedad de* $\mathcal{A}_{SVRG2}$*, se tiene para los pasos posteriores al primero:*

$$\mathbb{E}\delta_{t+1}^k \leq \mathbb{E}\delta_t^k + 2\mathbb{E}\delta_1^k + \frac{6\eta M}{n}.$$

Sumando desigualdades sobre el ciclo interior del algoritmo, se obtiene

$$\mathbb{E}\delta_1^{k+1} = \mathbb{E}\delta_{m+1}^k \leq (2m-1)\mathbb{E}\delta_1^k + \frac{(6m-4)\eta M}{n}.$$

Finalmente, sumando en el ciclo exterior del algoritmo:

$$\Rightarrow \mathbb{E}\delta_1^{K+1} \leq \frac{(6m-4)\eta M}{n}\sum_{i=0}^{K-1}(2m-1)^i \in O\left(\frac{m\eta M}{n}\cdot Km^K\right).$$

□

Por lo tanto, el Algoritmo 9 alcanza uniforme estabilidad $O\left(\frac{\eta M^2 T}{n} \cdot m^K\right)$ después de $T = Km$ iteraciones idéntica al *setting* suave para la elección del Algoritmo 8. Finalmente, se presenta el símil de la Proposición 4.1 para la versión alternativa del método, que alude al caso suave y fuertemente convexo.

Proposición A.3. *Sea una pérdida $f(\cdot, \xi) \in \mathcal{S}^{\mu}_{M,L}(\mathbb{R}^d)$ para todo $\xi \in \mathcal{Z}$ (Supuesto 1c) y $m \in \Theta(n)$. Luego:*

- *Para $\eta \in \left[\frac{2}{3\mu}, \frac{1}{L}\right]$, el Algoritmo 9 tiene uniforme estabilidad de argumentos $O\left(\frac{\eta M}{n}\right)$.*
- *Para $\eta < \frac{2}{3\mu}$, el Algoritmo 9 tiene uniforme estabilidad de argumentos $O\left(\frac{M}{\mu n} \cdot \left(\frac{1}{\eta\mu}\right)^K\right)$.*

DEMOSTRACIÓN. *Se toma la desigualdad obtenida en la demostración de la Proposición A.2:*

$$\begin{aligned} ||x^k_{t+1} - y^k_{t+1}|| \leq & \left|\left|x^k_t - y^k_t - \eta\left(\nabla f_{i_t}(x^k_t) - \nabla f'_{i_t}(y^k_t)\right)\right|\right| \\ & + \left|\left|\tilde{x}_k - \tilde{y}_k - \eta\left(\nabla f_{i_t}(\tilde{x}_k) - \nabla f'_{i_t}(\tilde{y}_k)\right)\right|\right| \\ & + ||\tilde{x}_k - \tilde{y}_k - \eta\left(\nabla F_S(\tilde{x}_k) - \nabla F_{S'}(\tilde{y}_k)\right)||\,. \end{aligned}$$

Por la fuerte convexidad de las pérdidas, para $\eta \leq \frac{1}{L}$ se cumple la $(1 - \eta\mu)$-expansividad de una actualización de tipo gradiente. Aplicando esta propiedad a los términos anteriores, se tiene

$$\delta^k_{t+1} \leq (1 - \eta\mu)\delta^k_t + 2\eta M \boldsymbol{r}_{\boldsymbol{i}_{k,t}} + 2(1 - \eta\mu)\delta^k_1 + 2\eta M \boldsymbol{r}_{\boldsymbol{i}_{k,t}} + \frac{2\eta M}{n}.$$

Luego, en valor esperado sobre la aleatoriedad del algoritmo se obtiene

$$\mathbb{E}\delta^k_{t+1} \leq (1 - \eta\mu)\mathbb{E}\delta^k_t + 2(1 - \eta\mu)\mathbb{E}\delta^k_1 + \frac{6\eta M}{n}.$$

Sumar las desigualdades sobre todo el ciclo interior del algoritmo entrega, para $m \geq 1$*,*

$$\begin{aligned}
\mathbb{E}\delta_{m+1}^k &\leq (1-\eta\mu)^m \mathbb{E}\delta_1^k + (1-\eta\mu)^{m-1}\frac{2\eta M}{n} \\
&\quad + \left[2(1-\eta\mu)\mathbb{E}\delta_1^k + \frac{6\eta M}{n}\right]\sum_{i=0}^{m-2}(1-\eta\mu)^i \\
&= \left[\frac{2(1-\eta\mu)}{\eta\mu} - \frac{2-\eta\mu}{\eta\mu}(1-\eta\mu)^m\right]\mathbb{E}\delta_1^k + \frac{2\eta M}{n}(1-\eta\mu)^{m-1} \\
&\quad + \frac{6\eta M}{n}\sum_{i=0}^{m-2}(1-\eta\mu)^i.
\end{aligned}$$

Como $m = \Theta(n)$*, para valores crecientes de* n *el factor que acompaña a la esperanza del lado derecho converge monótonamente creciente a la cantidad* $2(1-\eta\mu)/(\eta\mu)$*. Se separan dos casos:*

I) Sea $\eta \geq \frac{2}{3\mu}$*, luego el factor anterior es menor a uno. Con esta nueva restricción, se calcula la estabilidad de la ejecución completa:*

$$\mathbb{E}\delta_1^{k+1} \leq \left[\frac{2(1-\eta\mu)}{\eta\mu} - \frac{2-\eta\mu}{\eta\mu}(1-\eta\mu)^m\right]\mathbb{E}\delta_1^k + \frac{6\eta M}{n}\sum_{i=0}^{m-1}(1-\eta\mu)^i.$$

Notando que la sumatoria del lado derecho converge, se acota por el valor al límite. Finalmente, sumando en el ciclo exterior se obtiene una cota superior para el valor esperado de las distancias:

$$\begin{aligned}
&\leq \left[\frac{2(1-\eta\mu)}{\eta\mu} - \frac{2-\eta\mu}{\eta\mu}(1-\eta\mu)^m\right]\mathbb{E}\delta_1^k + \frac{6M}{n\mu} \\
\Rightarrow \mathbb{E}\delta_1^{K+1} &\leq \frac{6M}{n\mu}\sum_{i=0}^{K-1}\left[\frac{2(1-\eta\mu)}{\eta\mu} - \frac{2-\eta\mu}{\eta\mu}(1-\eta\mu)^m\right]^i \\
&\leq \frac{6M}{n\mu}\cdot\frac{\eta\mu}{3\eta\mu - 2 + (2-\eta\mu)(1-\eta\mu)^m},
\end{aligned}$$

deduciéndose una uniforme estabilidad de argumentos $O\left(\frac{\eta M}{n}\right)$*, debido al supuesto sobre* η*.*

II) Alternativamente, sea $\eta < \frac{2}{3\mu}$*. Por lo tanto, el factor que acompaña a la esperanza es mayor a uno, un cálculo directo muestra que esto implica crecimiento exponencial de la estabilidad. Se calcula la distancia de las trayectorias para la ejecución completa*

$$\mathbb{E}\delta_1^{k+1} \leq \left[\frac{2(1-\eta\mu)}{\eta\mu} - \frac{2-\eta\mu}{\eta\mu}(1-\eta\mu)^m\right]\mathbb{E}\delta_1^k + \frac{6M}{n\mu}$$
$$\Rightarrow \mathbb{E}\delta_1^{K+1} \leq \frac{6M}{n\mu}\sum_{k=0}^{K-1}\left[\frac{2(1-\eta\mu)}{\eta\mu} - \frac{2-\eta\mu}{\eta\mu}(1-\eta\mu)^m\right]^k,$$

obteniéndose una cota de orden $O\left(\frac{M}{\mu n}\cdot\left(\frac{1}{\eta\mu}\right)^K\right)$ *para la estabilidad de argumentos.* □

Dado que los órdenes de estabilidad son iguales a los planteados en la Proposición 4.1, se concluye que el efecto de la elección estocástica no es una mejora significativa en el análisis de estabilidad. En este caso se obtuvieron resultados del mismo orden de estabilidad para valores pequeños del *learning rate*, empeorándose el análisis en los casos donde se toma un paso más grande. Se obtiene una división similar al caso estocástico debido a la existencia de dos regímenes distintos, donde tomar un η mayor mejora el análisis, pero la elección determinística se comporta mejor en el caso límite. Sin embargo, la elección estocástica es necesaria para obtener convergencia del error de optimización, por lo que se limita el análisis del exceso de riesgo al uso del Algoritmo 8.

