Stochastic Approximation and Least-Squares Regression, with Applications to Machine Learning

Nicolas Flammarion

THÈSE DE DOCTORAT

de l'Université de recherche Paris Sciences Lettres
PSL Research University

Préparée à l'École normale supérieure

Stochastic Approximation and Least-Squares Regression, with Applications to Machine Learning

Approximation Stochastique et Régression par Moindres Carrés : Applications en Apprentissage Automatique

École doctorale n°386

ÉCOLE DOCTORALE DE SCIENCES MATHÉMATIQUES DE PARIS CENTRE

Spécialité MATHÉMATIQUES APPLIQUÉES

COMPOSITION DU JURY:

M. Jérôme Bolte
TSE Toulouse, Rapporteur

M. Shai Shalev-Shwartz
The Hebrew University of Jerusalem, Rapporteur (Absent)

M. Alexandre d'Aspremont
CNRS-ENS Paris, Directeur de thèse

M. Francis Bach
INRIA-ENS Paris, Directeur de thèse

M. Arnak Dalalyan
ENSAE Paris, Membre du Jury

M. Eric Moulines
CMAP EP Paris, Président du Jury

Soutenue par Nicolas Flammarion
le 24.07.2017

Dirigée par **Alexandre d'ASPREMONT**
et Francis BACH

ENS ÉCOLE NORMALE SUPÉRIEURE | PSL RESEARCH UNIVERSITY PARIS

Was Du für ein Geschenk hältst, ist
ein Problem, das Du lösen sollst.

L. Wittgenstein,
Vermischte Bemerkungen

What you are regarding as a gift is a problem for you to solve.

Dedicated to
my parents,
my sisters
and Adèle

Abstract

Many problems in machine learning are naturally cast as the minimization of a smooth function defined on a Euclidean space. For supervised learning, this includes least-squares regression and logistic regression. While small-scale problems with few input features may be solved efficiently by many optimization algorithms (e.g., Newton's method), large-scale problems with many high-dimensional features are typically solved with first-order techniques based on gradient descent, leading to algorithms with many cheap iterations.

In this manuscript, we consider the particular case of the quadratic loss. In the first part, we are interested in its minimization, considering that its gradients are only accessible through a stochastic oracle that returns the gradient at any given point plus a zero-mean finite variance random error. We propose different algorithms to efficiently solve these minimization problems in many cases. In the second part, we consider two applications of the quadratic loss in machine learning: unsupervised learning, specifically clustering and statistical estimation, specifically estimation with shape constraints.

In the first main contribution of the thesis, we provide a unified framework for optimizing non-strongly convex quadratic functions, which encompasses accelerated gradient descent, averaged gradient descent and the heavyball method. They are studied through second-order difference equations for which stability is equivalent to an $O(1/n^2)$ convergence rate. This new framework suggests an alternative algorithm that exhibits the positive behavior of both averaging and acceleration.

The second main contribution aims at obtaining the optimal prediction error rates for least-squares regression, both in terms of dependence on the noise of the problem and of forgetting the initial conditions. Our new algorithm rests upon averaged accelerated gradient descent and is analyzed under finer assumptions on the covariance matrix of the input data and the initial conditions of the algorithm which leads to tighter convergence rates expressed with dimension-free quantities.

The third main contribution of the thesis deals with the minimization of composite objective functions composed of the expectation of quadratic functions and a convex function. We show that stochastic dual averaging with a constant step-size has a convergence rate $O(1/n)$ without strong convexity assumption, extending earlier results on least-squares regression to any regularizer and any geometry represented by a Bregman divergence.

As a fourth contribution, we consider the problem of clustering high-dimensional data. We present a novel sparse extension of the discriminative clustering framework and propose a natural extension for the multi-label scenario. We also provide the first theoretical analysis of this formulation with a simple probabilistic model and

an efficient iterative algorithm with better running-time complexity than existing methods.

The fifth main contribution of the thesis deals with the seriation problem, which consists in permuting the rows of a given matrix in such way that all its columns have the same shape. We propose a statistical approach to this problem where the matrix of interest is observed with noise and study the corresponding minimax rate of estimation of the matrices. We also suggest a computationally efficient estimator whose performance is studied both theoretically and experimentally.

Keywords: Convex optimization, acceleration, averaging, stochastic gradient, least-squares regression, stochastic approximation, dual averaging, mirror descent, discriminative clustering, convex relaxation, sparsity, statistical seriation, permutation learning, minimax estimation, shape constraints.

Résumé

De nombreux problèmes en apprentissage automatique sont formellement équivalents à la minimisation d'une fonction lisse définie sur un espace euclidien. Plus précisément, dans le cas de l'apprentissage automatique supervisé, cela inclut la régression par moindres carrés et la régression logistique. Alors que les problèmes de petite taille, avec peu de variables, peuvent être résolus efficacement à l'aide de nombreux algorithmes d'optimisation (la méthode de Newton par exemple), les problèmes de grande échelle, avec de nombreuses données en grande dimension, sont, quant à eux, généralement traités à l'aide de méthodes du premier ordre, dérivées de la descente de gradient, conduisant à des algorithmes avec de nombreuses itérations peu coûteuses.

Dans ce manuscrit, nous considérons le cas particulier de la perte quadratique. Dans une première partie, nous nous intéressons à la minimisation de celle-ci dans l'hypothèse où nous accédons à ses gradients par l'intermédiaire d'un oracle stochastique. Celui-ci retourne le gradient évalué au point demandé plus un bruit d'espérance nulle et de variance finie. Nous proposons différents algorithmes pour résoudre efficacement ce problème dans de multiples cas. Dans une seconde partie, nous considérons deux applications différentes de la perte quadratique à l'apprentissage automatique : la première en apprentissage non-supervisé, plus spécifiquement en partitionnement des données, et la seconde en estimation statistique, plus précisément en estimation sous contrainte de forme.

La première contribution de cette thèse est un cadre unifié pour l'optimisation de fonctions quadratiques non-fortement convexes. Celui-ci comprend la descente de gradient accélérée, la descente de gradient moyennée et la méthode de la balle lourde. Ces méthodes sont étudiées grâce à des équations aux différences finies du second ordre dont la stabilité est équivalente à une vitesse de convergence $O(1/n^2)$ de la méthode étudiée. Ce nouveau cadre nous permet de proposer un algorithme alternatif qui combine les aspects positifs du moyennage et ceux de l'accélération.

La deuxième contribution est d'obtenir le taux optimal d'erreur de prédiction pour la régression par moindres carrés en fonction de la dépendance, à la fois au bruit du problème et à l'oubli des conditions initiales. Notre nouvel algorithme tire son origine de la descente de gradient accélérée et moyennée et nous l'analysons sous des hypothèses plus fines sur la matrice de covariance des données et sur les conditions initiales de l'algorithme. Cette nouvelle analyse aboutit à des taux de convergence plus tendus qui ne font pas intervenir la dimension du problème.

La troisième contribution de cette thèse traite du problème de la minimisation

de fonctions composites qui sont la somme de l'espérance de fonctions quadratiques et d'une régularisation convexe. Nous montrons qu'utilisée avec un pas constant, la méthode duale moyennée converge vers la solution du problème à la vitesse $O(1/n)$ sans hypothèse de forte convexité. Cela étend les résultats existants sur la régression par moindres carrés aux cas régularisés et aux différentes géométries induites par une divergence de Bregman.

Dans une quatrième contribution, nous considérons le problème de partitionnement de données de grande dimension. Nous présentons ainsi une nouvelle extension parcimonieuse du partitionnement discriminatif et son extension naturelle au cas de données avec de multiples labels. Nous analysons aussi cette formulation théoriquement, et ce pour la première fois, à l'aide d'un modèle probabiliste simple et nous proposons un nouvel algorithme itératif ayant une meilleure complexité que les méthodes existantes.

La dernière contribution de cette thèse aborde le problème de la sériation. Celui-ci consiste à permuter les lignes d'une matrice afin que ses colonnes aient toutes une forme identique. Nous adoptons une approche statistique et, dans le cas où la matrice est observée avec du bruit, nous étudions les taux d'estimation minimax correspondants. Nous proposons aussi un estimateur computationellement efficace et nous étudions ses performances d'un point de vue théorique et pratique.

Mots-clés : Optimisation convexe, accélération, moyennage, gradient stochastique, régression par moindres carrés, approximation stochastique, algorithme dual moyenné, descente miroir, partitionnement discriminatif, relaxation convexe, parcimonie, sériation statistique, apprentissage de permutation, estimation minimax, contraintes de forme.

Remerciements

Mes premières pensées vont à mes deux directeurs de thèse, Alexandre d'Aspremont et Francis Bach. Francis, tu as toujours été présent depuis la fin du MVA. Tu m'as tant appris, aussi bien humainement que mathématiquement. Je t'en serai toujours reconnaissant. Alexandre, même si nous n'avons pas encore eu l'occasion de réellement travailler ensemble, j'ai toujours pu compter sur tes conseils avisés et ton aide précieuse.

Jérôme Bolte et Shai Shalev-Shwartz m'ont fait l'immense honneur d'accepter de rapporter cette thèse. Je ne saurais assez les remercier.

Je tiens aussi à remercier chaleureusement Arnak Dalalyan et Eric Moulines d'avoir accepté d'être présents dans le jury. Eric, tes travaux avec Francis ont particulièrement inspiré ce manuscrit. Arnak, ton cours de master a très largement contribué à mon intérêt pour les statistiques.

Merci aussi à Philippe Rigollet de m'avoir accueilli près de quatre mois à Boston au printemps 2016. Philippe, tu es en partie responsable de mon envie de poursuivre mes recherches aux Etats-Unis l'année prochaine. J'espère un jour maîtriser aussi bien que toi les bornes inférieures statistiques et computationnelles, les barbecues et les bourbons. Mais j'en doute. La route sera longue.

Je tiens à remercier vivement Michael Jordan de m'accueillir à Berkeley. Une nouvelle aventure s'ouvre à moi.

J'ai également une pensée pour mes collaborateurs Aymeric Dieuleveut, Cheng Mao et Balamurugan Palaniappan. Comme d'autres activités, la recherche est plus intéressante à plusieurs et j'ai beaucoup profité de travailler avec vous. Je suis très fier d'avoir pu progresser à vos côtés et des résultats que nous avons obtenus ensemble.

Ce fut un réel plaisir d'effectuer ma thèse au département d'informatique de l'ENS et à l'INRIA. J'y ai bénéficié de conditions de travail exceptionnelles et d'une ambiance chaleureuse et je remercie pour cela son directeur Jean Ponce. Je suis heureux d'y avoir rencontré des collègues que je compte aujourd'hui parmi mes amis. Je remercie tous les doctorants que j'y ai côtoyés. En particulier mes différents cobureaux : mes mentors de la place d'Italie Vincent Delaitre et Edouard Grave, mes complices de la gare de Lyon, Aymeric Dieuleveut et Damien Garreau et les furtifs mais incisifs Rémy Degenne, Rémi Leblond, Horia Mania et John Weed. Mais aussi mes alcoolytes Jean-Baptiste Alayrac, Piotr Bojanowski, Nicolas Boumal, Guilhem Chéron, Théophile Dalens, Christophe Dupuy, Fajwel Fogel, Vadim Kantorov, Sesh Kumar, Rémi Lajugie, Loic Landrieu, Maxime Oquab, Julia Peyre, Anastasia Podosinnikova, Antoine Recanati, Vincent Roulet, Damien Scieur, Guillaume Seguin,

Nino Shervashidze, Matthew Trager et Gül Varol.

J'ai beaucoup voyagé pendant ces quelques années : pensées à mes amis du MIT, Afonso Bandeira, Victor-Emmanuel Brunel, Cheng Mao, Quentin Paris, Irène Waldspurger et John Weed. Mais aussi à mes collègues rencontrés lors de différentes occasions, Quentin Berthet, Alain Durmus, Pierre Gaillard et Vianney Perchet.

J'ai aussi pu toujours compter sur mes amis lyonnais, Clothilde F., Lucas H., Minh-Tu H., Nastya O., Quentin C., Selim G., Valentin H. et sur ceux de toujours Lisa P., Lucas B., Nicolas B., Paul M., Ulysse G. et Valentin M. Merci pour tout, pour votre aide, vos blagues, vos attentions et votre constance.

Last but not least, mes soeurs Sophie, Camille et Laure, ainsi que leurs compagnons et leurs enfants, d'année en année plus nombreux mais aussi et surtout mes parents Nadine et Jean-Noël sans qui je ne serais rien. Ils m'ont toujours soutenu et fait confiance en me laissant, depuis petit, une liberté totale. J'espère le leur rendre et qu'ils sont fiers de moi. J'ai aussi une pensée émue pour mes grands-parents aujourd'hui décédés mais qui ont égayé mon enfance et dont le souvenir ne me quitte pas. Et bien sûr pour Adèle, qui me rend meilleur, jour après jour.

Ce manuscrit a par ailleurs beaucoup profité des relectures attentives et bienveillantes de Fajwel, Lucas et Simon, du talent en dessin vectoriel de Camille ainsi que des aides cyrilliques de Dmitry Babichev, Tatiana Shpakova et Dmitry Zhukov.

Contents

Contributions

This thesis is divided into two parts. Chapter 2 to 4 discuss stochastic optimization of quadratic functions, while Chapter 5 and Chapter 6 concern applications of the quadratic loss in machine learning. Each chapter can be read independently of the others.

Chapter 1: This chapter is an introduction to statistical learning, convex optimization, stochastic approximation and online learning, which are the main topics of this manuscript. We overview the basic theoretical results through the unifying lens of the least-squares problem.

Chapter 2: This chapter considers a general framework for stochastic non-strongly convex quadratic optimization problems, including accelerated gradient descent, averaged gradient descent and the heavyball method. We provide a joint analysis explaining existing behavior and design a novel intermediate algorithm that exhibits the positive aspects of both acceleration (quick forgetting of initial conditions) and averaging (robustness to noise). This chapter is based on the article of Flammarion and Bach [2015], in the *Proceedings of the International Conference on Learning Theory.*

Chapter 3: This chapter presents a new algorithm, based on averaged accelerated regularized gradient descent, for optimizing quadratic objective functions whose gradients are only accessible through a stochastic oracle. We prove it achieves jointly the optimal prediction error rates for least-squares regression, both in terms of forgetting the initial conditions in $O(1/n^2)$, and in terms of dependence on the noise and the dimension d of the problem, as $O(d/n)$. We also analyze it through finer assumptions on the initial conditions and the Hessian matrix, leading to dimension-free quantities that may still be small while the "optimal" terms above are large. This chapter is based on the article of Dieuleveut et al. [2017], in the *Journal of Machine Learning Research.*

Chapter 4: This chapter considers the problem of minimizing composite objective functions composed of the expectation of quadratic functions and an arbitrary convex function. We prove the stochastic dual averaging algorithm converges at rate $O(1/n)$ without strong convexity assumptions. This extends earlier results on least-squares regression to all convex regularizers and all geometries induced by a Bregman diver-

gence. This chapter is based on the article of Flammarion and Bach [2017], in the *Proceedings of the International Conference on Learning Theory*.

Chapter 5: This chapter considers the problem of clustering high-dimensional data. The commonly used unsupervised learning algorithms have problems identifying the different clusters since they are easily perturbed by adding a few noisy dimensions to the data. This chapter considers the discriminative clustering formulation which aims at linearly separating noise from signal, i.e., finding a projection of the data that extracts the signal and removes the noise. We provide the first theoretical analysis of this formulation and a new efficient iterative algorithm with a complexity which depends only linearly on the number of observations, thus improving over previous results. We also propose a novel sparse extension to discriminative clustering to handle data with many irrelevant dimensions and we naturally extend these formulations to the multi-label scenarios where data share different labels, both potentially leading to interesting applications. This chapter is based on the article of Flammarion et al. [2017], in the *Journal of Machine Learning Research*.

Chapter 6: This chapter considers the seriation problems which consists in permuting the rows of a matrix in such way that all its columns have the same shape. While such problems are hard in general, it can be shown that some subproblems can be solved efficiently using spectral methods. However little is known about the robustness to noise of these methods. We study the minimax rate of estimation when the matrix is observed with noise by providing an upper bounds for the performance of the least-squares estimator together with corresponding lower bound. Unfortunately the least-squares estimator is intractable. Consequently we present a computationally efficient substitute estimator and analyze its rates of convergence both theoretically and numerically. This chapter is based on the article of Flammarion et al. [2016], under submission to *Bernoulli*.

N. Flammarion and F. Bach. From averaging to acceleration, there is only a step-size. In *Proceedings of the International Conference on Learning Theory (COLT)*, 2015.

A. Dieuleveut, N. Flammarion, and F. Bach. Harder, Better, Faster, Stronger Convergence Rates for Least-Squares Regression. In *Journal of Machine Learning Research*, 2017.

N. Flammarion and F. Bach. Stochastic Composite Least-Squares Regression with convergence rate $O(1/n)$. In *Proceedings of the International Conference on Learning Theory (COLT)*, 2017.

N. Flammarion, P. Balamurugan and F. Bach. Robust Discriminative Clustering with Sparse Regularizers. In *Journal of Machine Learning Research*, 2017.

N. Flammarion, C. Mao and P. Rigollet. Optimal Rates of Statistical Seriation. Under submission to *Bernoulli*, 2016.

Chapter 1

Introduction

1.1 Machine Learning

Machine learning is a recent scientific domain at the interface of applied mathematics, statistics and computer science. It aims at giving a machine the ability to produce a predictive analysis based on existing data and it applies to a tremendous variety of domains ranging from computer vision, natural language processing, advertising, bio-informatics, robotics, speech processing and economics. Its utmost challenge is to design general methods that can be applied across all of these domains. Machine learning has been particularly important in the current context of big data, that is a context of increasing data volumes, improved access to data and facilitated data collection.

Learning can be very loosely defined as the ability to answer a question after observing data. The related learning problems may be described through the interaction between the learner and the environment. An algorithm able to detect if a cat appears in an image, first needs to be trained on a dataset of examples. This dataset would consist of many pictures with a label saying if a cat is present or not. On the basis of this training it would design a rule to decide if there is a cat in a new picture. This is what is called *supervised learning*. In contrast for image segmentation, the learner receives all the images without label and he has to automatically detect the contours of objects in it. This is what is called *unsupervised learning*.

This thesis is centered around optimization methods for supervised learning. Nonetheless we will still consider extensions to unsupervised learning in Chapter 5. We note there are also other extensions (e.g., reinforcement learning) but they are outside the scope of this manuscript.

1.2 Supervised Machine Learning

Supervised machine learning aims at understanding the relationship between elements of an arbitrary input set $\mathcal{X}$ and elements of an arbitrary output set $\mathcal{Y}$. Typically $\mathcal{X} = \mathbb{R}^d$ for a large d and $\mathcal{Y}$ is a finite set or a subset of $\mathbb{R}$. For example:

— $\mathcal{X}$ is a set of images that contains a hand-written number and $\mathcal{Y}$ is the set of

associated numbers. A picture is coded through the grey level of its pixels and $\mathcal{X} = [0,1)^d$ for d pixels and $\mathcal{Y} = \{0, \dots, 9\}$.

— $\mathcal{X}$ summarizes the air pollution, the crime rate, the high-school quality, the percent of green spaces around certain neighborhoods and $\mathcal{Y} = \mathbb{R}$ depicts the housing prices.

Unfortunately an output $y \in \mathcal{Y}$ cannot be always expressed exactly as a function of an input $x \in \mathcal{X}$ since there may be some random noise or unknown factors. So, instead the couple (X, Y) is modeled as random variables. Therefore we aim to predict the output Y associated to the input X where it is given that (X, Y) is sampled from an unknown distribution $\mathcal{D}$. We do this prediction through a *predictor* which is defined as a measurable function $h : \mathcal{X} \to \mathcal{Y}$. The set of all predictors is denoted by $\mathcal{F}(\mathcal{X}, \mathcal{Y})$. A predictor is also called a *hypothesis* or an *estimator*. In order to measure the performance of a predictor we define a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ where $\ell(y, y')$ is the loss incurred when the true output is y whereas y' is predicted. For instance:

Classification: $\mathcal{Y}$ is a finite set and $\ell(y, y') = \mathbb{1}_{y \neq y'}$ is the $0-1$ loss. In binary classification this would be $\mathcal{Y} = \{0, 1\}$.

Least-squares regression: $\mathcal{X} = \mathbb{R}$ and $\ell(y, y') = |y - y'|^2$. Least-squares regression is the main topic of this thesis and one of the most classical problems of statistical learning. It will be used to illustrate numerous examples throughout this introduction. In particular we shall be considering the *parametric least-squares* framework: we shall assume a linear parameterization of the predictor, $h(x) = \langle \theta, \phi(x) \rangle$ where the features $\phi(x) \in \mathbb{R}^d$ are designed by experts using their knowledge of the phenomenon, or are independently learned, e.g., with neural networks [Bengio et al., 2013]. These features have a key importance in practice because linear predictors with relevant features may be more efficient than non-linear predictors. Furthermore we note that a linear parametrization in θ does not imply a linear parametrization in x since ϕ may be non-linear.

The quality of a prediction is measured by the *generalization error* (also called *risk*) of a predictor defined by

$$L(h) = \mathbb{E}_{(X,Y)\sim D} \ell(Y, h(X)).$$

This is the averaged loss incurred when the learner predicts Y by $h(X)$ and the data (X, Y) are sampled following $\mathcal{D}$. Thus the learner wants to solve the minimization problem

$$\min_{h \in \mathcal{F}(\mathcal{X}, \mathcal{Y})} L(h).$$

The solution of this problem (when it exists) is called the *Bayes predictor*. For least-squares regression the generalization error of a predictor h can be decomposed as

$$L(h) = \frac{1}{2}\mathbb{E}[Y - \mathbb{E}(Y|X)]^2 + \frac{1}{2}\mathbb{E}[\mathbb{E}[Y|X] - h(X)]^2.$$

In other words $\mathbb{E}[Y|X]$ is the Euclidean projection of Y onto the set $L^2(X)$ of the square integrable functions of X. Consequently $L(h)$ is always larger than $\frac{1}{2}\mathbb{E}[Y -$

$\mathbb{E}[Y|X]]^2$ and the Bayes predictor is (in this particular case)

$$h^*(x) = \mathbb{E}[Y|X = x].$$

Regrettably since the distribution $\mathcal{D}$ is unknown, this function is not computable. Therefore we assume that the learner is observing a finite *training set* of points $S_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$ which are sampled independently from the unknown law $\mathcal{D}$ and wants to use them to predict the output Y associated to the input X, where (X, Y) is sampled from $\mathcal{D}$ independently from S_n. This task is formalized through a *learning algorithm*: a function $A : \cup_{n\geq 1}(\mathcal{X} \times \mathcal{Y})^n \to \mathcal{F}(\mathcal{X} \times \mathcal{Y})$ which relates a training set S_n to a predictor $A(S_n)$. Since the training set S_n is random, the generalization error of the predictor $A(S_n)$ is a random variable and the quality of a learning algorithm A is measured by $\mathbb{E}_{S_n} L(A(S_n))$, where the expectation is measured with respect to the distribution of S_n. Alternatively $L(A(S_n))$ may be controlled in probability.

Let us stress that we adopt the *distribution-free* approach of Vapnik and Chervonenkis [1971, 1974]; see, e.g., Devroye et al. [1996]. There is no assumptions on the distribution $\mathcal{D}$ and the method has to be efficient independently of it. In contrast classical statistics first assumes a particular *statistical model* for the distribution $\mathcal{D}$ and then estimate the parameters of this model. This approach will be temporarily taken in Chapter 6. It is also different from the *probably approximately correct* (PAC) learning framework introduced by Valiant [1984] but the comparison is outside the scope of this introduction. Interested readers can see the monograph by Shalev-Shwartz and Ben-David [2014] for more precisions on these frameworks.

As the distribution $\mathcal{D}$ is unknown the Bayes predictor cannot be directly computed and the generalization error cannot even be minimized. Instead the training set S_n is used to approximate these objects. There are two principal ways to address this problem:

Local averaging methods: They estimate the Bayes predictor $\mathbb{E}[Y|X = x]$ by averaging the Y_i's corresponding to the X_i's close to x. This includes, for example, the Nadaraya-Watson estimator [Nadaraya, 1964, Watson, 1964] or the k-nearest neighbors algorithm Cover and Hart [1967]. These methods are well studied and efficient for dimensions of small to middle scale compared to n [see, e.g., Hastie et al., 2009].

Empirical Risk Minimization (ERM): It approximates the generalization error by the *training error* (also called *empirical risk*) $\hat{L}$ defined by the average error over the sample S_n

$$\hat{L}(g) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, h(Y_i)),$$

and considers its minimizer [see, e.g., Vapnik, 1998]. This method raises two principal issues:

Statistical problem: How well does the minimizer of the empirical risk perform? This question will be studied in Section 1.4.

Optimization problem: How to minimize the empirical risk $\hat{L}$? This question

will be examined in Section 1.5.

We will succinctly address these two questions through the lens of the least-squares regression. First of all we will introduce the mathematical framework adopted in this thesis.

1.3 Mathematical Framework

We shall now consider the Euclidean space $\mathbb{R}^d$ of dimension $d \in \mathbb{N}^*$ endowed with the natural inner product $\langle \cdot, \cdot \rangle$ and the Euclidean norm $\| \cdot \|_2$.

The central concept in this thesis is *convexity*:

Definition 1. *A set $\mathcal{C} \in \mathbb{R}^d$ is said to be convex if for all $\theta_1, \theta_2 \in \mathcal{C}$ and all $t \in (0,1)$,*

$$(1-t)\theta_1 + t\theta_2 \in \mathcal{C}.$$

While classical functional analysis does not attach the same importance to convex functions as to convex sets [Clarke, 2013], the former is a key concept for optimization theory:

Definition 2. *A extended-valued function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is said to be convex if for all $\theta_1, \theta_2 \in \mathbb{R}^d$ and all $t \in (0,1)$,*

$$f\big((1-t)\theta_1 + t\theta_2\big) \leq (1-t)f(\theta_1) + tf(\theta_2).$$

When f is twice differentiable this condition is equivalent to a symmetric positive-definite Hessian. It is worth noting that these concepts of convexity are independent of the notions of norm or even distance. Indeed convex sets and convex functions only need the basic operations of a vector space to be defined. The class of convex functions is very general and this thesis is restricted, for convenience, to the convex functions which are *closed* (their epigraphs $\mathrm{epi}(f) = \{(\theta, \alpha) \in \mathbb{R}^d \times \mathbb{R} : f(\theta) \leq \alpha\}$ are closed sets) and *proper* (not identically $+\infty$). We refer to the monographes by Rockafellar [1970], Hiriart-Urruty and Lemaréchal [2001] for more details on convex analysis.

Conceptually, convexity enables to turn a local information about the function into a global one. For example a local minimum is automatically a global minimum or even more importantly, the gradient $\nabla f(\theta)$ at a point $\theta \in \mathbb{R}^d$ provides a global linear lower-bound on the function f.

Functions met in machine learning applications often have additional properties. A continuously differentiable function f is said to be *L-smooth* for $L \in \mathbb{R}_+$ when the gradient of f is L-Lipschitz. When f is twice differentiable, this is equivalent to a global upper-bound on the Hessian of f. Thus this assumption provides a quadratic upper bound on the function value. For example, the empirical risk $\hat{L}(h) = \frac{1}{n}\sum_{i=1}^n \ell(Y_i, h(X_i))$ is usually smooth when the loss function ℓ is smooth and the data X_i are almost surely (abbreviated to from now on as a.s.) bounded.

On the other side a function f is *μ-strongly convex* for $\mu \in \mathbb{R}_+$ if $f - \frac{\mu}{2}\| \cdot \|_2^2$ is convex. When f is twice differentiable this is equivalent to a global lower-bound on

the Hessian of f and then this assumption provides a quadratic lower-bound on the function value. In this sense, μ measures the curvature of the function f.

strongly convex and smooth functions have several interesting properties we will not review here [see, e.g., Nesterov, 2004]. Moreover these two assumptions are connected: if f is μ-strongly convex then its Fenchel conjugate f^* is $1/\mu$-smooth [Hiriart-Urruty and Lemaréchal, 2001, Theorem 4.2.1]. On the other hand, contrary to convexity, both smoothness and strong convexity depend on the norm considered and Bauschke et al. [2016] recently relaxed this dependency.

1.4 Analysis of Empirical Risk Minimization

Since the data distribution $\mathcal{D}$ is unknown, the generalization error is approximated by the training error $\hat{L}(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, h(Y_i))$ which may be optimized to obtain a predictor h. Firstly we notice that the predictor $h(x) = y_i \mathbb{1}_{x_i}(x)$ has a training error $\hat{L}(h) = 0$ whereas its generalization error may be arbitrarily large. This is the *overfitting* phenomenon which appears when the prediction fits the data too closely and does not try to generalize enough. To prevent this effect, one may either adopt:

Constraint formulation: The hypothesis class is restricted to a smaller subset $\mathcal{G} \subset \mathcal{F}$. The empirical risk minimization for this class is the learning algorithm defined by

$$\hat{h} = \arg\min_{h \in \mathcal{G}} \hat{L}(h).$$

A bias may be created by restricting too much the class of predictor and the sub-class has to be fixed in advance (without observing the data).

Penalized formulation: A penalization $\phi(h)$ is added to the empirical risk $\hat{L}(h)$ and one solves the problem

$$\min_{h \in \mathcal{F}} \hat{L}(h) + \lambda \phi(h) \quad \text{for } \lambda \in \mathbb{R}_+.$$

The penalization $\phi(h)$ controls the complexity of the predictor h by implicitly inducing a tradeoff between predictors with a small training error but a large value of $\phi(h)$ and predictors with larger training error but smaller penalization $\phi(h)$. The penalization also induces a bias since the predictor would be the minimizer of the penalized problem which is different from the initial non-penalized problem.

Even if the penalized and constraint formulations are equivalent by convex duality [Borwein and Lewis, 2000, Sec. 4.3], the penalized formulation is easier to use in practice from an algorithmic point of view since (a) unconstrained optimization is easier, (b) there are efficient ways to set the value λ. On the other hand the constraint formulation is more appropriate for the theoretical analysis [Bach et al., 2012, p.7].

The generalization error of the ERM predictor $\hat{h}$ can be decomposed as follows:

$$L(\hat{h}) - L(h^*) = \underbrace{L(\hat{h}) - \min_{h \in \mathcal{G}} L(h)}_{\text{Estimation error}} + \underbrace{\min_{h \in \mathcal{G}} L(h) - L(h^*)}_{\text{Approximation error}}$$

Estimation error: It appears because the training error is an estimate of the generalization error. It depends on the size of the training set and on the complexity of the class $\mathcal{G}$.

Approximation error: It measures how much is lost from restricting the set of predictors. It does not depend on the sample size.

Thus there is a so-called *bias-variance tradeoff*: choosing a large class results in a small approximation error and a large estimation error since it leads to overfitting. We note that choosing a small class with dedicated structure may help in terms of computational complexity. It is also worth noting that only the estimation error depends on the predictor $\hat{h}$. We denote by $\tilde{h} = \arg\min_{h \in \mathcal{G}} L(h)$. The estimation error is uniformly bounded by

$$\begin{aligned} L(\hat{h}) - L(\tilde{h}) &= L(\hat{h}) - \hat{L}(\hat{h}) + \underbrace{\hat{L}(\hat{h}) - \hat{L}(\tilde{h})}_{\leq 0} + \hat{L}(\tilde{h}) - L(\tilde{h}) \\ &\leq 2 \sup_{g \in \mathcal{G}} |L(h) - \hat{L}(h)|. \end{aligned}$$

Under Lipschitz assumptions on the loss, this supremum is typically of order $O(1/\sqrt{n})$ [Boucheron et al., 2013]. When the loss function is, in addition, μ-strongly convex, Sridharan et al. [2009], Boucheron and Massart [2011] show that the estimation error is of order $O(1/\mu n)$ by directly bounding $L(\hat{h}) - L(\tilde{h})$ without using a uniform bound as above.

We only consider here *linear predictors*, i.e., linearly parameterized by $\theta \in \mathbb{R}^d$ as $h(x) = \langle \theta, \phi(x) \rangle$ for features $\phi(x) \in \mathbb{R}^d$. The error of a parameter θ is directly defined as $L(\theta) = L(\langle \theta, \phi(\cdot) \rangle)$ by a slight abuse of notation. The generalization error is assumed to be minimized among all linear predictors and we denote by θ^* one of its minimizer [1]:

$$\theta^* \in \arg\min_{\theta \in \mathbb{R}^d} L(\theta).$$

It is important to emphasize that the Bayes predictor h^* is never assumed to be linear as in the classical regression analysis (which assumes a *well-specified* linear model $Y = \langle \theta_*, X \rangle + \varepsilon$ for some independent zero-mean noise ε).

Parametric least-squares regression. For linear least-squares regression this bias-variance decomposition takes the form

$$\begin{aligned} L(\theta) - L(h^*) &= \frac{1}{2}\mathbb{E}[\mathbb{E}[Y|X] - \langle \theta_*, \phi(x) \rangle]^2 + \frac{1}{2}\langle \theta_* - \theta, \mathbb{E}[\phi(x) \otimes \phi(x)](\theta_* - \theta) \rangle \\ &= \frac{1}{2}\mathbb{E}[\mathbb{E}[Y|X] - \langle \theta_*, \phi(x) \rangle]^2 + \frac{1}{2}\|\theta_* - \theta\|_\Sigma^2, \end{aligned}$$

where the covariance matrix of the features is denoted by $\Sigma = \mathbb{E}[\phi(x) \otimes \phi(x)]$. By setting the gradient of the generalization error to 0, the optimum θ_* satisfies the

1. The generalization error does not always attain its global minimum, e.g., in logistic regression when the model is not well-specified [Bach, 2010].

normal equation

$$\Sigma\theta_* = \mathbb{E}[\phi(X)Y].$$

This implies, when Σ is full-rank, that $\theta_* = \Sigma^{-1}\mathbb{E}[\phi(X)Y]$. We will consider the *ordinary least-squares* estimator $\hat{\theta}$ which is the minimizer of the training error over linear predictors. Some authors also consider the *ridge* estimator [Hoerl, 1962] which is the solution of the ℓ_2-regularized training error minimization $\min_{\theta\in\mathbb{R}^d} \hat{L}(\theta) + \lambda\|\theta\|_2^2$.

We will now study the estimation error in the linear least-squares framework. It is a random variable which depends on the training-set S_n and its expectation can be decomposed as:

$$\frac{1}{2}\mathbb{E}\|\hat{\theta} - \theta_*\|_\Sigma^2 = \underbrace{\|\mathbb{E}\hat{\theta} - \theta_*\|_\Sigma^2}_{\text{Estimation bias}} + \underbrace{\mathbb{E}\|\mathbb{E}\hat{\theta} - \hat{\theta}\|_\Sigma^2}_{\text{Estimation variance}} .$$

Note that we only compare to the class of linear predictors. In particular, this predictor has a greater error than the Bayes predictor. This analysis becomes very simple when the input data X_n are assumed to be deterministic: this is the *fixed design* framework.

Fixed design analysis. We consider here that the input observations $(X_1, \dots, X_n)$ are deterministic. To emphasize this aspect, we denote them by $(x_1, \dots, x_n)$. The only randomness is thus in the sampling of the Y_i. We use the notation $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n \phi(x_i)\otimes\phi(x_i) = \Phi^\top\Phi$ and $\vec{Y} = [Y_1, \dots, Y_n]^\top \in \mathbb{R}^n$. For the sake of clarity, the design matrix Φ is assumed to be of rank d. This assumption may be relaxed using Moore-Penrose pseudo-inverse. Then

$$\theta_* = (\Phi^\top\Phi)^{-1}\Phi^\top\mathbb{E}[\vec{Y}] \quad \text{and} \quad \hat{\theta} = (\Phi^\top\Phi)^{-1}\Phi^\top\vec{Y}.$$

Therefore the estimation error is

$$L(\hat{\theta}) - L(\theta_*) = \left\|\Pi(\mathbb{E}\vec{Y} - \vec{Y})\right\|^2,$$

where the orthogonal projection matrix is denoted by $\Pi = \Phi(\Phi^\top\Phi)^{-1}\Phi^\top$. And the expected estimation error is

$$\mathbb{E}L(\hat{\theta}) - L(\theta_*) = \operatorname{tr}(\Pi \operatorname{var}\vec{Y}).$$

If $\vec{Y}$ satisfies $\operatorname{var}\vec{Y} \preccurlyeq \sigma^2 I$. Then

$$L(\hat{\theta}) - L(\theta_*) \leq \sigma^2 d/n.$$

Equality is obtained when $\operatorname{var}(Y_i) = \sigma^2$ and the full-rank assumption on the design matrix may be relaxed to provide a bound of the form $\sigma^2\operatorname{rank}(X)/n$. We also note that one obtains high-probability bounds using extra-assumptions on the distribution of the noise [van der Vaart and Wellner, 1996].

This result provides an upper-bound on the estimation error of the least-squares

estimator. However a different proof could improve upon this bound and a different estimator could also provide better performance. The *minimax* theory [Tsybakov, 2009] was developed as a way to handle at once all possible learning algorithms from a given class. This field which takes its roots in information theory [Cover and Thomas, 2006] provides a systematic way to show uniform bounds on their performance, even though this may at first appear daunting. An estimator $\hat{\theta}$ is *minimax optimal* and a positive sequence $(\psi_n)_{n\geq 0}$ is the *minimax rate of estimation* over the class Θ if there exist constants $c, C > 0$ such that

$$\sup_{\theta\in\Theta} \mathbb{E}\|\hat{\theta} - \theta\|_2^2 \leq c\psi_n,$$

and

$$\liminf_{n\to\infty} \inf_{\tilde{\theta}} \sup_{\theta\in\Theta} \psi_n^{-1} \mathbb{E}\|\tilde{\theta} - \theta\|_2^2 \geq C',$$

where the infimum is over all estimators (measurable functions of the data). In the fixed design setting, it is possible to show that $\sigma^2 d/n$ is optimal over $\mathbb{R}^d$ [Tsybakov, 2009].