ANEXO B. RESULTADOS COMPLEMENTARIOS PARA PEP BATCH

B.1. Cálculo del dual SDP *batch*

En esta sección, se calcula el problema dual a la versión semidefinida del problema (5.14) presentado en la Sección 5.1. El problema (5.14) modela el peor caso de una métrica $\mathcal{P}$ bSOLG-representable por la expresión $b^\top v + \text{Tr}(CX)$, que actúa sobre pérdidas pertenecientes a las clases generalizadas $\mathcal{F}_{M,L}$ o $\mathcal{F}_M$, realizando actualizaciones de algún método *batch* $\mathcal{M}$ expuesto en el Capítulo 2 y es aplicado al problema de minimización de riesgo empírico (1.3) de acuerdo a la reformulación semidefinida expuesta en el Capítulo 5.

Proposición B.1. *El dual al problema* (5.14) *es el problema semidefinido*

$$\begin{aligned}
&\underset{\lambda_{ij},\mu_i,\tau,\kappa}{\textit{inf}} && \tau R^2 + \sum_{i\in I_M} \mu_i M^2 + \kappa \\
&\textit{s.a} && \sum_{(i,j)\in\mathcal{I}} \lambda_{ij} A_{ij} + \sum_{i\in I_M} \mu_i A_{M_i} + \tau A_R + \kappa A_y - C \succeq O \\
& && \sum_{(i,j)\in\mathcal{I}} \lambda_{ij}(e_j - e_i) = b \\
& && \lambda_{ij}, \mu_i \geq 0 \quad ((i,j)\in\mathcal{I}) \\
& && \tau, \kappa \geq 0.
\end{aligned}$$

DEMOSTRACIÓN. *En primer lugar, se reescribe el problema* (5.14) *mediante el uso de variables de holgura y se lleva a una forma estándar para optimización semidefinida,*

$$\begin{aligned}
&\underset{v\in\mathbb{R}^{4T+10}, X\in\mathbb{S}^{4T+11}}{\textit{sup}} && b^\top v + \textit{Tr}(CX) \\
&\textit{s.a} && \textit{Tr}(A_{ij}X) + v_j - v_i + s_{ij} = 0 && (\forall i,j\in\mathcal{I}) \\
& && \textit{Tr}(A_{M_i}X) + s_i = M^2 && (\forall i\in I_M) \\
& && \textit{Tr}(A_R X) + s_R = R^2 \\
& && \textit{Tr}(A_y X) + s_y = 1.
\end{aligned} \tag{B.1}$$

Se utilzan las partes positiva y negativa de v separadamente, de la forma $v = v^+ + v^-$. Además, se considera la matriz diagonal por bloques de variables

$$\tilde{X} = \begin{pmatrix} diag(s) & O & O & O \\ O & diag(v^+) & O & O \\ O & O & diag(v^-) & O \\ O & O & O & X \end{pmatrix}$$

y el vector de holguras $s = \{s_x\}_{x \in \mathcal{I} \cup I_M \cup \{R,y\}}$. Luego, se pueden definir las matrices paramétricas diagonales por bloques

$$\tilde{A_{ij}} = \begin{pmatrix} diag(e_{ij}) & O & O & O \\ O & diag(e_j - e_i) & O & O \\ O & O & diag(e_i - e_j) & O \\ O & O & O & A_{ij} \end{pmatrix} \qquad (\forall (i,j) \in \mathcal{I})$$

$$\tilde{A_{M_i}} = \begin{pmatrix} diag(e_i) & O & O & O \\ O & O & O & O \\ O & O & O & O \\ O & O & O & A_{M_i} \end{pmatrix} \qquad (\forall i \in I_M)$$

$$\tilde{A_R} = \begin{pmatrix} diag(e_R) & O & O & O \\ O & O & O & O \\ O & O & O & O \\ O & O & O & A_R \end{pmatrix}$$

$$\tilde{A_y} = \begin{pmatrix} diag(e_y) & O & O & O \\ O & O & O & O \\ O & O & O & O \\ O & O & O & A_y \end{pmatrix}$$

$$\tilde{C} = \begin{pmatrix} O & O & O & O \\ O & diag(b) & O & O \\ O & O & diag(-b) & O \\ O & O & O & C \end{pmatrix}$$

quedando una versión del modelo primal anterior expresada solo mediante restricciones de igualdad:

$$\begin{aligned} \underset{s,v^+,v^- \geq 0; X \in \mathbb{S}^{4T+11}}{sup} \quad & Tr(\tilde{C}\tilde{X}) \\ s.a \quad & Tr(\tilde{A_{ij}}\tilde{X}) = 0 \qquad (\forall i,j \in \mathcal{I}) \\ & Tr(\tilde{A_{M_i}}\tilde{X}) = M^2 \qquad (\forall i \in I_M) \\ & Tr(\tilde{A_R}\tilde{X}) = R^2 \\ & Tr(\tilde{A_y}\tilde{X}) = 1. \end{aligned} \tag{B.2}$$

Esto permite plantear el modelo dual; considerando las variables duales κ, τ, μ_i y λ_{ij} asociadas respectivamente a las matrices por bloques $\tilde{A_y}, \tilde{A_R}, \tilde{A_{M_i}}$ y $\tilde{A_{ij}}$:

$$\begin{aligned} \underset{\{\lambda_{ij}\}_{(i,j)\in\mathcal{I}},\{\mu_i\}_{i\in I_M},\tau,\kappa}{\textit{inf}} \quad & \tau R^2 + \sum_{i\in I_M} \mu_i M^2 + \kappa \\ \textit{s.a} \quad & \sum_{(i,j)\in\mathcal{I}} \lambda_{ij}\tilde{A_{ij}} + \sum_{i\in I_M} \mu_i \tilde{A_{M_i}} + \kappa\tilde{A}_y + \tau\tilde{A}_R - \tilde{C} \succeq O. \end{aligned} \tag{B.3}$$