Random design analysis. The input observations are not assumed deterministic anymore. We start with the classical asymptotic study which assumes d fixed and n going to ∞. We denote by $\varepsilon = \langle \theta_*, \phi(X)\rangle - Y$ and $\xi = \varepsilon\phi(X)$. By the law of large numbers $\Phi^\top\Phi \xrightarrow{P} \Sigma$ and $\Phi Y \xrightarrow{P} \mathbb{E}[\phi(X)Y]$. We also assume here $\Sigma \succcurlyeq \mu I$ for $\mu > 0$. Therefore $\hat{\theta} \xrightarrow{P} \theta_*$ by the continuous mapping theorem. The *Delta method* [van der Vaart, 1998, Chapter 3] may be applied to obtain asymptotic normality:

$$\sqrt{n}(\hat{\theta} - \theta_*) \xrightarrow{D} N(0, \Sigma^{-1}\mathbb{E}[\xi \otimes \xi]\Sigma^{-1}).$$

This matches the Cramer-Rao bound [Casella and Berger, 1990] and therefore this estimator is asymptotically optimal. One proceeds in the same way to asymptotically bound the estimation error

$$n[L(\hat{\theta}) - L(\theta_*)] \xrightarrow{D} \frac{1}{2} \operatorname{tr} \mathcal{W}(\Sigma^{-1/2}\mathbb{E}[\xi \otimes \xi]\Sigma^{-1/2}, 1),$$

where $\mathcal{W}(\Sigma, k)$ is the Wishart distribution with variance matrix Σ and k degrees of freedom. Under the assumption $\mathbb{E}\xi\otimes\xi \preccurlyeq \sigma^2\Sigma$, the asymptotic variance is bounded by $\frac{\sigma^2 d}{n}$ as in the fixed design analysis. Using Taylor expansions, this asymptotic analysis can be extended to any loss functions.

The non-asymptotic study is more involved. It has been pursued by Györfi et al. [2006], Audibert and Catoni [2011], Hsu et al. [2014], Lecué and Mendelson [2016], Oliveira [2016]. Hsu et al. [2014] obtain results from fixed design analysis using that $\Sigma^{1/2}\hat{\Sigma}^{-1}\Sigma^{1/2}$ is concentrated around the identity. They consider $\bar{\theta} = \mathbb{E}[\hat{\theta}|X_1, \ldots, X_n]$, the conditional expectation of the least-squares estimator which is equal to $\bar{\theta} = \hat{\Sigma}^{-1}\hat{\mathbb{E}}[Xh^*(Y)]$, where the empirical expectation is denoted by $\hat{\mathbb{E}}$. We

sketch below the main lines of their proof which holds on the decomposition

$$\begin{aligned}\Sigma^{1/2}(\hat{\theta}-\theta_*) &= \Sigma^{1/2}(\hat{\theta}-\bar{\theta})+\Sigma^{1/2}(\bar{\theta}-\theta_*)\\ &= \Sigma^{1/2}\hat{\Sigma}^{-1/2}\hat{\mathbb{E}}[\hat{\Sigma}^{-1/2}X(Y-h^*(Y))]\\ &\quad +\Sigma^{1/2}\hat{\Sigma}^{-1}\Sigma^{1/2}\hat{\mathbb{E}}[\Sigma^{-1/2}X(h^*(X)-X^\top\theta_*)].\end{aligned}$$

They first observe that $\hat{\mathbb{E}}[\hat{\Sigma}^{-1/2}X(Y-h^*(Y))]$ is related to the fixed design excess loss and goes to $\mathbb{E}[Y-h^*(X)]=0$. The second term $\hat{\mathbb{E}}[\Sigma^{-1/2}X(h^*(X)-X^\top\theta_*)]$ corresponds to the approximation error and will be close to $\mathbb{E}[X(h^*(X)-X^\top\theta_*)]=0$ by optimality of θ_*. Under sub-Gaussian assumptions on the noise ε, Hsu et al. [2014] show that $L(\hat{\theta})-L(\theta_*)=O(x\sigma^2 d/n)$ probability with exponentially close to 1 (at least $1-\exp(-cdx)$). Similar results were obtained by Lecué and Mendelson [2013].

Under much weaker assumptions on the noise, Lecué and Mendelson [2016] provide a weak polynomial probability estimate:

Theorem 1. *There exist absolute constants c_0, c_1, c_2 for which the following holds. Assume there exists κ for which*

$$\mathbb{E}\langle\theta,\phi(X)\rangle^4 \le \kappa(\mathbb{E}\langle\theta,\phi(X)\rangle^2)^2 \quad \forall\theta\in\mathbb{R}^d,$$

and $\sigma=(\mathbb{E}(\varepsilon^4))^{1/4}<\infty$. Let $n\ge(c_0\kappa)^2 d$. Then, for every $\delta>0$ with probability $1-\exp(-n/c_1\kappa^2)-\delta$,

$$L(\hat{\theta})-L(\theta_*)\le\frac{c_2\kappa^6}{\delta}\frac{\sigma^2 d}{n}.$$

Oliveira [2016] obtains stronger polynomial probability under stronger assumptions on the noise. Even so Lecué and Mendelson [2016] explain it is impossible to get an exponential probability estimate without stronger moment assumptions.

Tsybakov [2003], Shamir [2015] show this rate is also optimal in the random design setting. Besides, it is worth noting a lower bound in the fixed design framework does not directly imply a lower bound in the random design setting [Tsybakov, 2009, Sec 2.7.2].

To conclude, for the parametric least-squares regression, the prediction error is $O(\sigma^2 d/n)$ under very light assumptions. Yet these sharp results are very challenging and their non asymptotic analysis is very recent. Even though (as will become clear later) this thesis is centered on techniques which avoid minimizing the empirical risk, we will now present some optimization methods which achieve that goal.

1.5 Complexity Results in Convex Optimization

This thesis is focused on convex optimization and mainly on the study of optimization algorithms for quadratic functions. Nowadays, optimization is indeed divided between convex problems (which can be efficiently solved) and non-convex problems (for which there are no efficient and generic methods).

Interestingly, this distinction was initially drawn between linear and non-linear problems. *Linear programming* is now essentially solved: (a) in practice since 1947

by Dantzig's simplex algorithm (with an exponential worst-case complexity); (b) in theory by Khachiyan [1979] who achieved polynomial complexity using the ellipsoid method (with very slow convergence in practice); and (c) in both theory and practice since Karmakar [1984] proposed the first efficient polynomial-time algorithm using *interior point methods*.

This winding path from practical to theoretical progress has fed and inspired convex optimization since the seminal works of von Neumann [1963] and those of Kuhn and Tucker [1951]. Nemirovski and Yudin [1979] first showed that all convex minimization problems with a *first-order oracle* can theoretically be solved in polynomial time using the ellipsoid method, but with slow convergence in practice. Then Nesterov and Nemirovskii [1994] extended the interior point method to efficiently solve a wider class of convex problems.

However when the complexity of problems becomes higher, this generic solver becomes inefficient and older *first-order methods* are preferred at the cost of precision to the advantage of many cheap iterations. Unfortunately there is no unified analysis for these large-scale problems and the algorithmic choices are constrained by the problem structure. This has prompted both theoreticians and practitioners to move away from *black-box optimization* and leverage the specifics of the problem being considered to design better adapted algorithms.

Therefore large-scale optimization contributes to blur the frontier between convexity and non-convexity by defining new common territories, i.e., algorithms which can efficiently solve different problems indifferently of their convex aspects [Ge et al., 2016, Jin et al., 2016].

1.5.1 Black Box Optimization and Lyapunov Analysis

We consider the general optimization problem:

$$\min_{\theta \in \mathcal{C}} f(\theta),$$

where $\mathcal{C}$ is a convex set and f is a convex function. In order to conceptualize the optimization task, we adopt the *black box* framework defined by Nemirovski and Yudin [1979]. The convex set $\mathcal{C}$ is known to the algorithm but not the objective function f. The only assumption made on f is that it belongs to some class of functions $\mathcal{F}$. The information about f is obtained through interacting with an *oracle*. When queried at a given point θ, the *first-order oracle* we consider here, answers by giving the function value $f(\theta)$ and the gradient $\nabla f(\theta)$. Therefore a first-order black box procedure is a sequence of mappings $(\phi_n)_{n\geq 0}$ where $\phi_n : \mathcal{C}^n \times (\mathbb{R}^d)^n \times \mathbb{R}^n \to \mathcal{C}$. The algorithm starts from $\theta_0 = \phi_0$ and then iterates:

$$\theta_n = \phi_{n-1}\big(\theta_0, \ldots, \theta_{n-1}, \nabla f(\theta_0), \ldots, \nabla f(\theta_{n-1}), f(\theta_0), \ldots, f(\theta_{n-1})\big).$$

Our aim is to determine the *oracle complexity* of the problem. This is the number of oracle queries necessary to solve the problem at a precision of ε, i.e., to find an admissible point $\theta \in \mathcal{C}$ such that $f(\theta) - \min_{\tilde{\theta}\in\mathcal{C}} f(\tilde{\theta}) \leq \varepsilon$. To that purpose we should:

(a) find an algorithm whose convergence rate matches the desired rate, this would imply an upper bound on the complexity; (b) provide a lower-bound on the complexity by showing that no admissible method can solve the problem faster. This is usually done by finding a particular function for which it is possible to bound from below the performance of any method.

We note that oracle complexity does not directly provide information about the *computational complexity of the method* since the oracle query and the mapping ϕ_n may be computationally demanding. Therefore we will also pay attention to the computational complexity of each method. This explains why we often restrict ourselves to algorithms with finite-order iteration of the form:

$$\theta_n = \phi_{n-1}(\theta_{n-k}, \dots, \theta_{n-1}, \nabla f(\theta_{n-k}), \dots, \nabla f(\theta_{n-1}), f(\theta_{n-k}) \dots, f(\theta_{n-1})).$$

They can be reformulated (after a change of variable $\Theta_n = (\theta_{n_1}, \dots, \theta_n)$) as iterative processes of the form

$$\Theta_n = F_{n-1}(\Theta_{n-1}).$$

To prove the convergence of these dynamical systems, it is common to rely on Lyapunov theory. It goes back to Lyapunov [1892] who showed that all the trajectories of the ordinary differential equation (ODE)

$$\dot{x}(t) = Ax(t)$$

goes to zero (the ODE is *stable*) if and only if there exists a symmetric positive-definite matrix P such that

$$A^\top P + PA \preccurlyeq 0.$$

This is the origin of *Lyapunov's first method* which is based on the equivalence for the linear iterative process

$$\theta_n = A\theta_{n-1}$$

between

$$\lim_{n\to\infty} \theta_n = 0 \overset{\text{[Oldenburger, 1940]}}{\iff} \rho(A) < 1 \overset{\text{[Stein, 1952]}}{\iff} \exists P \succcurlyeq 0;\ A^\top P A - P \preccurlyeq 0.$$

The matrix P solution of the Lyapunov inequality $A^\top P A - P \preccurlyeq 0$ can be found analytically using convex optimization [Boyd et al., 1994] and ensures the convergence of the iterative process. Lyapunov theory has been frequently applied to control theory with major works from Lur'e and Postnikov [1944], Popov [1961], Kalman [1963], Yakubovich [1971]. This method can be extended to non-linear processes, by linearizing the iterative procedure [Perron, 1929, Ostrowski, 1966]. It can also be applied directly to optimization [Polyak, 1987] or by using more complex control theory tools [Lessard et al., 2016].

Most stability results for difference equations [Ortega and Rheinboldt, 2000] have been obtained as discrete analogues of corresponding results for differential equations [see, e.g., LaSalle and Lefschetz, 1961]. *Lyapunov's second method* [Hahn, 1958, Kalman and Bertram, 1960] is the most common and general method to prove con-

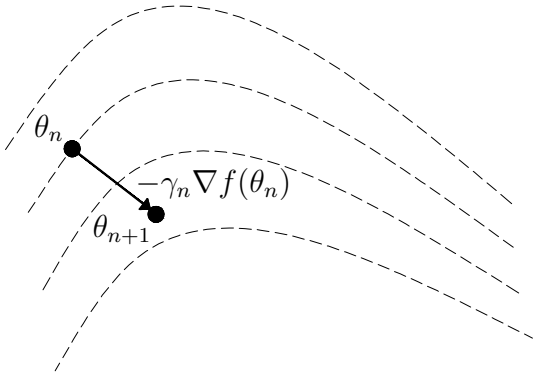


Figure 1-1 – Gradient descent.

vergence of iterative processes. Its idea is to introduce a nonnegative function V (the so-called *Lyapunov function*) which will decrease along the sequence of iterates, i.e., $V(\theta_{n+1}) \leq V(\theta_n)$ and therefore ensure the convergence of the method [see Polyak, 1987, Sec 2.2, for more details]. Therefore finding such a function V may prove the convergence of the iterative system and converse results ensure existence of a Lyapunov function for a stable iterative system [Halanay, 1963, Driver, 1965]. Unfortunately there is no systematic way to find Lyapunov functions in practice. There are some classical solutions such as $f(\theta) - f(\theta_*)$ or $\|\theta - \theta_*\|^2$, but these do not always work. In general, theoreticians have to use their experience to design them.

1.5.2 Smooth Optimization

In this section we present a panel of methods for optimizing a smooth function. We do not aim at being comprehensive and we refer to Nesterov [2004], Bubeck [2015] for surveys. We first present the gradient method dating back to Cauchy [1847] which will be the starting point for more refined first-order techniques.

Gradient descent. With a first-order oracle, the simplest minimization algorithm is to consider the Euler discretization of the continuous *gradient flow* $\dot{\eta} = -\nabla f(\eta)$. It starts from a vector θ_0, and iterates

$$\theta_n = \theta_{n-1} - \gamma_n \nabla f(\theta_{n-1}),$$

for a step-size sequence (γ_n). The gradient step is illustrated in Figure 1-1. An L-smooth function f is quadratically upper-bounded as

$$f(\theta) \leq f(\theta_{n-1}) + \langle \nabla f(\theta_{n-1}), \theta - \theta_{n-1} \rangle + \frac{L}{2} \|\theta - \theta_{n-1}\|^2,$$

and therefore the gradient update is the minimizer of this first-order approximation of f

$$\theta_n = \arg\min_{\theta\in\mathbb{R}^d} \Big\{ \langle \nabla f(\theta_{n-1}), \theta\rangle + \frac{1}{2\gamma_n}\|\theta-\theta_{n-1}\|^2 \Big\}. \tag{1.1}$$

This step is also the steepest descent direction since

$$\arg\min_{\|d\|\leq 1} \langle \nabla f(\theta_{n-1}), d\rangle = -\frac{\nabla f(\theta_{n-1})}{\|\nabla f(\theta_{n-1})\|}.$$

We now present the first convergence result.

Proposition 1. *Let f be a L-smooth convex function. Then gradient descent with $\gamma_n = 1/L$ satisfies*

$$f(\theta_n) - f(\theta_*) \leq \frac{L\|\theta_0-\theta_*\|^2}{n}.$$

Moreover, if f is μ-strongly convex,

$$f(\theta_n) - f(\theta_*) \leq (1-\mu/L)^n [f(\theta_0) - f(\theta_*)].$$

Convergence rates for gradient descent were first studied for quadratic functions by Kantorovitch [1945] and for convex functions by Vainberg [1960], Goldstein [1962], Polyak [1963], Levitin and Polyak [1966]. Proofs of convergence date back to seminal works of Temple [1939], Curry [1944], Crockett and Chernoff [1955]. We sketch the proof for quadratic convex functions since it will be used as a stepping stone when proving more general results later in the manuscript. We denote $f(\theta) = \frac{1}{2}\langle \theta, \Sigma\theta\rangle - \langle b, \theta\rangle$, which attains its global optimum at $\theta_* = \Sigma^{-1}b$. The gradient descent iteration is explicitly

$$\theta_n - \theta_* = \Big(I - \frac{1}{L}\Sigma\Big)(\theta_{n-1}-\theta_*) = \Big(I - \frac{1}{L},\Sigma\Big)^n(\theta_0-\theta_*)$$

and therefore

$$f(\theta_n) - f(\theta_*) = \frac{1}{2}\Big\|\Sigma^{1/2}\Big(I-\frac{1}{L}\Sigma\Big)^n(\theta_0-\theta_*)\Big\|^2 \leq (1-\mu/L)^{2n}[f(\theta_0)-f(\theta_*)],$$

which concludes the proof in the strongly convex case. Otherwise when $\mu = 0$, we use the inequality $\Sigma\big(I-\frac{1}{L}\Sigma\big)^{2n} \preccurlyeq L/nI$.

The gradient descent is adaptive to the problem difficulty. Indeed the step-size does not depend on the value of μ. On the other hand the result is proven for constant step-size, but line-search techniques are also possible [Boyd and Vandenberghe, 2004, Sec 9.3].

Lower-bounds on the convergence rates. After having shown that gradient descent algorithms yield convergence rate $O(\frac{1}{n})$ for L-smooth problems and $O((1-\mu/L)^n)$ when the function is also μ-strongly convex, it is natural to wonder whether these rates are optimal or if other algorithms exist with better convergence rates.

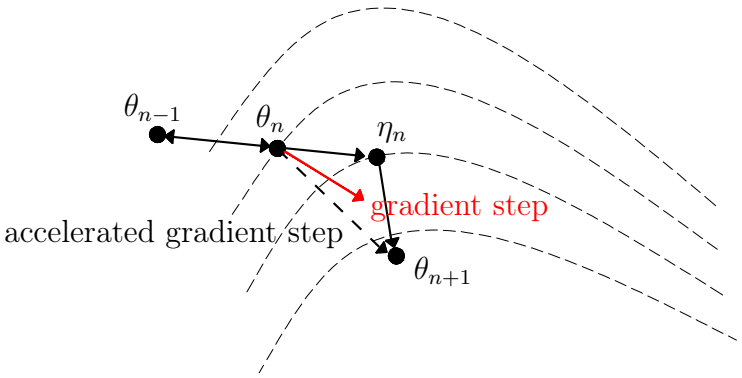


Figure 1-2 – Accelerated gradient descent.

This question was first answered positively by Nemirovski and Yudin [1979] and later presented, in a simplified form, by Nesterov [2004]. Interested readers can also see the recent work of Arjevani and Shamir [2016] on this topic.

Proposition 2. *Let $n \leq (d-1)/2$, $\theta_0 \in \mathbb{R}^d$ and $L > 0$. There exists a L-smooth convex function f with a global minimizer θ_* such that for any first-order method.*

$$f(\theta_n) - f(\theta_*) \geq \frac{3}{32} \frac{L\|\theta_* - \theta_0\|^2}{(n+1)^2}.$$

The number of iterations can not be too large compared to the dimension of the problem. Note this restriction is necessary since the *ellipsoid method* [Ben-Tal and Nemirovski, 2001, Sec. 5.2] converges at exponential rate for n larger than d. Moreover this lower-bound is still informative about the first steps, when n is small compared to d. Surprisingly, the worst-case function designed to prove this result is quadratic. In this sense, quadratic functions are not easier to optimize than smooth functions. Regarding the strongly convex case, Nesterov [2004] provides a lower bound $O\Big(\big(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\big)^n\Big)$ for the oracle complexity.

Clearly, there is a gap between the lower bounds and the performance of gradient descent in Proposition 1. But the lower bounds are not to blame as it turns out they are matched by a slightly more involved algorithm, which we describe now.

Accelerated gradient descent. Taking inspiration from conjugate gradient, Nemirovsky and Yudin [1983] proposed the first method with optimal convergence rates. However this method needed some kind of line-search and was not practical. Nesterov [1983] presented an optimal algorithm which only needs a first-order oracle. Nesterov's *accelerated gradient* can be described as follows: start from $\theta_0 = \eta_0 \in \mathbb{R}^d$ and then iterate

$$\begin{aligned} \theta_n &= \eta_{n-1} - \frac{1}{L}\nabla f(\eta_{n-1}) \\ \eta_n &= \theta_n + \delta_n(\theta_n - \theta_{n-1}), \end{aligned}$$

where the momentum coefficient $\delta_n \in \mathbb{R}$ is chosen to accelerate the convergence rate and has its roots in the heavy-ball algorithm from Polyak [1964]. Figure 1-2 illustrates the difference between gradient and accelerated gradient steps. For this algorithm one obtains the following result.

Proposition 3. *Let f be a L-smooth convex function. Then for $\delta_n = \frac{n-1}{n-2}$,*

$$f(\theta_n) - f(\theta_*) \leq \frac{2L\|\theta_0 - \theta_*\|^2}{(n+1)^2}.$$

Moreover let f be also μ strongly convex. Then for $\delta_n = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$,

$$f(\theta_n) - f(\theta_*) \leq L\|\theta_0 - \theta_*\|^2\Big(1 - \sqrt{\frac{\mu}{L}}\Big)^n.$$

The algorithm was initially proposed by Nesterov [1983, 2004] with a more complex momentum term. The simplified version we present here is due to Tseng [2008]. Beck and Teboulle [2009], Schmidt et al. [2011] both proposed very concise proofs of these optimal convergence rates. It is worth noting that the accelerated gradient method is not adaptive to strong convexity. Indeed the value of μ is needed to obtain the linear convergence rate. Nesterov [2013], O'Donoghue and Candès [2013] proposed the use of restart heuristics to resolve this issue. Indeed Arjevani and Shamir [2016] show that no methods with fixed sequences of step-sizes can achieve the accelerated rate, unless the parameter μ is known. This method is not a descent algorithm and often exhibits an oscillatory behavior. This will be further detailed in Chapter 2. Until recently, accelerated gradient descend lacked an intuitive interpretation, contrary to the standard gradient descent. This has been addressed by Allen-Zhu and Orecchia [2017], Bubeck et al. [2015], and in a series of works on the connection with differential equations by Su et al. [2014], Krichene et al. [2015], Wibisono et al. [2016], Wilson et al. [2016], Attouch et al. [2016a], Attouch and Peypouquet [2016].

1.5.3 Specificity of Quadratic Functions

Minimizing a quadratic function f is equivalent to solving a linear system of equations, i.e., $\nabla f(\theta) = 0$. For instance, in least-squares regression, the ERM predictor is directly given by the normal equation

$$\Phi^\top \Phi \hat{\theta} = \Phi \vec{Y}, \tag{1.2}$$

which is a set of d equations for d unknowns. Hence, the methods we presented so far may appear inefficient or needlessly complicated for solving such problems. Nevertheless the worst case function used to prove the lower bound in Proposition 2 is quadratic and accordingly quadratic problems are just as hard as any smooth problem. We introduce presently two methods dedicated to minimizing quadratic functions.

Numerical algebra. First one may prefer to directly solve the system in Eq. (1.2) by using numerical linear algebra algorithms. The complexities of these methods can usually be decomposed as $O(d^3+nd^2)$ to form $\Phi^\top\Phi$, $O(nd^2)$ to form $\Phi\vec{Y}$ and $O(d^3)$ to solve the linear system. Gaussian elimination could be used to solve the linear system in Eq. (1.2) as any linear system but this approach is oblivious to the specific form of the empirical covariance $\Phi^\top\Phi$. One may thus prefer methods based on singular value decomposition or QR decomposition when Φ is full rank [see, e.g., Golub and Van Loan, 2013]. These methods rely on well-known linear algebra, but can only be used for small to medium scale problems (typically $d \leq 10^4$). For large scale problems they are not suitable since they do not leverage the special structure of the problem and the *conjugate gradient* algorithm may be favored instead.

Conjugate gradient algorithm. Hestenes and Stiefel [1952] introduced the *conjugate gradient* method which corresponds to the *momentum* algorithm

$$\theta_n = \theta_{n-1} - \gamma_n \nabla f(\theta_{n-1}) + \delta_n(\theta_{n-1} - \theta_{n-2})$$

with (γ_n, δ_n) solutions of the two dimensional optimization problem:

$$\min_{\gamma,\delta} f\big(\theta_{n-1} - \gamma\nabla f(\theta_{n-1}) + \delta(\theta_{n-1} - \theta_{n-2})\big).$$

This line search procedure can be explicitly solved when f is a quadratic function with the same computational complexity as a gradient computation. The explicit method is detailed by Golub and Van Loan [2013]. Surprisingly this method converges to the solution θ_* of the problem in at most d steps and the behavior of the method is globally controlled by [Polyak, 1987]

$$f(\theta_n) - f(\theta_*) \leq \min\Big\{\frac{L\|\theta_0 - \theta_*\|_2^2}{8n^2}, 4\exp\left(-2\sqrt{\mu/L}n\right)[f(\theta_0) - f(\theta_*)]\Big\}.$$

We note that the method converges at the optimal linear rate without knowing the strong convexity constant μ, in contrast to the accelerated gradient method. Nonlinear extensions of the conjugate gradient are presented by Nocedal and Wright [2006].

1.5.4 Extension to Composite Optimization

When the function f is no longer smooth, gradient descent with constant step size fails be to consistent [Bertsekas, 1999]. Shor [1962], Ermoliev [1966], Polyak [1967] solve this issue by showing that the projected sub-gradient method with decreasing step size $O(1/\sqrt{n})$ converges at rate $O(1/\sqrt{n})$ when the average of the iterates is considered as output of the method [Shor et al., 1985]. When the function is also μ-strongly convex, the projected sub-gradient method with step-size $O(1/\mu n)$ converges at rate $O(\log(n)/\mu n)$ and at rate $O(1/\mu n)$ when non-uniform average is used. Moreover Nemirovsky and Yudin [1983] show these rates are optimal among first-order methods.

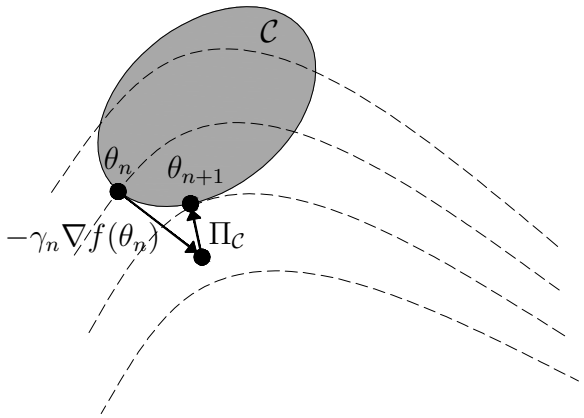


Figure 1-3 – Projected subgradient. We denote by $\Pi_{\mathcal{C}}$ the Euclidean projection on the convex set $\mathcal{C}$.

Yet there is a crucial difference between the class of smooth problems (with bounded smoothness constant) and the class of all convex problems since convergence rates are downgraded from $O(1/n^2)$ to $O(1/\sqrt{n})$. However the structure of the function as well as the reason for non-smoothness are often known. For instance the objective function is sometimes the sum of a smooth and a non-smooth function, as in Lasso or constrained problems. These are *composite problems* of the form:

$$\min_{\theta \in \mathbb{R}^d} f(\theta) + g(\theta),$$

where f is an L-smooth function accessible through a first-order oracle and g is a simple convex function in a sense that will be explained below.

As with the gradient descent for smooth functions, at each iteration the smooth function f is linearized around the current iterate θ_n. Then *proximal gradient* methods, also called forward-backward splitting methods [see, e.g., Beck and Teboulle, 2009, Wright et al., 2009, Combettes and Pesquet, 2011] consider the iterate

$$\theta_{n+1} = \arg\min_{\theta \in \mathbb{R}^d} \Bigg\{ \underbrace{f(\theta_n) + \langle \nabla f(\theta_n), \theta - \theta_n \rangle + \frac{L}{2}\|\theta - \theta_n\|_2^2}_{\text{Linearization of the smooth component}} + \underbrace{g(\theta)}_{\text{Non-smooth term}} \Bigg\}.$$

For instance when $g = \mathbb{1}_{\mathcal{C}}$ is the indicator function of a convex set $\mathcal{C}$, proximal gradient is tantamount to projected gradient as depicted in Figure 1-3. This update may be efficiently written in the terms of proximal operator defined by Moreau [1962] as $\mathrm{Prox}_g(\eta) = \arg\min_{\theta \in \mathbb{R}^d} \{\frac{1}{2}\|\theta - \eta\|_2^2 + g(\theta)\}$:

$$\theta_{n+1} = \arg\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{L} g(\theta) + \frac{1}{2} \left\| \theta - \Big(\theta_n - \frac{1}{L}\nabla f(\theta_n)\Big) \right\|_2^2 \right\} = \mathrm{Prox}_{\frac{1}{L}g}\Big(\theta_n - \frac{1}{L}\nabla f(\theta_n)\Big).$$

Therefore this method is only used in practice with simple functions g whose proximal

operator is computable effectively. This is the case if there exists a closed form expression and several examples are provided by Bach et al. [2012].

Surprisingly Beck and Teboulle [2009], Nesterov [2013] show convergence results similar to smooth optimization with the proofs following the same general guidelines.

1.5.5 Extension to Non-Euclidean Geometry

Until now, we have derived dimension-free convergence rates under Lipschitz assumptions on the function or on its gradient with respect to the Euclidean geometry. When these assumptions are with respect to different norms, the above convergence rates still apply but the dimension d of the space appears in the bound. For example, if the gradient of a function f is bounded in the ℓ_∞-norm by a constant $B > 0$, it is bounded in the ℓ_2-norm as $\|\nabla f(\theta)\|_2 \leq B\sqrt{d}$. Thus algorithms which work with different geometries may yield convergence rates with better dependency on the dimension.

Let $h : \mathbb{R}^d \to \mathbb{R}$ be a differentiable convex function such that its gradient ∇h is a bijection of $\mathbb{R}^d$. The *Bregman divergence* associated with the function h is defined as

$$D_h(\theta_1, \theta_2) = h(\theta_1) - h(\theta_2) - \langle \nabla h(\theta_2), \theta_1 - \theta_2 \rangle, \quad \theta_1, \theta_2 \in \mathbb{R}^d.$$

It may be interpreted as a generalized squared distance which provides a new geometry standing in for the Euclidean one. Incidentally, the latter is recovered by considering $h(\cdot) = \frac{1}{2}\|\cdot\|_2^2$.

With the aim of solving the constrained problem $\min_{\theta \in \mathcal{C}} f(\theta)$ under the Bregman geometry, Nemirovski and Yudin [1979] introduce the *mirror descent* algorithm which may be defined by the update

$$\theta_n = \operatorname*{arg\,min}_{\theta \in \mathcal{C}} \big\{ \gamma_n \langle \nabla f(\theta_{n-1}), \theta \rangle + D_h(\theta, \theta_{n-1}) \big\},$$

started from $\theta_0 \in \mathcal{C}$. The Bregman divergence has replaced the Euclidean distance in the definition of the gradient descent iteration in Eq. (1.1). This contemporary proximal definition is due to Beck and Teboulle [2003]. Mirror descent can be equivalently written as a *greedy gradient update* followed by a *projection* step

$$\begin{aligned} \nabla h(\phi_n) &= \nabla h(\theta_{n-1}) - \gamma_n \nabla f(\theta_{n-1}) && (\textit{greedy update}) \\ \theta_n &= \operatorname*{arg\,min}_{\theta \in \mathcal{C}} D_h(\theta, \phi_n). && (\textit{projection step}) \end{aligned}$$

Let us take a step backward and forget about the Euclidean structure. The gradient $\nabla f(\theta_{n-1})$ belongs to the *dual space* whereas the iterate θ_{n-1} lives in the *primal space*. Thus the gradient iteration no longer makes sense. Instead Nemirovski and Yudin [1979] propose first to map the current iterate θ_{n-1} to the dual space with $\nabla h(\theta_{n-1})$, and perform there the gradient update $\nabla h(\theta_{n-1}) - \gamma_n \nabla f(\theta_{n-1})$. The resulting point is then mapped back in the primal space and projected with regards to the Bregman divergence in the convex set $\mathcal{C}$. This is depicted in Figure 1-4.

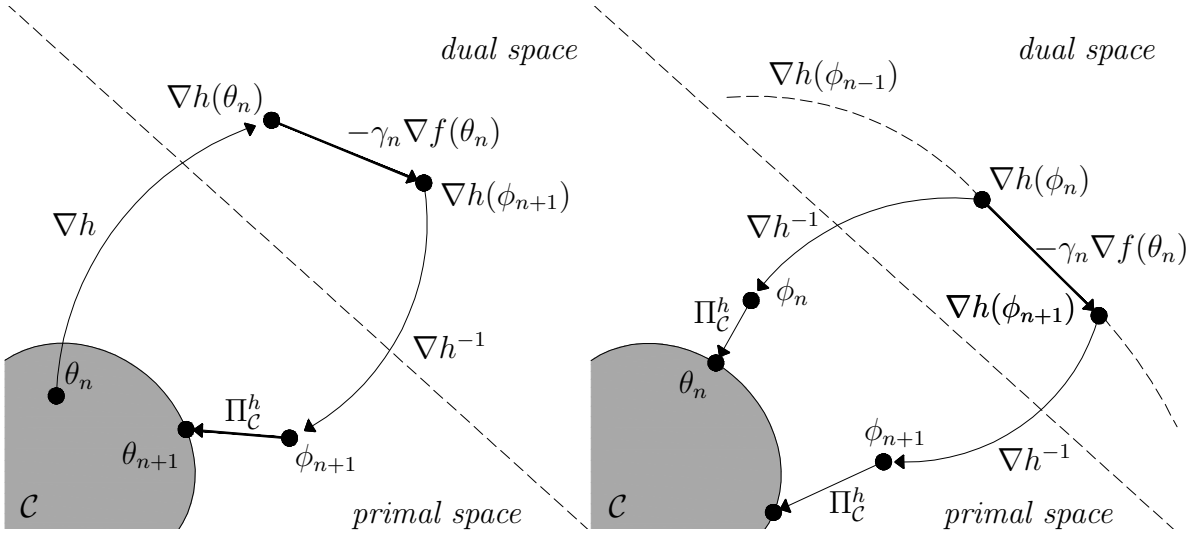


Figure 1-4 – Mirror descent versus dual averaging. Left: mirror descent. Right: dual averaging. We denote by $\Pi_{\mathcal{C}}^h$ the Bregman projection on the convex set $\mathcal{C}$ defined in the *projection step*.

Mirror descent has to be associated with its cousin *dual averaging* initially introduced by Nesterov [2009] which considers a *lazy gradient update* started from $\theta_0, \phi_0 \in \mathcal{C}$

$$\begin{aligned} &\nabla h(\phi_n) = \nabla h(\phi_{n-1}) - \gamma_n \nabla f(\theta_{n-1}) && (\textit{lazy update}) \\ &\theta_n = \operatorname*{arg\,min}_{\theta \in \mathcal{C}} D_h(\theta, \phi_n). && (\textit{projection step}) \end{aligned}$$

Taking ϕ_0 such that $\nabla h(\phi_0) = 0$, dual averaging can also be written under the following simpler expression (still started from $\theta_0 \in \mathcal{C}$)

$$\theta_n = \operatorname*{arg\,min}_{\theta \in \mathcal{C}} \Big\{ \big\langle \sum_{i=1}^{n-1} \gamma_{i+1} \nabla f(\theta_i), \theta \big\rangle + h(\theta) \Big\}.$$

The greedy and lazy aspects of these methods as well as what differentiates them will be explained in detail in Chapter 4. As initially desired, mirror descent and dual averaging algorithms adapt to the geometry provided by the Bregman divergence. When the function f is assumed to be B-Lipschitz with respect to an arbitrary norm $\|\cdot\|$ and the mirror map h is μ-strongly convex with respect to the same norm, then these methods yield an optimal convergence $O(BD/\sqrt{\mu n})$ where $D^2 = \sup_{\theta_1, \theta_2 \in \mathcal{C}} \{h(\theta_1) - h(\theta_2)\}$. Nesterov [2007], Krichene et al. [2015] have accelerated these algorithms when the function is smooth with respect to some non-Euclidean geometry. Bauschke et al. [2016] have recently redefined the concept of smoothness for non-Euclidean geometry. Furthermore these methods have also been extended to the composite setting by Duchi et al. [2010], Xiao [2010].

	Convex	**μ-Strongly convex**
B-Lipschitz		
Projected sub-gradient	$BD/\sqrt{n}$	$B^2D/\mu n$
L-Smooth		
Accelerated gradient	$L\|\theta_0 - \theta_*\|^2/n^2$	$(1 - \sqrt{L/\mu})^n\|\theta_0 - \theta_*\|^2$
Quadratic		
Conjugate gradient	$\min\{d, L\|\theta_0 - \theta_*\|^2/n^2\}$	$\min\{d, (1 - \sqrt{L/\mu})^n\|\theta_0 - \theta_*\|^2\}$

Table 1.1 – Convergence rates for deterministic optimization.

1.5.6 Conclusion

We summarize the different rates of convergence provided so far in Table 1.1. Our goal is to minimize the generalization error of a machine learning problem. But we have seen that the estimation error was $O(1/\sqrt{n})$ in the Lipschitz case and $O(1/n)$ in the strongly convex case. This is the key insight of Bottou and Bousquet [2008], Shalev-Shwartz and Srebro [2008] who explain there is no need to optimize below estimation error for machine learning applications. We present now a different approach which directly minimizes the generalization error.

1.6 Stochastic Approximation

1.6.1 Classical Stochastic Approximation

Stochastic approximation historically aims at finding a zero of a function $h : \mathbb{R}^d \to \mathbb{R}^d$ which cannot be directly computed but is accessible through samples. For instance in the dosage of chemical products, experiments make it possible to sample the function h at certain points: the result of the experiment is the true value perturbed by some random noise $h(\theta) + \varepsilon$.