Finalmente, se nota que la restricción semidefinida presenta sumas ponderadas de matrices diagonales por bloques, donde los tres primeros bloques diagonales corresponden a matrices diagonales, lo que equivale a tres restricciones separadas. La primera corresponde a la condición

$$\sum_{(i,j)\in\mathcal{I}} \lambda_{ij} s_{ij} + \sum_{i\in I_M} \mu_i s_i + \kappa s_y + \tau s_R \geq O,$$

que dada la no-negatividad de las holguras, es equivalente a que las variables duales tomen valores no-negativos. La segunda y tercera son equivalentes al par de desigualdades vectoriales

$$\begin{cases} \sum_{(i,j)\in\mathcal{I}} \lambda_{ij}(e_j - e_i) - b \geq 0 \\ \sum_{(i,j)\in\mathcal{I}} \lambda_{ij}(e_i - e_j) + b \geq 0, \end{cases}$$

por lo que se concluye la igualdad a cero. Por lo tanto, el problema dual puede reescribirse de la forma expresada en (5.15)*:*

$$\begin{aligned} \underset{\lambda_{ij},\mu_i,\tau,\kappa}{\textit{inf}} \quad & \tau R^2 + \sum_{i\in I_M} \mu_i M^2 + \kappa \\ \textit{s.a} \quad & \sum_{(i,j)\in\mathcal{I}} \lambda_{ij} A_{ij} + \sum_{i\in I_M} \mu_i A_{M_i} + \tau A_R + \kappa A_y - C \succeq O \\ & \sum_{(i,j)\in\mathcal{I}} \lambda_{ij}(e_j - e_i) = b \\ & \lambda_{ij}, \mu_i \geq 0 \quad ((i,j)\in\mathcal{I}) \\ & \tau, \kappa \geq 0, \end{aligned}$$

llegando al resultado enunciado. □

Se concluye que el dual a la versión semidefinida del PEP para métodos *Batch* es el problema obtenido, lo que permite su utilización para el estudio de la dualidad Lagrangiana. Además, el problema dual recién derivado se implementará computacionalmente, obteniéndose información complementaria a la obtenida de la implementación computacional del modelo primal, como se presenta en la Sección 5.2 para el método proximal.

B.2. Dualidad fuerte para SDP *batch*

En esta sección se muestran las pruebas de las garantías de dualidad fuerte para problemas *Batch* que no fueron expuestas en el Capítulo 5. Estas demostraciones garantizan dualidad fuerte para los métodos de subgradiente y de punto proximal cuando se desea maximizar una métrica de rendimiento que comprende los fenómenos de optimización (mediante el *gap* de optimalidad) y de estabilidad (mediante la diferencia en norma de las trayectorias) conjuntos. Se plantean nuevamente las Proposiciones 5.3 y 5.4 junto a sus pruebas respectivas y a una descripción de como extender los resultados al análisis de errores de estabilidad y optimización por separado.

Proposición B.2. *Sea* (5.14) *bSOLG-representación del problema de peor caso conjunto de optimización y estabilidad de una ejecución de T pasos del método de subgradiente. Sea* (5.15) *su dual semidefinido, donde ambos problemas son factibles. Entonces, ambos poseen un valor óptimo idéntico y existe un punto primal-factible que alcanza tal valor.*

DEMOSTRACIÓN. *Similar al caso expuesto para el método de gradiente en la demostración de la Proposición 5.2, se demuestra la dualidad fuerte mediante un certificado para la condición de Slater en el problema dual. Se nota que la matriz*

$$S = \tau A_R + \kappa A_y + \sum_{i \in I_M} \mu_i A_{M_i}$$

es una matriz diagonal cuyas columnas son la base canónica en $\mathbb{R}^{4T+10}$*. Se buscan valores apropiados de las variables duales de manera que la condición semidefinida del problema* (5.15) *se cumpla de manera estricta y también se cumpla la igualdad vectorial. En pos de esto, se calculan los valores propios de* C_{avGD} (5.12)*, notando que los máximos valores propios en valor absoluto son*

$$\pm \frac{M\eta}{2n}\sqrt{\frac{(T+1)(2T+1)(n^2-2n+2)}{3T}}.$$

Además, se acota la norma de las matrices asociadas a la interpolación convexa para el caso no-suave. Para $(i,j) \in \mathcal{I}$*:*

$$\begin{aligned} ||A_{ij}|| &= \sup_{||z||=1} z^\top A_{ij} z \\ &= \sup_{||z||=1} [\langle u_j, z\rangle \langle h_i - h_j, z\rangle] \\ &\leq ||h_i - h_j||. \end{aligned}$$

Luego, fijando los valores

$$\varepsilon_{ij} = \begin{cases} \frac{1}{|\mathcal{I}|}||h_i - h_j||^{-1} & (i \neq j) \\ \frac{1}{|\mathcal{I}|} & (i = j), \end{cases}$$

$$\lambda_{ij} = \begin{cases} \frac{1}{n} + \varepsilon_{ij} & (i = 4T+9, j = 4T+5) \\ \frac{n-1}{n} + \varepsilon_{ij} & (i = 4T+10, j = 4T+6) \\ \varepsilon_{ij} & \textit{e.o.c.} \end{cases}$$

se cumple la restricción de igualdad vectorial a b_{av} (5.9)*. Resta demostrar el cumplimento estricto de la restricción semidefinida, por lo que se acota superiormente la norma, utilizando desigualdad triangular y notando las desigualdades* $||A_{4T+9,4T+5}||, ||A_{4T+10,4T+6}|| \leq ||h_{4T+5}||$ *se tiene*

$$\begin{aligned}\left|\left|\sum_{(i,j)\in\mathcal{I}} \lambda_{ij}A_{ij} - C_{avGD}\right|\right| &\leq ||h_{4T+5}|| + \sum_{(i,j)\in\mathcal{I}, i\neq j} \varepsilon_{ij}||A_{ij}|| + ||C_{avGD}|| \\ &< ||h_{4T+5}|| + 1 + \frac{M\eta}{2n}\sqrt{\frac{(T+1)(2T+1)(n^2-2n+2)}{3T}} \\ &= \sqrt{\frac{2\eta^2(T+1)(2T+1)(n^2-2n+2)}{3Tn^2}} + 1 + 1 \\ &\quad + \frac{M\eta}{2n}\sqrt{\frac{(T+1)(2T+1)(n^2-2n+2)}{3T}},\end{aligned}$$

donde en la desigualdad estricta se utiliza el cálculo del valor propio máximo de C_{avGD} *y los valores fijados para las variables lambda. Finalmente, se considera* B *como la cota superior estricta antes calculada, que corresponde a una función solo de los parámetros del problema. Basta fijar el resto de las variables duales con valor* B *para concluir la prueba con la desigualdad estricta*

$$S + \sum_{(i,j)\in\mathcal{I}} \lambda_{ij}A_{ij} - C_{avGD} \succ O.$$

□

Proposición B.3. *Sea* (5.14) *bSOLG-representación del problema de peor caso conjunto de optimización y estabilidad de una ejecución de* T *pasos del método de punto proximal. Sea* (5.15) *su dual semidefinido, donde ambos son problemas factibles. Entonces, ambos poseen un valor óptimo idéntico y existe un punto primal-factible que alcanza tal valor.*

DEMOSTRACIÓN. *Similar a ambos casos anteriores, se demuestra la dualidad fuerte para el método proximal mediante un certificado para la condición de Slater en el problema dual. Se nota que la matriz*

$$S = \tau A_R + \kappa A_y + \sum_{i \in I_M} \mu_i A_{M_i}$$

es una matriz diagonal, igual que para las representaciones previas. Se buscan valores apropiados de las variables duales de manera que la condición semidefinida del problema (5.15) *se cumpla de manera estricta y adicionalmente se cumpla la igualdad vectorial. En pos de esto, se obtienen los valores propios de* C_{PPM} (5.13)*, notando que un cálculo rutinario entrega los máximos valores propios (en valor absoluto)*

$$\pm \frac{M\eta}{2n}\sqrt{2T(n^2 - 2n + 2)}.$$

Además, se acota la norma de las matrices asociadas a la interpolación convexa. Sea $(i, j) \in \mathcal{I}$*:*

$$||A_{ij}|| = \sup_{||z||=1} [\langle u_j, z\rangle \langle h_i - h_j, z\rangle] \leq ||h_i - h_j||.$$

Luego, fijando los valores

$$\varepsilon_{ij} = \begin{cases} \frac{1}{|\mathcal{I}|}||h_i - h_j||^{-1} & (i \neq j) \\ \frac{1}{|\mathcal{I}|} & (i = j), \end{cases}$$

$$\lambda_{ij} = \begin{cases} \frac{1}{n} + \varepsilon_{ij} & (i = 4T + 9, j = T + 1) \\ \frac{n-1}{n} + \varepsilon_{ij} & (i = 4T + 10, j = 2T + 2) \\ \varepsilon_{ij} & \textit{e.o.c.} \end{cases}$$

se cumple la restricción de igualdad vectorial a b_{last}*, definido en* (5.8)*. Resta mostrar el cumplimento estricto de la restricción semidefinida, por lo que se acota superiormente la norma siguiente, utilizando desigualdad triangular y notando los hechos* $||A_{4T+9,T+1}||, ||A_{4T+10,2T+2}|| \leq ||h_{T+1}||$ *se tiene*

$$\begin{aligned}\left|\left|\sum_{(i,j)\in\mathcal{I}} \lambda_{ij}A_{ij} - C_{PPM}\right|\right| &\leq ||h_{T+1}|| + \sum_{(i,j)\in\mathcal{I}, i\neq j} \varepsilon_{ij}||A_{ij}|| + ||C_{PPM}|| \\ &< ||h_{T+1}|| + 1 + \frac{M\eta}{2n}\sqrt{2T(n^2-2n+2)} \\ &= \underbrace{\sqrt{2\eta^2 T\frac{n^2-2n+2}{n^2}+1} + 1 + \frac{M\eta}{2n}\sqrt{2T(n^2-2n+2)}}_{B},\end{aligned}$$

donde en la desigualdad estricta utiliza el cálculo del valor propio máximo de C_{PPM} *y los valores fijados para las variables lambda. Finalmente, se considera* B *como la cota superior estricta a la norma, que corresponde a una función solo de los parámetros del problema. Basta fijar el resto de las variables duales con valor* B *para concluir la prueba con la desigualdad estricta*

$$S + \sum_{(i,j)\in\mathcal{I}} \lambda_{ij} - C_{PPM} \succ O.$$

□

COMENTARIO B.1. *Las Proposiciones 5.3 y 5.4 también pueden extenderse fácilmente a los casos de tomar una métrica de solo gap de optimalidad o de solo estabilidad, siguiendo la idea original. Se considera la matriz de parámetros* $C = O$ *en la función objetivo para el caso solo optimización y* $b = 0$ *para solo estabilidad. Ambos casos implican que* B *es una cota estricta válida, pero menos ajustada, al igual que en el caso del método de gradiente.*

Por lo tanto, las representaciones semidefinidas para los métodos de subgradiente y de punto proximal, ambos con su respectiva selección de gradientes de X y parámetros

de interpolación, cumplen la dualidad fuerte para el caso no-suave. Además, el Comentario B.1 entrega las extensiones a los problemas aislados de estabilidad y de optimización, mediante el uso de solo uno de los términos de la métrica compuesta, de la misma manera que fue expuesto en la Sección 5.1.

B.3. Resultados complementarios a la aplicación de la metodología PEP

Se presenta un resultado que complementa lo expuesto en la Subsección 5.2.3 para los problemas de maximizar el error de optimización y el error de estabilidad de una ejecución del método de punto proximal después de $T = n$ iteraciones, por sí mismos.

Para el primero, se comprueba la factibilidad dual de una asignación obtenida utilizando la metodología PEP para el modelo semidefinido asociado al problema de minimización ya mencionado. Esto se formaliza en la Proposición B.4, demostrada posteriormente.

Proposición B.4. *El punto dado por las asignaciones de variables* (5.16) *es factible para el problema* (5.15) *de maximizar el gap de optimalidad para el PPM.*

DEMOSTRACIÓN. *Se demuestra el cumplimiento de ambas restricciones duales, basándose en la demostración de Taylor et al. (2017a) para el Lema 2.11. Primero, se nota que la igualdad vectorial se cumple, teniéndose luego de un cálculo rutinario el cumplimiento de la igualdad para toda coordenada no trivial:*

$$
\begin{aligned}
&\lambda_{4n+9,2} - \lambda_{2,3} && = 0 && \\
&\lambda_{4n+10,n+3} - \lambda_{n+3,n+4} && = 0 && \\
&\lambda_{4n+9,i+1} + \lambda_{i,i+1} - \lambda_{i+1,i+2} && = 0 && (\forall i \in [2, n-1]) \\
&\lambda_{4n+10,n+2+i} + \lambda_{n+i+1,n+2+i} - \lambda_{n+2+i,n+3+i} && = 0 && (\forall i \in [2, n-1]) \\
&\lambda_{4n+9,n+1} + \lambda_{n,n+1} && = \frac{1}{n} && \\
&\lambda_{4n+10,2n+2} + \lambda_{2n+1,2n+2} && = \frac{n-1}{n}. &&
\end{aligned}
$$

Luego, resta mostrar el cumplimiento de la restricción semidefinida. Expresar las matrices asociadas a la interpolación convexa no-suave en términos de vectores canónicos entrega las igualdades

$$\begin{aligned}
A_{i+1,i+2} &= \frac{1}{2}u_{i+2}(h_{i+1}-h_{i+2})^\top + \frac{1}{2}(h_{i+1}-h_{i+2})u_{i+2}^\top \\
&= \frac{\eta}{n}e_{i+2}e_{i+2}^\top + \frac{\eta(n-1)}{2n}e_{i+2}e_{n+3+i}^\top + \frac{\eta(n-1)}{2n}e_{n+3+i}e_{i+2}^\top \\
A_{n+2+i,n+3+i} &= \frac{1}{2}u_{n+3+i}(h_{i+1}-h_{i+2})^\top + \frac{1}{2}(h_{i+1}-h_{i+2})u_{n+3+i}^\top \\
&= \frac{\eta(n-1)}{n}e_{n+3+i}e_{n+3+i}^\top + \frac{\eta}{2n}e_{i+2}e_{n+3+i}^\top + \frac{\eta}{2n}e_{n+3+i}e_{i+2}^\top \\
A_{4n+9,i+1} &= \frac{1}{2}u_{i+1}(-h_{i+1})^\top + \frac{1}{2}(-h_{i+1})u_{i+1}^\top \\
&= \frac{1}{2n}e_{i+1}\left(\eta\sum_{j=1}^{i}(e_{j+1}+(n-1)e_{n+j+2})\right)^\top \\
&\quad + \frac{1}{2n}\left(\eta\sum_{j=1}^{i}(e_{j+1}+(n-1)e_{n+j+2})\right)e_{i+1}^\top - \frac{1}{2}e_{i+1}e_{4n+10}^\top - \frac{1}{2}e_{4n+10}e_{i+1}^\top \\
A_{4n+10,n+i+2} &= \frac{1}{2}u_{n+i+2}(-h_{i+1})^\top + \frac{1}{2}(-h_{i+1})u_{n+i+2}^\top \\
&= \frac{1}{2n}e_{n+i+2}\left(\eta\sum_{j=1}^{i}(e_{j+1}+(n-1)e_{n+j+2})\right)^\top - \frac{n-1}{2}e_{n+i+2}e_{4n+10}^\top \\
&\quad + \frac{1}{2n}\left(\eta\sum_{j=1}^{i}(e_{j+1}+(n-1)e_{n+j+2})\right)e_{n+i+2}^\top - \frac{n-1}{2}e_{4n+10}e_{n+i+2}^\top.
\end{aligned}$$