Traditional deterministic techniques may be used to solve the non-linear system $h(\theta) = 0$. At each iteration, an estimate of the value of $h(\theta_n)$ is built, for example, by approximating it by its means over the samples. This is the *sample average approximation* approach by Kleywegt et al. [2002]; see, e.g., Shapiro et al. [2014] for a review. This technique is unstable and time-consuming because at each iteration enough samples need to be computed to properly approximate the function value. Instead Robbins and Monro [1951] propose to use each sample once by considering the noisy fixed point iteration:

$$\theta_n = \theta_{n-1} - \gamma_n[h(\theta_n) + \varepsilon_n],$$

where ε_n is typically assumed to be an i.i.d. sequence of zero-mean finite variance noise and (γ_n) is a sequence of step-size. Convergence results are usually asymptotic and assume the existence of a well behaving Lyapunov function [Polyak, 1976, Duflo, 1997]. Ermoliev [1969], Robbins and Siegmund [1971] show the almost-surely convergence of the iterates and Chung [1954], Sacks [1958], Fabian [1968] show their asymptotic

normality, both if the step-size satisfies the conditions

$$\sum_{i=1}^{\infty} \gamma_i = \infty \quad \text{and} \quad \sum_{i=1}^{\infty} \gamma_i^2 < \infty.$$

These results directly extend to minimization of a function f by finding a zero of its gradient ∇f (we note that no convexity is assumed).

Therefore the goal of *stochastic approximation* can be equivalently defined as minimizing a function f given only unbiased estimates $\nabla f_{n+1}(\theta_n)$ of its gradient $\nabla f(\theta_n)$ at certain points θ_n. Indeed in a lot of applications, the exact gradient of a function is not tractable because of errors in the measurements or in the Monte Carlo evaluation of expected values. Yet noisy unbiased estimates of the gradient are accessible and cheap. As noticed by Bottou and Le Cun [2005] this directly includes the usual machine learning situation of minimization of the generalization error

$$L(\theta) = \mathbb{E} f_n(\theta) \quad \text{for} \quad f_n(\theta) = l(y_n, \langle \theta, x_n \rangle),$$

observing estimate $\nabla f_n(\theta) = l'(y_n, \langle \theta, x_n \rangle) x_n$ of the true gradient $\nabla L(\theta) = \mathbb{E} \nabla f_n(\theta)$. But its applicability is much broader and goes far beyond convex optimization [Benveniste et al., 1990, Kushner and Yin, 2003].

1.6.2 Convex Stochastic Optimization

When we want to minimize a function which is only available through unbiased estimates of the function values or its gradients, *stochastic gradient descent* is the key algorithm. It takes the form

$$\theta_n = \theta_{n-1} - \gamma_n \nabla f_n(\theta_{n-1}),$$

where γ_n is a step-size sequence chosen depending on the situation. Here we consider the classical stochastic approximation framework [Kushner and Yin, 2003, Borkar, 2008]. That is, we let $(\mathcal{F}_n)_{n \geq 0}$ be an increasing family of σ-fields such that for each $\theta \in \mathbb{R}^d$ and for all $n \geq 1$ the random variable $\nabla f_n(\theta)$ is square-integrable and $\mathcal{F}_n$-measurable with $\mathbb{E}[\nabla f_n(\theta) | \mathcal{F}_{n-1}] = \nabla f(\theta)$. Polyak [1990], Ruppert [1988] consider instead the average of the iterates $\bar{\theta}_n = \frac{1}{n} \sum_{i=1}^{n-1} \theta_i$ as the output of the algorithm. This is the *averaged stochastic gradient descent*.

In machine learning applications, when the objective function is the expectation of a loss function, the running-time complexity of the stochastic gradient is $O(nd)$ after n iterations. This has to be compared with the $O(n^2 d)$ complexity of gradient descent on the training error computed with n data (computing a gradient which is an average of n terms costs $O(nd)$).

The global minimax rates of convergence are known since Nemirovsky and Yudin [1983] for convex non-smooth problems and Tsypkin and Polyak [1974], Nazin [1989] for strongly convex non-smooth problems. For convex problems the minimax rate is $O(1/\sqrt{n})$ and it is attained by averaged stochastic gradient descent with step-size $\gamma_n =$

$C/\sqrt{n}$ [Zhang, 2004]. When in addition, the function is assumed μ-strongly convex, the minimax rate becomes $O(1/(\mu n))$ and is still attained by averaged stochastic gradient but with a smaller step-size $\gamma_n = C/\mu n$ and a non-uniform average [Rakhlin et al., 2012, Lacoste-Julien et al., 2012]. Yet the minimax rate is the same for non-smooth functions as it is in deterministic optimization. Moreover the proofs in the two settings are very similar. The proof techniques are different for the lower-bounds and use tools from statistics and information theory. They are clearly presented and extended by Agarwal et al. [2012], Raginsky and Rakhlin [2011].

The stochastic gradient algorithm has been extended to the proximal case by Duchi and Singer [2009], Hu et al. [2009], Xiao [2010] and to the non-Euclidean setting by Nemirovski et al. [2009], Lan [2012].

1.6.3 Smooth and Strongly Convex Stochastic Optimization

We have seen that smoothness plays a central role in the context of deterministic optimization but for stochastic optimization, smoothness only leads to improvements on constants, not on the rate itself. The minimax rate remains $O(1/\sqrt{n})$ for non-strongly convex problems. Moreover accelerated gradient descent is notorious for not being robust to random or deterministic noise in the gradient [d'Aspremont, 2008, Schmidt et al., 2011, Devolder et al., 2014] and Hu et al. [2009], Lan [2012] advocate small step-size $C/n^{3/2}$ to obtain convergence rate $O(1/\sqrt{n})$.

The story is different for strongly convex problems. When the function is non-smooth, stochastic gradient descent is not adaptive to strong convexity since the step-size has to depend on μ in order to obtain fast rates. Moreover the choice of the constant in the step-size $C/\mu n$ is very sensitive as noted by Nemirovski et al. [2009].

Yet for smooth and strongly convex functions, Polyak and Juditsky [1992] show that averaged stochastic gradient with larger step size $\gamma_n = Cn^{-\alpha}$ for $\alpha \in (1/2, 1)$ has a convergence rate of order $O(1/n)$ independent of the strong convexity constant. This idea is developed by Bach and Moulines [2011] who provide a non-asymptotic analysis of averaging for smooth functions. They show that averaged stochastic gradient with step-size $\gamma_n = C/n^\alpha$ has a rate of convergence $O(1/n + 1/(\mu n)^2)$ for μ-strongly convex problems and $O(1/n^{1-\alpha})$ for convex problems. Therefore this method with step-size $\gamma_n = C/\sqrt{n}$ is adaptive to strong convexity with convergence rate $O(\min(1/\mu n), 1/\sqrt{n})$ for all smooth problems. Bach [2014] extends these results to logistic regression which is not globally strongly convex. We also note than the benefits of averaging may be understood through the lens of two-time scale stochastic approximation [Borkar, 1997].

Typical high-dimensional machine learning problems have very correlated variables so that the strong convexity constant μ may be arbitrarily small. This makes convergence rates $O(1/(\mu n))$ useless and makes non-strongly convex scenario more suitable. Thus we aim at obtaining algorithms with convergence rate $O(1/n)$ that are robust to arbitrarily small strong convexity constants.

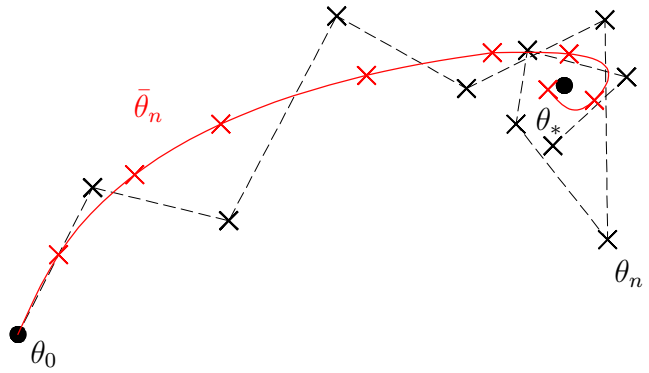


Figure 1-5 – Markov chain interpretation.

1.6.4 Least-Mean-Squares Algorithm

When applied to least-squares $f(\theta) = \frac{1}{2}\mathbb{E}[(y_n - \langle \phi(x_n), \theta \rangle)^2]$, stochastic gradient descent is known as the least-mean-square algorithm [Macchi, 1995]. In signal processing theory, it is usually studied (a) without averaging, (b) with decreasing step-size and (c) with strong convexity assumption $\Sigma = \mathbb{E}[\phi(X) \otimes \phi(X)] \succcurlyeq \mu I$. Bach and Moulines [2013], following Györfi and Walk [1996], propose a non-asymptotic analysis of the averaged least-mean-square algorithm with constant-step size

$$\theta_n = \theta_{n-1} - \gamma(\langle \phi(x_n), \theta_{n-1} \rangle - y_n)\phi(x_n). \tag{1.3}$$

The sequence (θ_n) is a *homogeneous Markov chain* [Meyn and Tweedie, 2009] which, under appropriate conditions, converges in distribution to its unique stationary distribution π_γ. The average of the stationary distribution is denoted by $\bar{\theta}_\gamma = \int \theta \pi_\gamma(d\theta)$. Taking the expectation in Eq. (1.3) and remembering that $\theta_* = \Sigma^{-1}\mathbb{E}[\phi(x)y]$ one obtains

$$\mathbb{E}\theta_n = \mathbb{E}\theta_{n-1} - \gamma\Sigma(\mathbb{E}\theta_{n-1} - \theta_*).$$

Recall that $\lim_{n\to\infty} \mathbb{E}\theta_n = \bar{\theta}_\gamma$ and Σ is invertible. It follows that $\bar{\theta}_\gamma = \theta_*$. The ergodic theorem then shows that $\bar{\theta}_n$ converges almost surely to $\bar{\theta}_\gamma$. The central limit theorem for Markov chains also implies that $n\mathbb{E}\|\bar{\theta}_n - \theta_*\|_\Sigma^2$ has a finite limit. Therefore, the Markov chain interpretation explains the averaged iterates converge to θ_* pointwise, and that the rate of convergence of $\mathbb{E}[f(\theta_n) - f(\theta_*)]$ is of order $O(1/n)$, whereas θ_n does not converge to θ_* but oscillates around it. This is summarized in Figure 1-5.

In addition Bach and Moulines [2013] show that under similar assumptions to that of the analysis of least-squares with random design and without assuming strong convexity,

$$\mathbb{E}f(\bar{\theta}_n) - f(\theta_*) \leq \frac{4R^2\|\theta_0 - \theta_*\|^2}{n} + \frac{4\sigma^2 d}{n}.$$

The performance of this algorithm is then decomposed as the sum of: (a) a bias term which depends on the deviation between the initial point of the algorithm and the solution of the problem and characterizes how fast initial conditions are forgotten and

(b) a variance term, which depends on the covariance of the noise in the gradients and describes the effect of this noise.

We have already seen that the variance term $O(\sigma^2 d/n)$ is optimal over all estimators. However the bias term $O(1/n)$ is suboptimal since the optimal optimization algorithm forgets the initial condition as $O(1/n^2)$. Therefore there is a mismatch between the optimization and the statistics theories which will be investigated to a great extent in Chapter 3. This is relevant in practice since the bias term can be significantly larger than the variance term. It may consequently be useful to accelerate its convergence.

1.6.5 Finite Sum Problems

We end this section with an intermediate problem, related both to deterministic optimization and to stochastic approximation, which has gained a lot of attention these past years. We consider here a finite sum problem of the form:

$$f(\theta) = \frac{1}{k} \sum_{i=1}^{k} f_i(\theta),$$

where $f_1, \dots, f_k$ are L-smooth-functions, possibly μ-strongly convex. An important example is the empirical risk studied in Section 1.4.

The function f is deterministic, thus contrary to stochastic approximation we have a full access to its gradient. The function f can therefore be optimized with gradient descent:

$$\theta_n = \theta_{n-1} - \frac{\gamma_{n-1}}{k} \sum_{i=1}^{k} \nabla f_i(\theta_{n-1})$$

with a linear convergence rate $O(\exp(-n\mu/L))$. However the complexity of computing $\nabla f(\theta_{n-1})$ is $O(dk)$ at each iteration. The running time complexity is then $O(dkL/\mu \log(1/\varepsilon))$ and can be decreased to $O(dk\sqrt{L/\mu} \log(1/\varepsilon))$ using accelerated gradient descent.

Stochastic gradient descent can also be applied to f, sampling with replacement, at each iteration, $i(n) \in \{1, \dots, n\}$ (see recent work by Gürbüzbalaban et al. [2015], Shamir [2016] for results on sampling without replacement) and iterating

$$\theta_n = \theta_{n-1} - \gamma_{n-1} \nabla f_{i(n)}(\theta_{n-1}).$$

The convergence rate $O(L/(\mu n))$ is slower but the running time complexity $O(Ld/(\mu\varepsilon))$ is independent of k. So stochastic gradient descent achieves rapidly low accuracy whereas deterministic algorithm should be preferred when high accuracy is desired.

Schmidt et al. [2017], Shalev-Shwartz and Zhang [2013] propose two new methods, respectively stochastic averaged gradient (SAG) and stochastic dual coordinate ascent (SDCA) which both combine linear convergence and $O(d)$ complexity by iteration. These methods have a running time complexity $O(d(k + L/\mu) \log(1/\varepsilon))$. Initially, SDCA applied only to a restricted set of problems (minimization ℓ_2-regularized train-

	Convex	**μ-Strongly convex**
B-Lipschitz		
Averaged gradient descent	$BD/\sqrt{n}$	$B^2D/\mu n$
L-Smooth		
Averaged gradient descent	$O(1/\sqrt{n})$	$O(1/\mu n)$
Least-squares		
Averaged gradient descent	$O(\sigma^2 d/n + R^2\|\theta_*\|_2^2)$	[Jain et al., 2016]
Finite sum		
Accelerated SDCA, SVRG	$O\Big(\exp\big(-\frac{n}{k}\big) + \frac{\sqrt{Lk}}{n^2}\Big)$	$O\Big(\exp\big(-\frac{n}{k+\sqrt{Lk/\mu}}\big)\Big)$

Table 1.2 – Convergence rates for stochastic approximation.

ing error with linear predictions) and required duality but these two assumptions have been recently lifted by Shalev-Shwartz [2016a]. Johnson and Zhang [2013] obtain the same running time complexity through a different method that can be understood in terms of variance reduction.. These methods do not contradict lower bounds seen in the previous section since they hold due to the special structure of the objective function.

Shalev-Shwartz and Zhang [2014], Zhang and Xiao [2015], Nitanda [2014], Lin et al. [2015] obtain accelerated versions of these algorithms with running time complexities $O(d(k + \sqrt{kL/\mu})\log(1/\varepsilon))$ and $O(k\log 1/\varepsilon) + \sqrt{kL/\varepsilon})$ when the function is not strongly convex. They are shown to be optimal by Agarwal and Bottou [2015], Lan and Zhou [2015], Arjevani and Shamir [2016], Woodworth and Srebro [2016].

1.6.6 Conclusion

The different rates of convergence for stochastic convex optimization are summarized in Table 1.2. Stochastic approximation techniques enable to directly optimize the generalization error and provide predictors with optimal estimation error. Moreover their estimation rates are much simpler to prove than analyzing (a) the statistical performance of the minimizer of the empirical risk as in Section 1.4 and (b) the convergence rate of the optimization method (studied in Section 1.5).

Shalev-Shwartz [2016b] points out that the minimizer of the empirical risk may have better statistical properties than the predictor obtained with stochastic gradient descent even if they have an estimation error of the same order. In particular this latter seems to be more efficient to predict rare events. Therefore this fact emphasizes the importance to efficiently optimize finite sum objective functions with dedicated methods (SAG, SDCA, SVRG) with the intended target of rapidly minimizing the empirical risk [Frostig et al., 2015].

Stochastic approximation methods have the significant advantage of being online and new data are instantly taken into account. Yet the i.i.d. assumption on the streamed data is quite strong. We will explain now how *online learning* lifts the assumption that the sequence being observed is independently sampled from an unknown distribution.

1.7 Online Convex Optimization

Online learning studies sequential decision processes. It may be viewed as an extension of statistical learning to a sequential setting without independence of the data. The setting is even harder: an adversary maliciously and adaptively picks the data in order to confuse the learner. A comprehensive survey on this very active field, closely related to *game theory*, was written by Cesa-Bianchi and Lugosi [2006]. We present here the *online convex optimization* framework defined by Zinkevich [2003] following Gordon [1999]: let $\mathcal{C}$ be a non empty convex set. At each iteration n the learner

— predicts a vector θ_n

— observes a convex loss function $f_n : \mathcal{C} \to \mathbb{R}$

— suffers the loss $f_n(\theta_n)$.

He wants to minimize his regret:

$$R_n = \sum_{i=1}^{n} f_i(\theta_i) - \inf_{\theta \in \mathcal{C}} \sum_{i=1}^{n} f_i(\theta).$$

Following Shalev-Shwartz [2012], Hazan [2012], only linear functions $f_n(\theta) = \langle x_n, \theta \rangle$ for $x_n \in \mathbb{R}^d$ are considered here. Indeed results on linear functions are straightforwardly lifted to general convex functions, using the inequality $f_n(\theta_n) - f_n(\theta) \leq \langle \nabla f_n(\theta_n), \theta_n - \theta \rangle$ for all $\theta \in \mathcal{C}$ and considering the linear proxy $\hat{f}_n(\theta) = \langle \nabla f_n(\theta_n), \theta \rangle$.

Hannan [1957] initially proposes to predict $\theta_n = \arg\min_{\theta \in \mathcal{C}} \sum_{i=1}^{n-1} f_i(\theta)$, the best vector so far. However Kalai and Vempala [2005] show that the regret R_n of this simple strategy (they name *follow the leader*) can sometimes be $O(n)$. Following the time-varying potential method presented by Cesa-Bianchi and Lugosi [2006, Sec. 11.6], Shalev-Shwartz and Singer [2007] and Abernethy et al. [2008] propose to add a regularization $h : \mathcal{C} \to \mathbb{R}$ to the previous approach. The *follow the regularized leader* (FTRL) algorithm considers at each iteration n the update

$$\theta_n = \arg\min_{\theta \in \mathcal{C}} \sum_{i=1}^{n-1} f_i(\theta) + h(\theta).$$

It may be directly interpreted as the online extension of the dual averaging algorithm presented in Section 1.5.5. Nevertheless the primal-dual approach of Shalev-Shwartz and Singer [2007] is more general and may lead to different algorithms. We also note FTRL recovers for special choices of constraint set $\mathcal{C}$ and regularization h the *online gradient descent* by Zinkevich [2003] and the *exponentiated gradient* by Kivinen and Warmuth [1997].

We can also take a different iterative approach and consider the online extensions of mirror descent and dual averaging algorithms introduced Section 1.5.5. In online learning, they are respectively denoted by *greedy* and *lazy* online mirror descent. As their deterministic counterparts, these two methods consider two sequences (ϕ_n) and

(θ_n) and respectively update:

$$\nabla h(\phi_n) = \nabla h(\theta_{n-1}) - \eta f_{n-1} \qquad (\textit{greedy update})$$
$$\nabla h(\phi_n) = \nabla h(\phi_{n-1}) - \eta f_{n-1}, \qquad (\textit{lazy update})$$

Then θ_n is obtained by projecting according to the Bregman divergence

$$\theta_n = \arg\min_{\theta \in \mathcal{C}} D_h(\theta, \phi_n).$$

In the linear case, the lazy online mirror descent is equivalent to the FTRL approach as remarked by Hazan and Kale [2010].