Utilizando la notación simplificada $\lambda_i := \lambda_{i,i+1}$, $\nu_i := \lambda_{4n+10,i}$ y notando que las asignaciones a las variables duales referentes a restricciones de F son múltiplos de las asignaciones relativas a f, se puede reescribir la restricción semidefinida por bloques. Primero,

se define la matriz

$$S = \frac{\eta}{2n}\begin{pmatrix} 2\nu_2 & \nu_3 & \nu_4 & \dots & \nu_{n+1} \\ \nu_3 & 2\nu_3+2\lambda_2 & \nu_4 & \dots & \nu_{n+1} \\ \nu_4 & \nu_4 & 2\nu_4+2\lambda_3 & \dots & \nu_{n+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \nu_{n+1} & \nu_{n+1} & \nu_{n+1} & \dots & 2\nu_{n+1}+2\lambda_n \end{pmatrix},$$

que puede reescribirse como sigue, de acuerdo con las igualdades $\nu_2 = \lambda_2$, $\nu_i + \lambda_{i-1} = \lambda_i$ *para todo* $i \in [3,n]$ *y* $\nu_{n+1} + \lambda_n = \frac{1}{n}$*:*

$$= \frac{\eta}{2n}\begin{pmatrix} 2\lambda_2 & \nu_3 & \nu_4 & \dots & \nu_{n+1} \\ \nu_3 & 2\lambda_3 & \nu_4 & \dots & \nu_{n+1} \\ \nu_4 & \nu_4 & 2\lambda_4 & \dots & \nu_{n+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \nu_{n+1} & \nu_{n+1} & \nu_{n+1} & \dots & 2/n \end{pmatrix}.$$

En segundo lugar, se define el vector

$$\mathbf{v} = -\frac{1}{2}(\nu_{i+1})_{i\in[n]}.$$

Luego, eliminando las filas nulas de la restricción semidefinida, se tiene la restricción equivalente en términos de S, $\mathbf{v}$ *y* τ*:*

$$\begin{pmatrix} S & (n-1)S & \mathbf{v} \\ (n-1)S & (n-1)^2 S & (n-1)\mathbf{v} \\ \mathbf{v}^\top & (n-1)\mathbf{v}^\top & \tau \end{pmatrix} \succeq O.$$

Utilizando el complemento de Schur, se obtiene la equivalencia con la restricción:

$$\begin{pmatrix} S & (n-1)S \\ (n-1)S & (n-1)^2 S \end{pmatrix} - \frac{1}{\tau}\begin{pmatrix} \mathbf{v} \\ (n-1)\mathbf{v} \end{pmatrix}\begin{pmatrix} \mathbf{v} \\ (n-1)\mathbf{v} \end{pmatrix}^\top \succeq O,$$

que un cálculo sencillo muestra es equivalente a la restricción semidefinida

$$S - \frac{1}{\tau}\mathbf{v}\mathbf{v}^\top \succeq O.$$

Se demuestra que la resta del lado izquierdo es una matriz diagonal dominante y cuya diagonal es positiva, condiciones suficientes para concluir lo enunciado. Primero, se muestra que los términos fuera de la diagonal son no-positivos. Sean $1 \leq j < i \leq n$*:*

$$\frac{\eta\nu_{i+1}}{2n} - \frac{1}{4\tau}\nu_{i+1}\nu_{j+1} = \frac{\eta\nu_{i+1}}{2n}\left(1 - 2n^2\nu_{j+1}\right) = \frac{\eta\nu_{i+1}}{2n}\left(1 - \frac{2n}{2n-j}\frac{2n}{2n+1-j}\right) < 0.$$

Luego, se muestra que la matriz es diagonal dominante, pudiendo utilizarse la suma directa de una fila. Para esto, se separan dos casos exhaustivos, En primer lugar, sea $i = n$*:*

$$\begin{aligned}\sum_{j\neq i}\left(S - \frac{1}{\tau}\mathbf{v}\mathbf{v}^\top\right)_{ij} &= \frac{\eta}{2n}\sum_{j=1}^{n-1}\nu_{n+1} - \frac{1}{4\tau}\nu_{n+1}\sum_{j=1}^{n-1}\nu_{j+1} \\ &= \frac{1}{2n}\left((n-1)\eta\nu_{n+1} - \frac{n\nu_{n+1}}{2\tau}\sum_{j=1}^{n-1}\nu_{j+1}\right).\end{aligned}$$

Utilizando la igualdad $\lambda_n + \nu_{n+1} = \frac{1}{n}$ *y la consecuencia directa de la restricción vectorial* $\sum_{j=2}^{n}\nu_j = \lambda_n$*:*

$$\begin{aligned}&= \frac{1}{2n}\left((n-1)\eta\nu_{n+1} - \frac{n\nu_{n+1}}{2\tau}\lambda_n\right) \\ &= \frac{1}{2n}\left((n-1)\eta\nu_{n+1} - \frac{\nu_{n+1}}{2\tau}(1 - n\nu_{n+1})\right) \\ &= \frac{\nu_{n+1}^2}{4\tau} + \frac{(n-1)\eta\nu_{n+1}}{2n} - \frac{\nu_{n+1}}{4n\tau}.\end{aligned}$$

Finalmente, reemplazando ν_{n+1} *y* τ*:*

$$= \frac{\nu_n^2}{4\tau} - \frac{\eta}{n^2},$$

que corresponde al inverso aditivo de la última entrada diagonal. Luego, para el caso contrario, $i < n$, la suma de los términos no diagonales es

$$\sum_{j\in[n],i\neq j}\left(S-\frac{1}{\tau}\mathbf{v}\mathbf{v}^\top\right)_{ij} = \frac{\eta}{2n}\sum_{j=1}^{i-1}\nu_{i+1}+\frac{\eta}{2n}\sum_{j=i+1}^{n}\nu_{j+1}-\frac{1}{4\tau}\nu_{i+1}\sum_{j\neq i}\nu_{j+1}$$
$$= \frac{\eta}{2n}\sum_{j=1}^{i-1}\nu_{i+1}+\frac{\eta}{2n}\sum_{j=i+1}^{n}\nu_{j+1}-\frac{1}{4\tau}\nu_{i+1}\left(\sum_{j=1}^{i-1}\nu_{j+1}+\sum_{j=i+1}^{n}\nu_{j+1}\right).$$

Utilizando las siguientes igualdades derivadas de la igualdad vectorial: $\sum_{j=2}^{i}\nu_j = \lambda_i$, $\sum_{j=i}^{n}\nu_{j+1} = \frac{1}{n}-\lambda_i$ *y* $\lambda_i-\lambda_{i-1}=\nu_i$

$$= \frac{\eta}{2n}(i-1)\nu_{i+1}+\frac{\eta}{2n}\left(\frac{1}{n}-\lambda_{i+1}\right)-\frac{1}{4\tau}\nu_{i+1}\left(\frac{1}{n}-\nu_{i+1}\right)$$
$$= \frac{1}{4\tau}\nu_{i+1}^2+\nu_{i+1}\left(\frac{\eta(i-1)}{2n}-\frac{1}{4n\tau}\right)+\frac{\eta}{2n}\left(\frac{1}{n}-\lambda_{i+1}\right).$$

Finalmente, por las definiciones de λ_i, ν_i y τ se cumplen las igualdades

$$= \frac{1}{4\tau}\nu_{i+1}^2+\frac{-2n\eta}{2n^2(2n-i)}+\frac{2\eta(n-i)}{2n^2(2n-i)}$$
$$= \frac{1}{4\tau}\nu_{i+1}^2-\frac{\eta}{n}\underbrace{\frac{i}{n(2n-i)}}_{\lambda_{i+1}}.$$

Se concluye que en ambos casos la suma de los valores de cada fila es igual a cero, donde es posible descartar los valores absolutos para la condición de diagonal-dominancia por la negatividad de los valores que no pertenecen a la diagonal y positividad de la diagonal. Por ende, la matriz es semidefinida positiva y el punto generado por las restricciones de (5.16) *es dual factible.* □

En cambio, para el problema de estabilidad se conjetura un resultado obtenible a través de la metodología PEP, mostrando una demostración parcial de tal hipótesis junto a una

discusión de la evidencia que apoya esta afirmación. Se plantea nuevamente la conjetura expuesta en el Capítulo 5:

Conjetura B.1. *El punto dado por las restricciones* (5.17) *es dual factible para el problema de peor caso de estabilidad luego de* $T = n$ *iteraciones del método de punto proximal.*

COMENTARIO B.2. *Siguiendo lo expuesto en la Proposición anterior para el caso de optimización, se quiere demostrar que la elección de valores para las variables cumplen con ambas restricciones del problema dual. Primero, se nota que la elección de las variables lambda es simétrica, concluyéndose que se cumple la igualdad*

$$\sum_{j:(i,j)\in\mathcal{I}} \lambda_{ji} - \lambda_{ij} = 0$$

para toda elección de $i \in I_M$.

Luego, resta la prueba para la restricción semidefinida. Nuevamente, se quiere explotar la estructura de la matriz que se quiere probar semidefinida positiva. Se buscan expresiones para las matrices asociadas a cada variable activa lambda en términos de vectores canónicos:

$$\begin{aligned} A_{i+n+2,i+3n+4} &= \frac{1}{2}u_{i+3n+4}(h_{i+n+2} - h_{i+3n+4})^\top + \frac{1}{2}(h_{i+n+2} - h_{i+3n+4})u_{i+3n+4}^\top \\ &= \frac{1}{2}u_{i+3n+4}(h_{i+1} - h_{i+2n+3})^\top + \frac{1}{2}(h_{i+1} - h_{i+2n+2+3})u_{i+3n+4}^\top \\ A_{i+3n+4,i+n+2} &= \frac{1}{2}u_{i+n+2}(h_{i+2n+3} - h_{i+1})^\top + \frac{1}{2}(h_{i+1} - h_{i+2n+3})u_{i+n+2}^\top \end{aligned}$$

$$
\begin{aligned}
&\Rightarrow A_{i+n+2,i+3n+4} + A_{i+3n+4,i+n+2} \\
&= -\frac{1}{2}(u_{i+n+2} - u_{i+3n+4})(h_{i+2n+3} - h_{i+1})^\top - \frac{1}{2}(h_{i+1} - h_{i+2n+3})(u_{i+n+2} - u_{i+3n+4})^\top \\
&= \frac{\eta}{2n}\left(e_{i+n+2} - e_{i+3n+4}\right)\left(\sum_{j=1}^{i} e_{j+1} + (n-1)\sum_{j=1}^{i} e_{j+n+2} - \sum_{j=1}^{i} e_{j+2n+3} - (n-1)\sum_{j=1}^{i} e_{j+3n+4}\right)^\top \\
&\quad + \frac{\eta}{2n}\left(\sum_{j=1}^{i} e_{j+1} + (n-1)\sum_{j=1}^{i} e_{j+n+2} - \sum_{j=1}^{i} e_{j+2n+3} - (n-1)\sum_{j=1}^{i} e_{j+3n+4}\right)\left(e_{i+n+2} - e_{i+3n+4}\right)^\top .
\end{aligned}
$$

La restricción semidefinida puede reescribirse como una matriz por bloques, reduciendo la dimensión de ésta tras eliminar filas nulas. Se define la matriz L de $n \times n$ para mostrar esta representación equivalente, utilizando la simplificación de la notación para las variables lambda de la forma $\lambda_i := \lambda_{i+n+2,i+3n+4}$,

$$
L = \begin{pmatrix} \lambda_1 & 0 & 0 & \dots & 0 \\ \lambda_2 & \lambda_2 & 0 & \dots & 0 \\ \lambda_3 & \lambda_3 & \lambda_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \lambda_n & \lambda_n & \lambda_n & \dots & \lambda_n \end{pmatrix}.
$$

Luego, la restricción semidefinida puede reescribirse como

$$
\frac{\eta}{2n}\begin{pmatrix} I & L^\top & O & -L^\top & -M\mathbf{1} \\ L & (n-1)(L+L^\top) & -L & (1-n)(L+L^\top) & (1-n)M\mathbf{1} \\ O & -L^\top & I & L^\top & M\mathbf{1} \\ -L & (1-n)(L+L^\top) & L & (n-1)(L+L\top) & (n-1)M\mathbf{1} \\ -M\mathbf{1}^\top & (1-n)M\mathbf{1}^\top & M\mathbf{1}^\top & (n-1)M\mathbf{1}^\top & 2nM^2 \end{pmatrix} \succeq O.
$$

La evidencia computacional sugiere que la restricción se cumple y que los valores propios de esta matriz son positivos, pudiendo obtenerse expresiones simbólicas para la mayoría de los valores propios de la matriz. Sin embargo, no se logró una prueba certera encontrando el total de valores propios o utilizando algún método alternativo para verificar el cumplimiento de la restricción.

A continuación se proveen los vectores propios encontrados, partiendo por aquellos que son nulos. El núcleo de la matriz contiene al subespacio generado por el vector

$$(\alpha\mathbf{1}^\top, \beta^\top, -\alpha\mathbf{1}^\top, \beta^\top, \alpha/M)^\top,$$

para variables $\alpha \in \mathbb{R}$ y $\beta \in \mathbb{R}^n$. Esto se comprueba mediante el producto matriz-vector

$$\begin{pmatrix} I & L^\top & O & -L^\top & -M\mathbf{1} \\ L & (n-1)(L+L^\top) & -L & (1-n)(L+L^\top) & (1-n)M\mathbf{1} \\ O & -L^\top & I & L^\top & M\mathbf{1} \\ -L & (1-n)(L+L^\top) & L & (n-1)(L+L\top) & (n-1)M\mathbf{1} \\ -M\mathbf{1}^\top & (1-n)M\mathbf{1}^\top & M\mathbf{1}^\top & (n-1)M\mathbf{1}^\top & 2nM^2 \end{pmatrix} \begin{pmatrix} \alpha\mathbf{1} \\ \beta \\ -\alpha\mathbf{1} \\ \beta \\ \alpha/M \end{pmatrix} = \mathbf{0}_{4n+1},$$

utilizando la igualdad $L\mathbf{1} = \frac{n-1}{2}\mathbf{1}$ en el cálculo. Luego, se tienen al menos $n+1$ valores propios nulos. Además, las implementaciones computacionales indican que hay solo esa cantidad de valores propios cero. Resta, sin embargo, comprobar esta afirmación teóricamente.

Por último, para cualquier elección arbitraria de $\gamma \in \mathbb{R}^n$ se tiene la igualdad

$$\frac{\eta}{2n}\begin{pmatrix} I & L^\top & O & -L^\top & -M\mathbf{1} \\ L & (n-1)(L+L^\top) & -L & (1-n)(L+L^\top) & (1-n)M\mathbf{1} \\ O & -L^\top & I & L^\top & M\mathbf{1} \\ -L & (1-n)(L+L^\top) & L & (n-1)(L+L\top) & (n-1)M\mathbf{1} \\ -M\mathbf{1}^\top & (1-n)M\mathbf{1}^\top & M\mathbf{1}^\top & (n-1)M\mathbf{1}^\top & 2nM^2 \end{pmatrix} \begin{pmatrix} \gamma \\ \mathbf{0}_n \\ \gamma \\ \mathbf{0}_n \\ 0 \end{pmatrix} = \frac{\eta}{2n}\begin{pmatrix} \gamma \\ \mathbf{0}_n \\ \gamma \\ \mathbf{0}_n \\ 0 \end{pmatrix},$$

deduciéndose que existe un valor propio $\frac{\eta}{2n}$ con multiplicidad n. Igualmente, puede deducirse de los resultados computacionales que ésta es la máxima multiplicidad para tal valor propio. Sin embargo, debido a no tener conocimiento de los $2n$ valores propios restantes, no se puede concluir formalmente tal afirmación.