Under strong convexity assumptions on the regularization h, and B-Lipschitz assumptions on the functions f_n, Shalev-Shwartz [2007], Abernethy et al. [2008] show that the regret of FTRL is $R_n = O(B\sqrt{n})$. Moreover when the loss functions f_n are also μ-strongly convex, Hazan et al. [2007], Shalev-Shwartz and Kakade [2009] show the regret becomes logarithmic $R_n = O(B^2 \log n/\mu)$. These rates are known to be optimal [Takimoto and Warmuth, 2000, Shalev-Shwartz, 2007].

Furthermore Littlestone [1989] explains online learning results can be converted into statistical learning results by considering the average $\bar{\theta}_n = \frac{1}{n}\sum_{i=0}^{n-1} \theta_i$ as output of the learning algorithm. This has been strengthened by Cesa-Bianchi et al. [2004], Zhang [2005] to results on high probability.

Application to online linear regression. We illustrate this setting on *online linear regression* where we observe sequentially deterministc data $(x_1, y_1, \ldots, x_n, y_n) \in (\mathcal{X} \times \mathcal{Y})^n$ and want to design a predictor which achieves small regret

$$R_n = \sum_{i=1}^{n} (\langle x_i, \theta_i \rangle - y_i)^2 - \inf_{\theta \in \mathcal{C}} \sum_{i=1}^{n} (\langle x_i, \theta \rangle - y_i)^2.$$

This is the online learning setting applied to the least-squares functions $f_n(\theta) = (\langle x_n, \theta \rangle - y_n)^2$. The FTRL approach may be used with an ℓ_2-regularization $h(\theta) = \frac{1}{2}\|\theta\|_2^2$ to obtain a predictor

$$\theta_n = \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n-1} (\langle \theta, x_i \rangle - y_i)^2 + \|\theta\|^2$$

similar to the one of the *ridge regression* [Hoerl and Kennard, 1970]. Its regret is not generally bounded, though [Cesa-Bianchi and Lugosi, 2006, Theorem 11.7].

Vovk [2001], Azoury and Warmuth [2001] solve this issue making the predictor non linear. The *Vovk-Azoury-Warmuth* predictor is the non linear predictor $h_n(x) = \langle \theta_n^x, x \rangle$ where θ_n^x is given by

$$\theta_n^x = \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n-1} (\langle \theta, x_i \rangle - y_i)^2 + \|\theta\|^2 + \langle \theta, x \rangle^2.$$

The additional term (compared with the ridge regression predictor) $\langle \theta, x\rangle^2$ may be interpreted as the loss being incurred at time n when y_n is estimated by 0. $\theta_n^{x_n}$ has the following close form:

$$\theta_n^{x_n} = \Big(I + \sum_{i=1}^{n} x_i \otimes x_i\Big)^{-1} \Big(\sum_{i=1}^{n-1} y_i x_i\Big),$$

which can be sequentially computed with complexity $O(d^2)$ thanks to the Sherman-Morrison formula. For this predictor, the following convergence result holds [Cesa-Bianchi and Lugosi, 2006, Theorem 11.8].

Proposition 4. *The Vovk-Azoury-Warmuth predictor on* $(x_1, y_1, \dots, x_n, y_n) \in (\mathbb{R}^d \times \mathbb{R})^n$ *satisfies*

$$\frac{1}{n}\sum_{i=1}^{n}(\langle \theta_i^{x_i}, x_i\rangle - y_i)^2 - \inf_{\theta\in\mathbb{R}^d}\Big\{\frac{1}{n}\sum_{i=1}^{n}(\langle \theta, x_i\rangle - y_i)^2 + \frac{1}{n}\|\theta\|_2^2\Big\} \leq \frac{dB^2\log(1+nR^2/d)}{n}, \quad (1.4)$$

where $B^2 = \max_{i\in\{1,\dots,n\}} y_i^2$ *and* $R^2 = \max_{i\in\{1,\dots,n\}} \|x_i\|_2^2$.

Therefore in the setting of online learning, with no assumption on the generation of the data, we obtain an $O(d\log(n)/n)$ bound on the regret for least-squares problems. Furthermore Cesa-Bianchi and Lugosi [2006] show the optimality of this bound. We also note Rakhlin and Sridharan [2014] extend this method to online non-parametric regression.

1.8 Digest of Least-Squares Regression

We summarize here the different results seen so far for the particular example of the least-squares regression:

$$\min_{\theta\in\mathbb{R}^d} L(\theta) = \frac{1}{2}\mathbb{E}(\langle \phi(X), \theta\rangle - Y)^2,$$

where (X, Y) are random variables of unknown distribution $\mathcal{D}$. We remind that the covariance matrix is denoted by $\Sigma = \mathbb{E}[\phi(X)\otimes\phi(X)]$ and is only assumed invertible. Thus its eigenvalues may be arbitrarily small, in contrast with a recent work by Jain et al. [2016]. The global minimum of $L(\theta)$ is denoted by θ_*.

We consider not being limited in terms of sample complexity and are instead interested in the running time complexity of the different methods.

Empirical risk minimization. In Section 1.4, the generalization error L is first approximated by the training error $\hat{L}(\theta) = \frac{1}{2n}\sum_{i=1}^{n}(\langle \phi(x_i), \theta\rangle - y_i)^2$. Its minimizer $\hat{\theta}$

	Estimation error	Oracle complexity	Running-time complexity
ERM			
Conjugate gradient	$O(\frac{\sigma^2 d}{n})$	d	$O(\frac{\sigma^2 d^3}{\varepsilon})$
Regularized ERM			
Accelerated SDCA	$O(\frac{\sigma^2 d}{n})$	$\tilde{O}(\frac{\sigma^2 d + \|\theta_*\|_2 R\sigma d^{1/2}}{\varepsilon})$	$\tilde{O}(\frac{\sigma^2 d^2 + \|\theta_*\|_2 R\sigma d^{3/2}}{\varepsilon})$
S.A.			
SGD	$O(\frac{\sigma^2 d + R^2\|\theta_*\|^2}{n})$	$O(\frac{\sigma^2 d + R^2\|\theta_*\|^2}{\varepsilon})$	$O(\frac{\sigma^2 d^2 + dR^2\|\theta_*\|^2}{\varepsilon})$
Online Learning			
$\bar{h}_n$ predictor	$O(\frac{\sigma^2 d\log(n) + R^2\|\theta_*\|^2}{n})$	$\tilde{O}(\frac{\sigma^2 d + R^2\|\theta_*\|^2}{\varepsilon})$	?

Table 1.3 – Complexities of least-squares regression. $\tilde{O}$ ignores logarithmic factors. S.A. denotes stochastic approximation.

achieves under reasonable assumptions on (X, Y):

$$L(\hat{\theta}) - L(\theta_*) = O\Big(\frac{\sigma^2 d}{n}\Big).$$

This quadratic function is efficiently solved with the conjugate gradient algorithm presented in Section 1.5.3. The solution $\hat{\theta}$ is obtained in at most d iterations and the complexity of each iteration is $O(nd)$ (the cost of computing a gradient of $\hat{L}$). Therefore the total running time complexity is $O(\frac{\sigma^2 d^3}{\varepsilon})$.

Regularized risk minimization. The ℓ_2-regularized training error $\hat{L}_\lambda(\theta) = \lambda\|\theta\|_2^2 + \frac{1}{2n}\sum_{i=1}^n(\langle\phi(x_i), \theta\rangle - y_i)^2$ may also be considered. Hsu et al. [2014] show its minimizer $\hat{\theta}_\lambda$ achieves, under suitable assumptions, a similar estimation rate

$$L(\hat{\theta}_\lambda) - L(\theta_*) = O\Big(\frac{\sigma^2 d}{n} + \lambda\|\theta_*\|_2^2\Big).$$

Thus $\lambda = O(\frac{\sigma^2 d}{\|\theta_*\|_2^2 n})$ is taken to obtain $L(\hat{\theta}_\lambda) - L(\theta_*) = O(\sigma^2 d/n)$. Furthermore the regularization may be leveraged by using fast optimization methods dedicated to finite sum of strongly convex functions. Accelerated-SDCA has a running time complexity $O(d(n + R\sqrt{n/\lambda}\log(1/\varepsilon)))$. Replacing with the value of λ, the global running time complexity becomes $\tilde{O}(\frac{\sigma^2 d^2 + \|\theta_*\|_2 R\sigma d^{3/2}}{\varepsilon})$.

Stochastic approximations. The generalization error may also be directly optimized using stochastic gradient descent with constant step size and averaging, as in Section 1.6.4. One directly obtains a running time complexity $O(\frac{\sigma^2 d^2 + dR^2\|\theta_*\|^2}{\varepsilon})$.

Online learning. The Vovk-Azoury-Warmuth predictor presented in Section 1.7 guarantees an $O(d\log(n)/n)$ bound on the regret for any sequence $(x_1, y_1, \ldots, x_n, y_n)$. When the data are actually i.i.d., the average predictor $\bar{h}_n$ defined by

$$\bar{h}_n(x) = \frac{1}{n}\sum_{i=1}^{n}\langle \theta_i^x, x\rangle$$

can be taken into account. Classic online to batch conversion techniques ensure an estimation error $O(\sigma^2 d/n + R^2\|\theta_*\|_2^2/n)$. The main downside of this method is that it is unclear how to efficiently compute this average. It may be relaxed in $\tilde{h}_n(x) = \langle \tilde{\theta}_n, x\rangle$ where $\tilde{\theta}_n$ is defined by $\tilde{\theta}_n = \frac{1}{n}\sum_{i=1}^{n}\theta_i^{x_i}$ but the independence between $\tilde{\theta}_n$ and x_n is then lost and the statistical properties of this estimator are not well understood.

Hence the best running time complexity of these methods is $O(\frac{\sigma^2 d^2 + dR^2\|\theta_*\|^2}{\varepsilon})$ with a minimax optimal variance term $\sigma^2 d^2/\varepsilon$ and a suboptimal bias term $O(\frac{dR^2\|\theta_*\|^2}{\varepsilon})$. Actually a bias term with a decrease in $1/\sqrt{\varepsilon}$ is expected since it is the optimal rate of convergence for first-order optimization of smooth non-strongly convex quadratic functions. This issue will be the quest of the two following chapters of this manuscript.

Part I

Stochastic Approximation and Least-Squares Regression

Chapter 2

Multistep Methods for Quadratic Optimization

Abstract

We show that accelerated gradient descent, averaged gradient descent and the heavy-ball method for quadratic non-strongly convex problems may be reformulated as constant parameter second-order difference equation algorithms, where stability of the system is equivalent to convergence at rate $O(1/n^2)$, where n is the number of iterations. We provide a detailed analysis of the eigenvalues of the corresponding linear dynamical system, showing various oscillatory and non-oscillatory behaviors, together with a sharp stability result with explicit constants. We also consider the situation where noisy gradients are available, where we extend our general convergence result, which suggests an alternative algorithm (i.e., with different step sizes) that exhibits the good aspects of both averaging and acceleration.

This chapter is based on our work: "From Averaging to Acceleration, There is Only a Step-size", N. Flammarion and F. Bach, published in the *Proceedings of the International Conference on Learning Theory (COLT)*, 2015.

2.1 Introduction

Many problems in machine learning are naturally cast as convex optimization problems over a Euclidean space; for supervised learning this includes least-squares regression, logistic regression, and the support vector machine. Faced with large amounts of data, practitioners often favor first-order techniques based on gradient descent, leading to algorithms with many cheap iterations. For smooth problems, two extensions of gradient descent have had important theoretical and practical impacts: acceleration and averaging.

Acceleration techniques date back to Nesterov [1983] and have their roots in momentum techniques and conjugate gradient [Polyak, 1987]. For convex problems, with an appropriately weighted momentum term which requires to store two iterates, Nesterov [1983] showed that the traditional convergence rate of $O(1/n)$ for the function

values after n iterations of gradient descent goes down to $O(1/n^2)$ for accelerated gradient descent, such a rate being optimal among first-order techniques that can access only sequences of gradients [Nesterov, 2004]. Like conjugate gradient methods for solving linear systems, these methods are however more sensitive to noise in the gradients; that is, to preserve their improved convergence rates, significantly less noise may be tolerated [d'Aspremont, 2008, Schmidt et al., 2011, Devolder et al., 2014].

Averaging techniques which consist in replacing the iterates by the average of all iterates have also been thoroughly considered, either because they sometimes lead to simpler proofs, or because they lead to improved behavior. In the noiseless case where gradients are exactly available, they do not improve the convergence rate in the convex case; worse, for strongly convex problems, they are not linearly convergent while regular gradient descent is. Their main advantage comes with random unbiased gradients, where it has been shown that they lead to better convergence rates than the unaveraged counterparts, in particular because they allow larger step-sizes [Polyak and Juditsky, 1992, Bach and Moulines, 2011]. For example, for least-squares regression with stochastic gradients, they lead to convergence rates of $O(1/n)$, even in the non-strongly convex case [Bach and Moulines, 2013].

In this chapter, we show that for quadratic problems, both averaging and acceleration are two instances of the same second-order finite difference equation, with different step-sizes. They may thus be analyzed jointly, together with a non-strongly convex version of the heavy-ball method [Polyak, 1987, Section 3.2]. In presence of random zero-mean noise on the gradients, this joint analysis allows to design a novel intermediate algorithm that exhibits the good aspects of both acceleration (quick forgetting of initial conditions) and averaging (robustness to noise).

In this chapter, we make the following contributions:

- We show in Section 2.2 that accelerated gradient descent, averaged gradient descent and the heavy-ball method for quadratic non-strongly convex problems may be reformulated as constant parameter second-order difference equation algorithms, where stability of the system is equivalent to convergence at rate $O(1/n^2)$.
- In Section 2.3, we provide a detailed analysis of the eigenvalues of the corresponding linear dynamical system, showing various oscillatory and non-oscillatory behaviors, together with a sharp stability result with explicit constants.
- In Section 2.4, we consider the situation where noisy gradients are available, where we extend our general convergence result, which suggests an alternative algorithm (i.e., with different step sizes) that exhibits the good aspects of both averaging and acceleration.
- In Section 2.5, we illustrate our results with simulations on synthetic examples.

2.2 Second-Order Iterative Algorithms for Quadratic Functions

In this chapter, we consider minimizing a convex quadratic function $f : \mathbb{R}^d \to \mathbb{R}$ defined as:

$$f(\theta) = \tfrac{1}{2}\langle \theta, H\theta \rangle - \langle q, \theta \rangle, \tag{2.1}$$

with $H \in \mathbb{R}^{d\times d}$ a symmetric positive semi-definite matrix and $q \in \mathbb{R}^d$. Without loss of generality, H is assumed invertible (by projecting onto the orthogonal of its null space), though its eigenvalues could be arbitrarily small. The solution is known to be $\theta_* = H^{-1}q$, but the inverse of the Hessian is often too expensive to compute when d is large. The excess cost function may be simply expressed as $f(\theta_n) - f(\theta_*) = \frac{1}{2}\langle \theta_n - \theta_*, H(\theta_n - \theta_*)\rangle$.

2.2.1 Second-order Algorithms

In this chapter we study second-order iterative algorithms of the form:

$$\theta_{n+1} = A_n\theta_n + B_n\theta_{n-1} + c_n, \tag{2.2}$$

started with $\theta_1 = \theta_0$ in $\mathbb{R}^d$, with $A_n \in \mathbb{R}^{d\times d}$, $B_n \in \mathbb{R}^{d\times d}$ and $c_n \in \mathbb{R}^d$ for all $n \in \mathbb{N}^*$. We impose the natural restriction that the optimum θ_* is a stationary point of this recursion, that is, for all $n \in \mathbb{N}^*$:

$$\theta_* = A_n\theta_* + B_n\theta_* + c_n. \qquad (\theta_*\text{-stationarity})$$

By letting $\phi_n = \theta_n - \theta_*$ we then have $\phi_{n+1} = A_n\phi_n + B_n\phi_{n-1}$, started from $\phi_0 = \phi_1 = \theta_0 - \theta_*$. Thus, we restrict our problem to the study of the convergence of an iterative system to 0.

In connection with accelerated methods, we are interested in algorithms for which

$$f(\theta_n) - f(\theta_*) = \tfrac{1}{2}\langle \phi_n, H\phi_n\rangle \tag{2.3}$$

converges to 0 at a speed of $O(1/n^2)$. Within this context we impose that A_n and B_n have the form:

$$A_n = \tfrac{n}{n+1}A \text{ and } B_n = \tfrac{n-1}{n+1}B \quad \forall n \in \mathbb{N} \text{ with } A, B \in \mathbb{R}^{d\times d}. \qquad (n\text{-scalability})$$

By letting $\eta_n = n\phi_n = n(\theta_n - \theta_*)$, we can now study the simple iterative system with *constant* terms $\eta_{n+1} = A\eta_n + B\eta_{n-1}$, started at $\eta_0 = 0$ and $\eta_1 = \theta_0 - \theta_*$. Showing that the function values $f(\eta_n)$ remain bounded, we directly have from Eq. (2.3), the convergence of $f(\theta_n)$ to $f(\theta_*)$ at the speed $O\left(1/n^2\right)$. Thus the n-scalability property allows to switch from a convergence problem to a stability problem.

For feasibility concerns the method can only access H through matrix-vector products. Therefore A and B should be polynomials in H and c a polynomial in H times q, if possible of low degree. The following theorem clarifies the general form of iterative systems which share these three properties (see proof in Appendix 2.B).

Theorem 2. *Let $(P_n, Q_n, R_n) \in (\mathbb{R}[X])^3$ for all $n \in \mathbb{N}$, be a sequence of polynomials. If the iterative algorithm defined by Eq. (2.2) with $A_n = P_n(H)$, $B_n = Q_n(H)$ and $c_n = R(H)q$ satisfies the θ_*-stationarity and n-scalability properties, there are polynomials $(\bar{A}, \bar{B}) \in (\mathbb{R}[X])^2$ such that:*

$$A_n \;=\; 2\tfrac{n}{n+1}\big(I - \tfrac{1}{2}\big(\bar{A}(H) + \bar{B}(H)\big)H\big),$$

$$B_n \;=\; -\tfrac{n-1}{n+1}\big(I - \bar{B}(H)H\big) \quad \textit{and} \quad c_n = \tfrac{1}{n+1}\big(n\bar{A}(H) + \bar{B}(H)\big)q.$$

Note that our result prevents A_n and B_n from being zero, thus requiring the algorithm to strictly be of second order. This illustrates the fact that first-order algorithms as gradient descent do not have the convergence rate in $O(1/n^2)$.