ANEXO C. RESULTADOS COMPLEMENTARIOS PARA PEP INCREMENTAL

C.1. Cálculo del dual SDP *batch*

En esta sección se realiza el cálculo del problema dual del modelo PEP incremental (6.12) presentado en la Sección 6.1. El problema (6.12) modela el peor caso real de una métrica $\mathcal{P}$ iSOLG-representable de estabilidad y convergencia del *gap* de optimalidad que actúa sobre pérdidas en la clase generalizada $\mathcal{F}_M$, para dimensiones $d \geq (2K+3)n+2$, realizando actualizaciones de algún método incremental $\mathcal{M}$ aplicado al problema de minimización de riesgo empírico (1.3), de acuerdo a la reformulación semidefinida expuesta en el Capítulo 6.

Se presenta la Proposición C.1, resultado que indica el dual al problema mencionado.

Proposición C.1. *El dual al problema* (6.12) *es el problema semidefinido*

$$
\begin{aligned}
&\inf_{\lambda_{ij},\mu_i,\tau,\kappa,\varphi} \quad \tau R^2 + \sum_{i\in I_M} \mu_i M^2 + \kappa \\
&\textit{s.a} \qquad \sum_{(i,j)\in\mathcal{I}} \lambda_{ij} A_{ij} + \sum_{i\in I_M} \mu_i A_{M_i} + \tau A_R + \kappa A_y + \varphi A_* - C \succeq O \\
&\qquad\quad \sum_{(i,j)\in\mathcal{I}} \lambda_{ij}(e_j - e_i) = b \\
&\qquad\quad \lambda_{ij}, \mu_i \geq 0 \quad ((i,j)\in\mathcal{I}) \\
&\qquad\quad \tau, \kappa \geq 0; \varphi \in \mathbb{R}.
\end{aligned}
$$

DEMOSTRACIÓN. *Se lleva el problema primal a una forma estándar para optimización semidefinida. Primero, se agregan las variables de holgura $s_{(\cdot)}$ correspondientes a*

cada restricción,

$$\begin{aligned} \underset{v\in\mathbb{R}^{(2K+3)n}, X\in\mathbb{S}^{(2K+3)n+2}}{\sup} \quad & b^\top v + \mathit{Tr}(CX) \\ \text{s.a} \quad & \mathit{Tr}(A_{ij}X) + v_j - v_i + s_{ij} = 0 \qquad (\forall i,j \in \mathcal{I}) \\ & \mathit{Tr}(A_{M_i}X) + s_i = M^2 \qquad (\forall i \in I_M) \\ & \mathit{Tr}(A_R X) + s_R = R^2 \\ & \mathit{Tr}(A_y X) + s_y = 1 \\ & \mathit{Tr}(A_* X) = 0. \end{aligned} \tag{C.1}$$

Para lograr la reformulación, se separan las partes positiva y negativa de v, $v = v^+ + v^-$. *Además, se considera la matriz diagonal por bloques de variables*

$$\tilde{X} = \begin{pmatrix} \mathit{diag}(s) & O & O & O \\ O & \mathit{diag}(v^+) & O & O \\ O & O & \mathit{diag}(v^-) & O \\ O & O & O & X \end{pmatrix}$$

y el vector de holguras $s = \{s_x\}_{x\in\mathcal{I}\cup I_M\cup\{R,y,*\}}$. *Luego, se definen las matrices paramétricas diagonales por bloques*

$$\tilde{A_{ij}} = \begin{pmatrix} \mathit{diag}(e_{ij}) & O & O & O \\ O & \mathit{diag}(e_j - e_i) & O & O \\ O & O & \mathit{diag}(e_i - e_j) & O \\ O & O & O & A_{ij} \end{pmatrix} \qquad (\forall (i,j) \in \mathcal{I})$$

$$\tilde{A_{M_i}} = \begin{pmatrix} \mathit{diag}(e_i) & O & O & O \\ O & O & O & O \\ O & O & O & O \\ O & O & O & A_{M_i} \end{pmatrix} \qquad (\forall i \in I_M)$$

$$\tilde{A_R} = \begin{pmatrix} \mathit{diag}(e_R) & O & O & O \\ O & O & O & O \\ O & O & O & O \\ O & O & O & A_R \end{pmatrix}$$

$$\tilde{A_y} = \begin{pmatrix} \mathit{diag}(e_y) & O & O & O \\ O & O & O & O \\ O & O & O & O \\ O & O & O & A_y \end{pmatrix}$$

$$\tilde{A_*} = \begin{pmatrix} O & O & O & O \\ O & O & O & O \\ O & O & O & O \\ O & O & O & A_* \end{pmatrix}$$

$$\tilde{C} = \begin{pmatrix} O & O & O & O \\ O & \mathit{diag}(b) & O & O \\ O & O & \mathit{diag}(-b) & O \\ O & O & O & C \end{pmatrix},$$

quedando una versión estándar del modelo primal anterior, expresada solo mediante restricciones de igualdad

$$\begin{aligned} \underset{s,v^+,v^- \geq 0; X \in \mathbb{S}^{4T+11}}{sup} \quad & Tr(\tilde{C}\tilde{X}) \\ s.a \quad & Tr(\tilde{A_{ij}}\tilde{X}) = 0 \qquad (\forall i,j \in \mathcal{I}) \\ & Tr(\tilde{A_{M_i}}\tilde{X}) = M^2 \qquad (\forall i \in I_M) \\ & Tr(\tilde{A_R}\tilde{X}) = R^2 \\ & Tr(\tilde{A_y}\tilde{X}) = 1 \\ & Tr(\tilde{A_*}\tilde{X}) = 0. \end{aligned} \tag{C.2}$$

Luego, se plantea el modelo dual estándar, considerando las variables duales $\kappa, \tau, \mu_i, \varphi$ y λ_{ij} asociadas respectivamente a las matrices por bloques $\tilde{A_y}, \tilde{A_R}, \tilde{A_{M_i}}, \tilde{A_*}$ y $\tilde{A_{ij}}$,

$$\begin{aligned} \underset{\{\lambda_{ij}\}_{(i,j)\in\mathcal{I}}, \{\mu_i\}_{i\in I_M}, \tau, \kappa, \varphi}{inf} \quad & \tau R^2 + \sum_{i \in I_M} \mu_i M^2 + \kappa \\ s.a \quad & \sum_{(i,j)\in\mathcal{I}} \lambda_{ij}\tilde{A_{ij}} + \sum_{i\in I_M} \mu_i \tilde{A_{M_i}} + \kappa\tilde{A_y} + \tau\tilde{A_R} + \varphi\tilde{A_*} - \tilde{C} \succeq O. \end{aligned} \tag{C.3}$$