We now restrict our class of algorithms to lowest possible order polynomials, that is, $\bar{A} = \alpha I$ and $\bar{B} = \beta I$ with $(\alpha, \beta) \in \mathbb{R}^2$, which correspond to the fewest matrix-vector products per iteration, leading to the *constant-coefficient* recursion for $\eta_n = n\phi_n = n(\theta_n - \theta_*)$:

$$\eta_{n+1} = (I - \alpha H)\,\eta_n + (I - \beta H)\,(\eta_n - \eta_{n-1})\,. \tag{2.4}$$

Expression with gradients of f. The recursion in Eq. (2.4) may be written with gradients of f in multiple ways. In order to preserve the parallel with accelerated techniques, we rewrite it as:

$$\theta_{n+1} = \tfrac{2n}{n+1}\theta_n - \tfrac{n-1}{n+1}\theta_{n-1} - \tfrac{n\alpha+\beta}{n+1} f'\big(\tfrac{n(\alpha+\beta)}{n\alpha+\beta}\theta_n - \tfrac{(n-1)\beta}{n\alpha+\beta}\theta_{n-1}\big). \tag{2.5}$$

It may be interpreted as a modified gradient recursion with two potentially different affine (i.e., with coefficients that sum to one) combinations of the two past iterates. This reformulation will also be crucial when using noisy gradients. The allowed values for $(\alpha, \beta) \in \mathbb{R}^2$ will be determined in the following sections.

2.2.2 Examples

Averaged gradient descent. We consider averaged gradient descent (referred to from now on as "Av-GD") [Polyak and Juditsky, 1992] with step-size $\gamma \in \mathbb{R}$ defined by $\psi_{n+1} = \psi_n - \gamma f'(\psi_n)$ and $\theta_{n+1} = \frac{1}{n+1}\sum_{i=1}^{n+1}\psi_i$. When computing the average online as $\theta_{n+1} = \theta_n + \frac{1}{n+1}(\psi_{n+1} - \theta_n)$ and seeing the average as the main iterate, the algorithm becomes (see proof in Appendix 2.B.2):

$$\theta_{n+1} = \tfrac{2n}{n+1}\theta_n - \tfrac{n-1}{n+1}\theta_{n-1} - \tfrac{\gamma}{n+1} f'\big(n\theta_n - (n-1)\theta_{n-1}\big).$$

This corresponds to Eq. (2.5) with $\alpha = 0$ and $\beta = \gamma$.

Accelerated gradient descent. We consider the accelerated gradient descent (referred to from now on as "Acc-GD") [Nesterov, 1983] with step-sizes $(\gamma, \delta_n) \in \mathbb{R}^2$:

$$\theta_{n+1} = \omega_n - \gamma f'(\omega_n), \qquad \omega_n = \theta_n + \delta_n(\theta_n - \theta_{n-1}).$$

For smooth optimization, the accelerated literature [Nesterov, 2004, Beck and Teboulle, 2009] uses the step-size $\delta_n = 1 - \frac{3}{n+1}$ and their results are not valid for bigger step-size δ_n. However $\delta_n = 1 - \frac{2}{n+1}$ is compatible with the framework of Lan [2012] and is more convenient for our set-up. This corresponds to Eq. (2.5) with $\alpha = \gamma$ and $\beta = \gamma$. Note that accelerated techniques are more generally applicable, e.g., to composite optimization with smooth functions [Nesterov, 2013, Beck and Teboulle, 2009].

Heavy ball. We consider the heavy-ball algorithm (referred to from now on as "HB") [Polyak, 1964] with step-sizes $(\gamma, \delta_n) \in \mathbb{R}^2$:

$$\theta_{n+1} = \theta_n - \gamma f'(\theta_n) + \delta_n(\theta_n - \theta_{n-1}),$$

when $\delta_n = 1 - \frac{2}{n+1}$ (we note that typically δ_n is constant for strongly convex problems). This corresponds to Eq. (2.5) with $\alpha = \gamma$ and $\beta = 0$.

2.3 Convergence with Noiseless Gradients

We study the convergence of the iterates defined by Eq. (2.4). This is a second-order iterative system with constant coefficients that it is standard to cast in a linear framework [see, e.g., Ortega and Rheinboldt, 2000]. We may rewrite it as:

$$\Theta_n = F\Theta_{n-1}, \quad \text{with } \Theta_n = \begin{pmatrix} \eta_n \\ \eta_{n-1} \end{pmatrix} \text{ and } F = \begin{pmatrix} 2I - (\alpha+\beta)\,H & \beta H - I \\ I & 0 \end{pmatrix} \in \mathbb{R}^{2d \times 2d}.$$

Thus $\Theta_n = F^n\Theta_0$. Following O'Donoghue and Candès [2013], if we consider an eigenvalue decomposition of H, i.e., $H = P\mathrm{Diag}(h)P^\top$ with P an orthogonal matrix and (h_i) the eigenvalues of H, sorted in decreasing order: $h_d = L \geq h_{d-1} \geq \cdots \geq h_2 \geq h_1 = \mu > 0$, then Eq. (2.4) may be rewritten as:

$$P^\top\eta_{n+1} = \left(I - \alpha\mathrm{Diag}\,(h)\right)P^\top\eta_n + \left(I - \beta\mathrm{Diag}\,(h)\right)\left(P^\top\eta_n - P^\top\eta_{n-1}\right). \tag{2.6}$$

Thus there is no interaction between the different eigenspaces and we may consider, for the analysis only, d different recursions with $\eta_n^i = p_i^\top\eta_n$, $i \in \{1, ..., d\}$, where $p_i \in \mathbb{R}^d$ is the i-th column of P:

$$\eta_{n+1}^i = (1 - \alpha h_i)\,\eta_n^i + (1 - \beta h_i)\left(\eta_n^i - \eta_{n-1}^i\right). \tag{2.7}$$

2.3.1 Characteristic Polynomial and Eigenvalues

In this section, we consider a fixed $i \in \{1, \ldots, d\}$ and study the stability in the corresponding eigenspace. This linear dynamical system may be analyzed by studying the eigenvalues of the 2×2-matrix $F_i = \begin{pmatrix} 2 - (\alpha+\beta)h_i & \beta h_i - 1 \\ 1 & 0 \end{pmatrix}$. These eigenvalues are the roots of its characteristic polynomial which is:

$$\det(xI - F_i) = \left(x\left(x - 2 + (\alpha+\beta)h_i\right) + 1 - \beta h_i\right) = x^2 - 2x\big(1 - \big(\tfrac{\alpha+\beta}{2}\big)h_i\big) + 1 - \beta h_i.$$

To compute the roots of the second-order polynomial, we compute its discriminant:

$$\Delta_i = \big(1 - \big(\tfrac{\alpha+\beta}{2}\big)h_i\big)^2 - 1 + \beta h_i = h_i\big(\big(\tfrac{\alpha+\beta}{2}\big)^2 h_i - \alpha\big).$$

Depending on the sign of the discriminant Δ_i, there will be two real distinct eigenvalues ($\Delta_i > 0$), two complex conjugate eigenvalues ($\Delta_i < 0$) or a single real eigen-

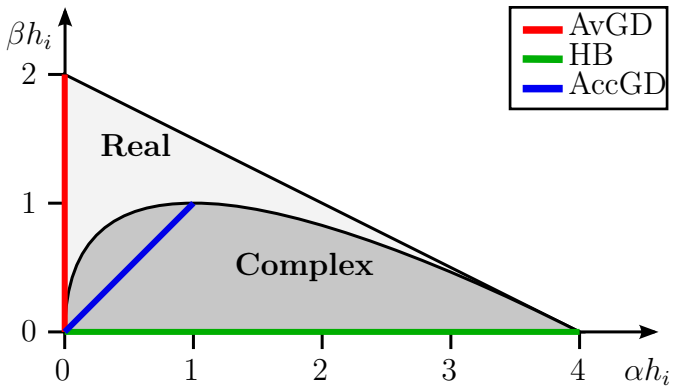


Figure 2-1 – Area of stability of the algorithm, with the three traditional algorithms represented. In the interior of the triangle, the convergence is linear.

value ($\Delta_i = 0$).

We will now study the sign of Δ_i. In each different case, we will determine under what conditions on α and β the modulus of the eigenvalues is less than one, which means that the iterates $(\eta_n^i)_n$ remain bounded and the iterates $(\theta_n)_n$ converge to θ_*. We may then compute function values from Eq. (2.3) as $f(\theta_n) - f(\theta_*) = \frac{1}{2n^2}\sum_{i=1}^d (\eta_n^i)^2 h_i = \frac{1}{2}\sum_{i=1}^d (\phi_n^i)^2 h_i$.

The various regimes are summarized in Figure 2-1: there is a triangle of values of $(\alpha h_i, \beta h_i)$ for which the algorithm remains stable (i.e., the iterates $(\eta_n)_n$ do not diverge), with either complex or real eigenvalues. In the following lemmas (see proof in Appendix 2.C), we provide a detailed analysis that leads to Figure 2-1.

Lemma 1 (Real eigenvalues)**.** *The discriminant Δ_i is strictly positive and the algorithm is stable if and only if*

$$\alpha \geq 0, \qquad \alpha + 2\beta \leq 4/h_i, \qquad \alpha + \beta > 2\sqrt{\alpha/h_i}.$$

We then have two real roots $r_i^{\pm} = r_i \pm \sqrt{\Delta_i}$, with $r_i = 1 - (\frac{\alpha+\beta}{2})h_i$. Moreover, we have:

$$(\phi_n^i)^2 h_i = \frac{(\phi_1^i)^2 h_i}{4n^2} \frac{\left[(r_i + \sqrt{\Delta_i})^n - (r_i - \sqrt{\Delta_i})^n\right]^2}{\Delta_i}. \tag{2.8}$$

Therefore, for real eigenvalues, $((\phi_n^i)^2 h_i)_n$ will converge to 0 at a speed of $O(1/n^2)$ however the constant Δ_i may be arbitrarily small (and thus the scaling factor arbitrarily large). Furthermore we have linear convergence if the inequalities in the lemmas are strict.

Lemma 2 (Complex eigenvalues)**.** *The discriminant Δ_i is stricly negative and the algorithm is stable if and only if*

$$\alpha \geq 0, \qquad \beta \geq 0, \qquad \alpha + \beta < \sqrt{\alpha/h_i}.$$

We then have two complex conjugate eigenvalues: $r_i^{\pm} = r_i \pm \sqrt{-1}\sqrt{-\Delta_i}$. *Moreover, we have:*

$$(\phi_n^i)^2 h_i = \frac{(\phi_1^i)^2}{n^2} \frac{\sin^2(\omega_i n)}{\left(\alpha - (\frac{\alpha+\beta}{2})^2 h_i\right)} \rho^{2n}. \tag{2.9}$$

with $\rho_i = \sqrt{1-\beta h_i}$, *and* ω_i *defined through* $\sin(\omega_i) = \sqrt{-\Delta_i}/\rho_i$ *and* $\cos(\omega_i) = r_i/\rho_i$.

Therefore, for complex eigenvalues, there is a linear convergence if the inequalities in the lemma are strict. Moreover, $((\phi_n^i)^2 h_i)_n$ oscillates to 0 at a speed of $O(1/n^2)$ even if h_i is arbitrarily small.

Coalescing eigenvalues. When the discriminants go to zero in the explicit formulas of the real and complex cases, both the denominator and numerator of $((\phi_n^i)^2 h_i)_n$ will go to zero. In the limit case, when the discriminant is equal to zero, we will have a double real eigenvalue. This happens for $\beta = 2\sqrt{\alpha/h_i} - \alpha$. Then the eigenvalue is $r_i = 1 - \sqrt{\alpha h_i}$, and the algorithm is stable for $0 < \alpha < 4/h_i$, we then have $(\phi_n^i)^2 h_i = h_i(\phi_1^i)^2(1-\sqrt{\alpha h_i})^{2(n-1)}$. This can be obtained by letting Δ_i goes to 0 in the real and complex cases (see also Appendix 2.C.3).

Summary. To conclude, the iterate $(\eta_n^i)_n = (n(\theta_n^i - \theta_*^i))_n$ will be stable for $\alpha \in [0, 4/h_i]$ and $\beta \in [0, 2/h_i - \alpha/2]$. According to the values of α and β this iterate will have a different behavior. In the complex case, the roots are complex conjugate with magnitude $\sqrt{1-\beta h_i}$. Thus, when $\beta > 0$, $(\eta_n^i)_n$ will converge to 0, oscillating, at rate $\sqrt{1-\beta h_i}$. In the real case, the two roots are real and distinct. However the product of the two roots is equal to $\sqrt{1-\beta h_i}$, thus one will have a higher magnitude and $(\eta_n^i)_n$ will converges to 0 at rate higher than in the complex case (as long as α and β belong to the interior of the stability region).

Finally, for a given quadratic function f, all the d iterates $(\eta_n^i)_n$ should be bounded, therefore we must have $\alpha \in [0, 4/L]$ and $\beta \in [0, 2/L - \alpha/2]$. Then, depending on the value of h_i, some eigenvalues may be complex or real.

2.3.2 Classical Examples

For particular choices of α and β, displayed in Figure 2-1, the eigenvalues are either all real or all complex, as shown in the table below.

	Av-GD	**Acc-GD**	**Heavy ball**
α	0	γ	γ
β	γ	γ	0
Δ_i	$(\gamma h_i)^2$	$-\gamma h_i(1-\gamma h_i)$	$-\gamma h_i(1-\frac{\gamma h_i}{4})$
$r_i^{\pm}$	$1,\ 1-\gamma h_i$	$\sqrt{1-\gamma h_i}e^{\pm i\omega_i}$	$e^{\pm i\omega_i}$
$\cos(\omega_i)$		$\sqrt{1-\gamma h_i}$	$1-\frac{\gamma}{2}h_i$
ρ_i		$\sqrt{1-\gamma h_i}$	1

Averaged gradient descent loses linear convergence for strongly convex problems, because $r_i^+ = 1$ for all eigensubspaces. Similarly, the heavy ball method is not adaptive to strong convexity because $\rho_i = 1$. However, accelerated gradient descent, although designed for non-strongly convex problems, is adaptive because $\rho_i = \sqrt{1-\gamma h_i}$ depends on h_i while α and β do not. These last two algorithms have an oscillatory behavior which can be observed in practice and has been already studied [Su et al., 2014].

Note that all the classical methods choose step-sizes α and β either having all the eigenvalues real or complex; whereas we will see in Section 2.4 that it is significant to combine both behaviors in the presence of noise.

2.3.3 General Bound

Even if the exact formulas in Lemmas 1 and 2 are computable, they are not easily interpretable. In particular when the two roots become close, the denominator will go to zero, which prevents us from bounding them easily. When we further restrict the domain of (α, β), we can always bound the iterate by the general bound (see proof in Appendix 2.D):

Theorem 3. *For $\alpha \leq 1/h_i$ and $0 \leq \beta \leq 2/h_i - \alpha$, we have*

$$(\eta_n^i)^2 \leq \min\left\{\frac{2(\eta_1^i)^2}{\alpha h_i}, \frac{8(\eta_1^i)^2 n}{(\alpha+\beta)h_i}, \frac{16(\eta_1^i)^2}{(\alpha+\beta)^2 h_i^2}\right\}. \tag{2.10}$$

These bounds are shown by dividing the set of (α, β) in three regions where we obtain specific bounds. They do not depend on the regime of the eigenvalues (complex or real); this enables us to get the following general bound on the function values, our main result for the deterministic case.

Corollary 1. *For $0 \leq \alpha \leq 1/L$ and $0 \leq \beta \leq 2/L - \alpha$:*

$$f(\theta_n) - f(\theta_*) \leq \min\left\{\frac{\|\theta_0 - \theta_*\|^2}{\alpha n^2}, \frac{4\|\theta_0 - \theta_*\|^2}{(\alpha+\beta)n}\right\}. \tag{2.11}$$

We can make the following observations:

- The first bound $\frac{\|\theta_0-\theta_*\|^2}{\alpha n^2}$ corresponds to the traditional acceleration result, and is only relevant for $\alpha > 0$ (that is, for Nesterov acceleration and the heavy-ball method, but not for averaging). We recover the traditional convergence rate of second-order methods for quadratic functions in the singular case, such as conjugate gradient [Polyak, 1987, Section 6.1].
- While the result above focuses on function values, like most results in the non-strongly convex case, the distance to optimum $\|\theta_n - \theta_*\|^2$ typically does not go to zero (although it remains bounded in our situation).
- When $\alpha = 0$ (averaged gradient descent), then the second bound $\frac{4\|\theta_0-\theta_*\|^2}{(\alpha+\beta)n}$ provides a convergence rate of $O(1/n)$ if no assumption is made regarding the starting point

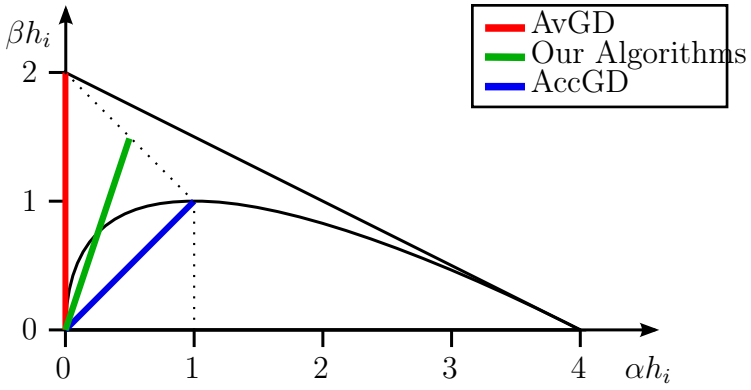


Figure 2-2 – Trade-off between averaged and accelerated methods for noisy gradients.

θ_0, while the last bound of Theorem 3 would lead to a bound $\frac{8\|H^{-1/2}(\theta_0-\theta_*)\|^2}{(\alpha+\beta)^2 n^2}$, that is a rate of $O(1/n^2)$, only for some starting points.

– As shown in Appendix 2.E by exhibiting explicit sequences of quadratic functions, the inverse dependence in αn^2 and $(\alpha+\beta)n$ in Eq. (2.11) is not improvable.

2.4 Quadratic Optimization with Additive Noise

In many practical situations, the gradient of f is not available for the recursion in Eq. (2.5), but only a noisy version. In this chapter, we only consider additive uncorrelated noise with finite variance.

2.4.1 Stochastic Difference Equation

We now assume that the true gradient is not available and we rather have access to a noisy oracle for the gradient of f in Eq. (2.5). We assume that the oracle outputs a noisy gradient $f'\big(\frac{n(\alpha+\beta)}{n\alpha+\beta}\theta_n - \frac{(n-1)\beta}{n\alpha+\beta}\theta_{n-1}\big) - \varepsilon_{n+1}$. The noise (ε_n) is assumed to be uncorrelated zero-mean with bounded covariance, i.e., $\mathbb{E}[\varepsilon_n \otimes \varepsilon_m] = 0$ for all $n \neq m$ and $\mathbb{E}[\varepsilon_n \otimes \varepsilon_n] \preccurlyeq C$, where $A \preccurlyeq B$ means that $B - A$ is positive semi-definite.