Finalmente, se nota que la restricción semidefinida presenta sumas ponderadas de matrices diagonales por bloques, donde los tres primeros bloques de la diagonal corresponden a matrices diagonales, lo que entrega tres restricciones separadas. La primera corresponde a la condición

$$\sum_{(i,j)\in\mathcal{I}} \lambda_{ij} s_{ij} + \sum_{i\in I_M} \mu_i s_i + \kappa s_y + \tau s_R \succeq O,$$

que dada la no-negatividad de las holguras, es equivalente a que las variables duales, salvo φ, tomen valores no-negativos. La segunda y tercera son equivalentes al par de desigualdades vectoriales

$$\begin{cases} \sum_{(i,j)\in\mathcal{I}} \lambda_{ij}(e_j - e_i) - b \geq 0 \\ \sum_{(i,j)\in\mathcal{I}} \lambda_{ij}(e_i - e_j) + b \geq 0, \end{cases}$$

donde se puede concluir la igualdad a cero. Por lo tanto, el problema dual puede reescribirse de la forma expresada en (5.15)*:*

$$
\begin{aligned}
&\underset{\lambda_{ij},\mu_i,\tau,\kappa,\varphi}{\textit{inf}} && \tau R^2 + \sum_{i\in I_M} \mu_i M^2 + \kappa \\
&\textit{s.a} && \sum_{(i,j)\in\mathcal{I}} \lambda_{ij} A_{ij} + \sum_{i\in I_M} \mu_i A_{M_i} + \tau A_R + \kappa A_y + \varphi A_* - C \succeq O \\
& && \sum_{(i,j)\in\mathcal{I}} \lambda_{ij}(e_j - e_i) = b \\
& && \lambda_{ij}, \mu_i \geq 0 \quad ((i,j)\in\mathcal{I}) \\
& && \tau, \kappa \geq 0 \\
& && \varphi \in \mathbb{R}.
\end{aligned}
$$

concluyendo el resultado enunciado. □

La Proposición C.1 garantiza que el dual de problema incremental planteado para ambos SGD incremental e *IncrementalProx* es el presentado en el Capítulo 6. Este resultado permite su utilización para el estudio de acuerdo con la teoría de dualidad Lagrangiana. En la sección siguiente, se presenta un resultado de dualidad fuerte válido para ambos métodos y para las distintas métricas de rendimiento cuantificando el peor caso.

C.2. Dualidad fuerte para SDP incremental

En esta sección se presentan una garantía de dualidad fuerte válida para todos los modelos PEP incrementales desarrollados en el Capítulo 6. Esta demostración garantiza dualidad fuerte para el problema de maximizar una métrica de rendimiento que comprende los fenómenos de optimización (mediante el *gap* de optimalidad) y de estabilidad (mediante la diferencia en norma de las trayectorias) del peor caso de una implementación de los métodos SGD incremental e *IncrementalProx*. Se plantean nuevamente el resultado de dualidad fuerte para el modelo incremental ya introducido, junto a su prueba y a una descripción de cómo extender el resultado al análisis de errores de estabilidad y optimización por separado.

Proposición C.2. *Sea* (5.14) *iSOLG-representación del problema de peor caso conjunto de optimización y estabilidad de una ejecución de K rondas del método de gradiente estocástico incremental o el método IncrementalProx. Sea* (5.15) *su dual semidefinido, donde ambos problemas son factibles. Entonces, ambos poseen un valor óptimo idéntico y existe un punto primal-factible que alcanza tal valor.*

DEMOSTRACIÓN. *Se demuestra la dualidad fuerte mediante un certificado para la condición de Slater en el problema dual. Se nota que la matriz*

$$S = \tau A_R + \kappa A_y + \sum_{i \in I_M} \mu_i A_{M_i}$$

es una matriz diagonal y de rango completo para $\tau, \kappa, \mu_i > 0$. Se buscan valores apropiados de las variables duales de manera que la condición semidefinida del problema (6.13) *se cumpla de manera estricta, y adicionalmente, se cumpla la igualdad vectorial. En pos de esto, se buscan los valores propios de C (definido en* (6.11)*), notando que un cálculo rutinario entrega que los máximos valores propios en valor absoluto son*

$$\pm \frac{M\eta}{2} \sqrt{\frac{n(K+1)(2K+1)}{3K}}.$$

Además, se acota superiormente la norma de las matrices asociadas a la interpolación convexa para el caso no-suave. Sea $(i,j) \in \mathcal{I}$:

$$\begin{aligned} ||A_{ij}|| &= \sup_{||z||=1} \left[\langle u_j, z \rangle \langle h_i - h_j, z \rangle \right] \\ &\leq ||h_i - h_j||. \end{aligned}$$

Luego, fijando los valores

$$\varepsilon_{ij} = \begin{cases} \frac{1}{|\mathcal{I}|}||h_i - h_j||^{-1} & (i \neq j) \\ \frac{1}{|\mathcal{I}|} & (i = j) \end{cases}$$

$$\lambda_{ij} = \begin{cases} \frac{1}{n} + \varepsilon_{ij} & (i = (2K+2)n + l, j = (2K+2)n + l; l \in [n]) \\ \varepsilon_{ij} & \textit{e.o.c.} \end{cases},$$

se cumple la restricción de igualdad vectorial a b, definido en (6.10)*. Resta demostrar el cumplimento estricto de la restricción semidefinida, por lo que se acota superiormente la norma, notando el hecho que $h_{(2K+2)n+i} = 0$ para todo $i \in [n]$ y utilizando la desigualdad triangular se tiene*

$$\begin{aligned} \left|\left| \sum_{(i,j)\in\mathcal{I}} \lambda_{ij} A_{ij} - C \right|\right| &\leq ||h_{2Kn+1}|| + \sum_{(i,j)\in\mathcal{I}, i\neq j} \varepsilon_{ij}||A_{ij}|| + ||C|| \\ &< ||h_{2Kn+1}|| + 1 + \frac{M\eta}{2}\sqrt{\frac{n(K+1)(2K+1)}{3K}}. \end{aligned}$$

El cálculo de la norma restante entrega

$$\left|\left| \sum_{(i,j)\in\mathcal{I}} \lambda_{ij} A_{ij} - C \right|\right| < \sqrt{1 + \frac{n\eta^2(K+1)(2K+1)}{3K}} + 1 + \frac{M\eta}{2}\sqrt{\frac{n(K+1)(2K+1)}{3K}},$$

donde en la desigualdad estricta se utiliza el valor propio máximo de C y los valores fijados para las variables $\lambda_{i,j}$ más arriba. Finalmente, se considera B como la cota superior estricta calculada, que corresponde a una función solo de los parámetros del problema. Basta fijar el resto de las variables duales que conforman la diagonal con valor B y fijar $\varphi = 0$ para concluir la prueba con la desigualdad estricta

$$S + \sum_{(i,j)\in\mathcal{I}} \lambda_{ij} A_{ij} - C \succ O.$$

□

COMENTARIO C.1. *Al igual que las pruebas presentadas para modelos PEP batch, esta prueba es similar para las métricas separadas de estabilidad y optimización. En el caso de optimización,* $C = O$ *es mayorado estrictamente por el máximo valor propio presentado, siendo la demostración un certificado válido para este caso alternativo.*

En cambio, para el caso de estabilidad, la elección de $\lambda_{ij} = \varepsilon_{ij}$ *en todo par de* $\mathcal{I}$ *basta para lograr la igualdad vectorial a* $b = 0$*. Consecuentemente,* B *es una cota superior más laxa, pero aún válida.*

La prueba anterior funciona para los dos métodos incrementales presentados, ya que la elección de parámetros $h_{2Kn+i}, h_{(2K+2)n+i}$ y C es común a ambos. Se destaca el hecho que la elección de método, en términos del problema semidefinido, depende solo de los vectores $h_{(\cdot)}$ fijados. Se concluye que en ambos casos el *gap* de dualidad es nulo y el valor primal se alcanza.