For quadratic functions, for the reduced variable $\eta_n = n\phi_n = n(\theta_n - \theta_*)$, we get:

$$\eta_{n+1} = (I - \alpha H)\eta_n + (I - \beta H)(\eta_n - \eta_{n-1}) + [n\alpha + \beta]\varepsilon_{n+1}. \tag{2.12}$$

Note that algorithms with $\alpha \neq 0$ will have an important level of noise because of the term $n\alpha\varepsilon_{n+1}$. We denote by $\xi_{n+1} = \begin{pmatrix} [n\alpha+\beta]\varepsilon_{n+1} \\ 0 \end{pmatrix}$ and we now have the recursion:

$$\Theta_{n+1} = F\Theta_n + \xi_{n+1}, \tag{2.13}$$

which is a standard noisy linear dynamical system [see, e.g., Arnold, 1998] with un-

correlated noise process (ξ_n). We may thus express Θ_n directly as $\Theta_n = F^{n-1}\Theta_1 + \sum_{k=2}^n F^{n-k}\xi_k$, and its expected second-order moment as

$$\mathbb{E}\big(\Theta_n\Theta_n^\top\big) = F^{n-1}\Theta_1\Theta_1^\top(F^{n-1})^\top + \sum_{k=2}^n F^{n-k}\mathbb{E}\big(\xi_k\xi_k^\top\big)(F^{n-k})^\top.$$

In order to obtain the expected excess cost function, we simply need to compute $\operatorname{tr}\begin{pmatrix} 0 & 0 \\ 0 & H \end{pmatrix}\mathbb{E}\big(\Theta_n\Theta_n^\top\big)$, which thus decomposes as a term that only depends on initial conditions (which is exactly the one computed and studied in Section 2.3.3), and a new term that depends on the noise.

2.4.2 Convergence Result

For a quadratic function f with arbitrarily small eigenvalues and uncorrelated noise with finite covariance, we obtain the following convergence result (see proof in Appendix 2.F); since we will allow the parameters α and β to depend on the time we stop the algorithm, we introduce the horizon N:

Theorem 4 (Convergence rates with noisy gradients)**.** *With $\mathbb{E}[\varepsilon_n \otimes \varepsilon_n] = C$ for all $n \in \mathbb{N}$, for $\alpha \leq \frac{1}{L}$ and $0 \leq \beta \leq \frac{2}{L} - \alpha$. Then for any $N \in \mathbb{N}$, we have:*

$$\mathbb{E}f(\theta_N) - f(\theta_*) \leq \min\left\{\frac{\|\theta_0-\theta_*\|^2}{\alpha N^2} + \frac{(\alpha N+\beta)^2}{\alpha N}\operatorname{tr}(C), \frac{4\|\theta_0-\theta_*\|^2}{(\alpha+\beta)N} + \frac{4(\alpha N+\beta)^2}{\alpha+\beta}\operatorname{tr}(C)\right\}.$$

We can make the following observations:

- Although we only provide an upper-bound, the proof technique relies on direct moment computations in each eigensubspace with few inequalities, and we conjecture that the scalings with respect to N are tight.
- For $\alpha = 0$ and $\beta = 1/L$ (which corresponds to averaged gradient descent), the second bound leads to $\frac{4L\|\theta_0-\theta_*\|^2}{N} + \frac{4\operatorname{tr}(C)}{L}$, which is bounded but not converging to zero. We recover a result from Bach and Moulines [2011, Theorem 1].
- For $\alpha = \beta = 1/L$ (which corresponds to Nesterov's acceleration), the first bound leads to $\frac{L\|\theta_0-\theta_*\|^2}{N^2} + \frac{(N+1)\operatorname{tr}(C)}{L}$, and our bound suggests that the algorithm diverges, which we have observed in our experiments in Appendix 2.A.
- For $\alpha = 0$ and $\beta = 1/L\sqrt{N}$, the second bound leads to $\frac{4L\|\theta_0-\theta_*\|^2}{\sqrt{N}} + \frac{4\operatorname{tr}(C)}{L\sqrt{N}}$, and we recover the traditional rate of $1/\sqrt{N}$ for stochastic gradient in the non-strongly convex case.
- When the values of the bias and the variance are known we can choose α and β such that the trade-off between the bias and the variance is optimal in our bound, as the following corrollary shows. Note that in the bound below, taking a non-zero β enables the bias term to be adaptive to hidden strong convexity.

Corollary 2. *For* $\alpha = \min\left\{\frac{\|\theta_0-\theta_*\|}{2\sqrt{\operatorname{tr} C}N^{3/2}}, 1/L\right\}$ *and* $\beta \in [0, \min\{N\alpha, 1/L\}]$*, we have:*

$$\mathbb{E}f(\theta_N) - f(\theta_*) \leq \frac{2L\|\theta_0-\theta_*\|^2}{N^2} + \frac{4\sqrt{\operatorname{tr} C}\|\theta_0-\theta_*\|}{\sqrt{N}}.$$

2.4.3 Structured Noise and Least-Squares Regression

When only the noise total variance $\operatorname{tr}(C)$ is considered, as shown in Section 2.4.4, Corollary 2 recovers existing (more general) results. Our framework however leads to improved result for *structured noise processes* frequent in machine learning, in particular in least-squares regression which we now consider but also in others problems [see, e.g. Bach and Moulines, 2013].

Assume we observe independent and identically distributed pairs $(x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$ and we want to minimize the expected loss $f(\theta) = \frac{1}{2}\mathbb{E}[(y_n - \langle\theta, x_n\rangle)^2]$. We denote by $H = \mathbb{E}(x_n \otimes x_n)$ the covariance matrix which is assumed invertible. The global minimum of f is attained at $\theta_* \in \mathbb{R}^d$ defined as before and we denote by $r_n = y_n - \langle\theta_*, x_n\rangle$ the statistical noise, which we assume bounded by σ. We have $\mathbb{E}[r_n x_n] = 0$. In an online setting, we observe the gradient $(x_n \otimes x_n)(\theta - \theta_*) - r_n x_n$, whose expectation is the gradient $f'(\theta)$. This corresponds to a noise in the gradient of $\varepsilon_n = (H - x_n \otimes x_n)(\theta - \theta_*) + r_n x_n$. Given θ, if the data (x_n, y_n) are almost surely bounded, the covariance matrix of this noise is bounded by a constant times H. This suggests to characterize the noise convergence by $\operatorname{tr}(CH^{-1})$, which is bounded even though H has arbitrarily small eigenvalues.

However, our result will not apply to stochastic gradient descent (SGD) for least-squares, because of the term $(H - x_n \otimes x_n)(\theta - \theta_*)$ which depends on θ, but to a "semi-stochastic" recursion where the noisy gradient is $H(\theta - \theta_*) - r_n x_n$, with a noise process $\varepsilon_n = r_n x_n$, which is such that $\mathbb{E}[\varepsilon_n \otimes \varepsilon_n] \preccurlyeq \sigma^2 H$, and has been used by Bach and Moulines [2011] and Dieuleveut and Bach [2015] to prove results on regular stochastic gradient descent. We conjecture that our algorithm (and results) also applies in the regular SGD case, and we provide encouraging experiments in Section 2.5.

For this particular structured noise we can take advantage of a large β:

Theorem 5 (Convergence rates with structured noisy gradients)**.** *Let* $\alpha \leq \frac{1}{L}$ *and* $0 \leq \beta \leq \frac{3}{2L} - \frac{\alpha}{2}$*. For any* $N \in \mathbb{N}$*,* $\mathbb{E}f(\theta_N) - f(\theta_*)$ *is upper-bounded by:*

$$\min\left\{\frac{\|\theta_0-\theta_*\|^2}{N^2\alpha} + \frac{(\alpha N+\beta)^2}{\alpha\beta N^2}\operatorname{tr}(CH^{-1}), \frac{4L\|\theta_0-\theta_*\|^2}{(\alpha+\beta)N} + \frac{8(\alpha N+\beta)^2\operatorname{tr}(CH^{-1})}{(\alpha+\beta)^2 N}\right\}.$$

We can make the following observations:

- For $\alpha = 0$ and $\beta = 1/L$ (which corresponds to averaged gradient descent), the second bound leads to $\frac{4L\|\theta_0-\theta_*\|^2}{N} + \frac{8\operatorname{tr}(CH^{-1})}{N}$. We recover a result from Bach and Moulines [2013, Theorem 1]. Note that when $C \preccurlyeq \sigma^2 H$, $\operatorname{tr}(CH^{-1}) \leq \sigma^2 d$.
- For $\alpha = \beta = 1/L$ (which corresponds to Nesterov's acceleration), the first bound

leads to $\frac{L\|\theta_0-\theta_*\|^2}{N^2} + \operatorname{tr}(CH^{-1})$, which is bounded but not converging to zero (as opposed to the the unstructured noise where the algorithm may diverge).

- For $\alpha = 1/(LN^a)$ with $0 \leqslant a \leqslant 1$ and $\beta = 1/L$, the first bound leads to $\frac{L\|\theta_0-\theta_*\|^2}{N^{2-a}} + \frac{\operatorname{tr}(CH^{-1})}{N^a}$. We thus obtain an explicit bias-variance trade-off by changing the value of a.
- When the values of the bias and the variance are known we can choose α and β with an optimized trade-off, as the following corrollary shows:

Corollary 3. *For* $\alpha = \min\left\{\frac{\|\theta_0-\theta_*\|}{\sqrt{L\operatorname{tr}(CH^{-1})N}}, 1/L\right\}$ *and* $\beta = \min\{N\alpha, 1/L\}$ *we have:*

$$\mathbb{E}f(\theta_N) - f(\theta_*) \leqslant \max\left\{\frac{5\operatorname{tr}(CH^{-1})}{N}, \frac{5\sqrt{\operatorname{tr}(CH^{-1})L}\|\theta_0-\theta_*\|}{N}, \frac{2\|\theta_0-\theta_*\|^2 L}{N^2}\right\}.$$

2.4.4 Related Work

Acceleration and noisy gradients. Several authors [Lan, 2012, Hu et al., 2009, Xiao, 2010] have shown that by using a step-size proportional to $1/N^{3/2}$ accelerated methods with noisy gradients lead to the same convergence rate of $O\big(\frac{L\|\theta_0-\theta_*\|^2}{N^2} + \frac{\|\theta_0-\theta_*\|\sqrt{\operatorname{tr}(C)}}{\sqrt{N}}\big)$ as in Corollary 2, for smooth functions. Thus, for unstructured noise, our analysis provides insights in the behavior of second-order algorithms, without improving bounds. We get significant improvements for structured noises.

Least-squares regression. When the noise is structured as in least-square regression and more generally in linear supervised learning, Bach and Moulines [2011] have shown that using averaged stochastic gradient descent with constant step-size leads to the convergence rate of $O\big(\frac{L\|\theta_0-\theta_0\|^2}{N} + \frac{\sigma^2 d}{N}\big)$. It has been highlighted by Défossez and Bach [2015] that the bias term $\frac{L\|\theta_0-\theta_*\|^2}{N}$ may often be the dominant one in practice. Our result in Corollary 3 leads to an improved bias term in $O(1/N^2)$ with the price of a potentially slightly worse constant in the variance term. However, with optimal constants in Corollary 3, the new algorithm is always an improvement over averaged stochastic gradient descent in all situations. If constants are unknown, we may use $\alpha = 1/(LN^a)$ with $0 \leqslant a \leqslant 1$ and $\beta = 1/L$ and we choose a depending on the emphasis we want to put on bias or variance.

Minimax convergence rates. For noisy quadratic problems, the convergence rate nicely decomposes into two terms, a bias term which corresponds to the noiseless problem and the variance term which corresponds to a problem started at θ_*. For each of these two terms, lower bounds are known. For the bias term, if $N \leqslant d$, then the lower bound is, up to constants, $L\|\theta_0-\theta_*\|^2/N^2$ [Nesterov, 2004, Theorem 2.1.7]. For the variance term, for the general noisy gradient situation, we show in Appendix 2.H that for $N \leqslant d$, it is $(\operatorname{tr} C)/(L\sqrt{N})$, while for least-squares regression, it is $\sigma^2 d/N$ [Tsybakov, 2009]. Thus, for the two situations, we attain the two lower bounds *simultaneously* for situations where respectively $L\|\theta_0-\theta_*\|^2 \leqslant (\operatorname{tr} C)/L$ and

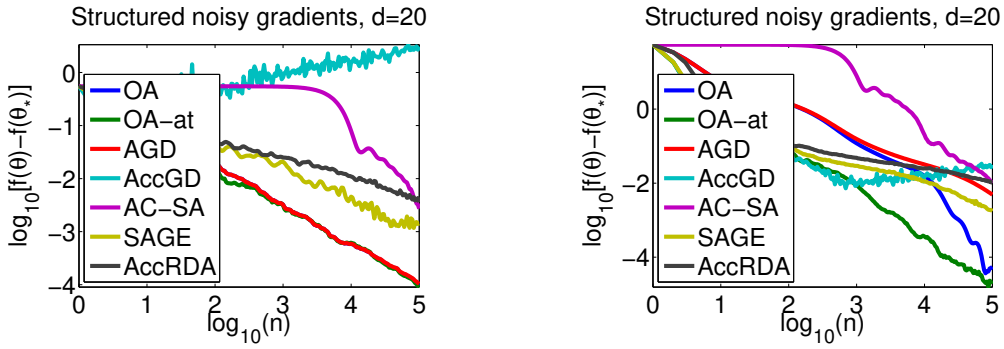


Figure 2-3 – Quadratic optimization with regression noise. Left $\sigma = 1$, $r = 1$. Right $\sigma = 0.1$, $r = 10$.

$L\|\theta_0 - \theta_*\|^2 \leq d\sigma^2$. It remains an open problem to achieve the two minimax terms in all situations.

Other algorithms as special cases. We also note as shown in Appendix 2.G that in the special case of quadratic functions, the algorithms of Lan [2012], Hu et al. [2009], Xiao [2010] could be unified into our framework (although they have significantly different formulations and justifications in the smooth case).

2.5 Experiments

In this section, we illustrate our theoretical results on synthetic examples. We consider a matrix H that has random eigenvectors and eigenvalues $1/k^m$, for $k = 1, \ldots, d$ and $m \in \mathbb{N}$. We take a random optimum θ_* and a random starting point θ_0 such that $r = \|\theta_0 - \theta_*\| = 1$ (unless otherwise specified). In Appendix 2.A, we illustrate the noiseless results of Section 2.3, in particular the oscillatory behaviors and the influence of all eigenvalues, as well as unstructured noisy gradients. In this section, we focus on noisy gradients with structured noise (as described in Section 2.4.3), where our new algorithms (referred to as "OA") show significant improvements.

We compare our algorithm to other stochastic accelerated algorithms, that is, AC-SA [Lan, 2012], SAGE [Hu et al., 2009] and Acc-RDA [Xiao, 2010] which are presented in Appendix 2.G. For all these algorithms (and ours) we take the optimal step-sizes defined in these papers. We show results averaged over 10 replications.

Homoscedastic noise. We first consider an i.i.d. zero mean noise whose covariance matrix is proportional to H. We also consider a variant "OA-at" of our algorithm with an any-time step-size function of n rather than N (for which we currently have no proof of convergence). In Figure 2-3, we take into account two different set-ups. In the left plot, the variance dominates the bias (with $r = \|\theta_0 - \theta_*\| = \sigma$). We see that (a) Acc-GD does not converge to the optimum but does not diverge either, (b) Av-GD and our algorithms achieve the optimal rate of convergence of $O(\sigma^2 d/n)$, whereas (c) other accelerated algorithms only converge at rate $O(1/\sqrt{n})$. In the right plot, the

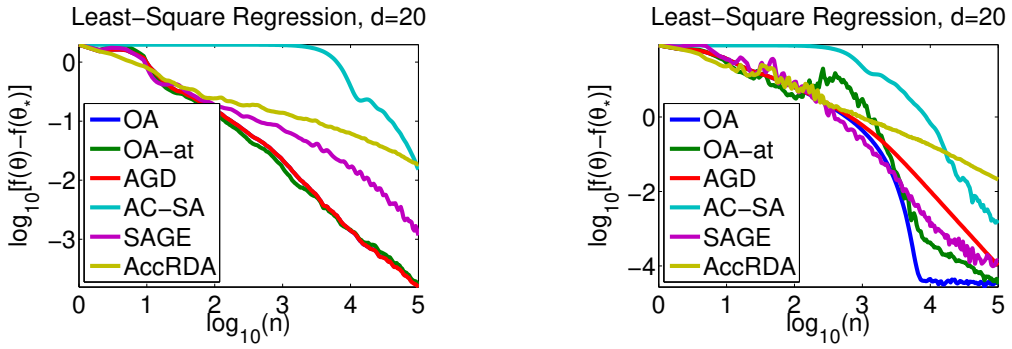


Figure 2-4 – Least-Square Regression. Left $\sigma = 1$, $r = 1$. Right $\sigma = 0.1$, $r = 10$.

bias dominates the variance ($r = 10$ and $\sigma = 0.1$). In this situation our algorithm outperforms all others.

Application to least-squares regression. We now see how these algorithms behave for least-squares regressions and the regular (non-homoscedastic) stochastic gradients described in Section 2.4.3. We consider normally distributed inputs. The covariance matrix H is the same as before. The outputs are generated from a linear function with homoscedatic noise with a signal-to-noise ratio of σ. We consider $d = 20$. We show results averaged over 10 replications. In Figure 2-4, we consider again a situation where the variance dominates the bias (left) and vice versa (right). We see that our algorithm has the same good behavior as in the homoscedastic noise case and we conjecture that our bounds also hold in this situation.

2.6 Conclusion

We have provided a joint analysis of averaging and acceleration for non-strongly convex quadratic functions in a single framework, both with noiseless and noisy gradients. This allows us to define a class of algorithms that can benefit simultaneously from the known improvements of averaging and acceleration: faster forgetting of initial conditions (for acceleration), and better robustness to noise when the noise covariance is proportional to the Hessian (for averaging).

Our current analysis of our class of algorithms in Eq. (2.5), that considers two different affine combinations of previous iterates (instead of one for traditional acceleration), is limited to quadratic functions; an extension of its analysis to all smooth or self-concordant-like functions would widen its applicability. Similarly, an extension to least-squares regression with natural heteroscedastic stochastic gradient, as suggested by our simulations, would be an interesting development. At this point, it is tempting to consider algorithms which use the last three iterates rather than the last two. In particular, we investigate the effect of blending acceleration and averaging in Chapter 3.

Appendix

2.A Additional Experimental Results

In this appendix, we provide additional experimental results to illustrate our theoretical results.

2.A.1 Deterministic Convergence

Comparaison for $d = 1$. In Figure 2.A.1, we minimize a one-dimensional quadratic function $f(\theta) = \frac{1}{2}\theta^2$ for a fixed step-size $\alpha = 1/10$ and different step-sizes β. In the left plot, we compare Acc-GD, HB and Av-GD. We see that HB and Acc-GD both oscillate and that Acc-GD leverages strong convexity to converge faster. In the right plot, we compare the behavior of the algorithm for different values of β. We see that the optimal rate is achieved for $\beta = \beta_*$ defined to be the one for which there is a double coalescent eigenvalue, where the convergence is linear at speed $O(1 - \sqrt{\alpha L})^n$. When $\beta > \beta_*$, we are in the real case and when $\beta < \beta_*$ the algorithm oscillates to the solution.

Comparison between the different eigenspaces. Figure 2.A.2 shows interactions between different eigenspaces. In the left plot, we optimize a quadratic function of dimension $d = 2$. The first eigenvalue is $L = 1$ and the second is $\mu = 2^{-8}$. For Av-GD the convergence is of order $O(1/n)$ since the problem is "not" strongly convex

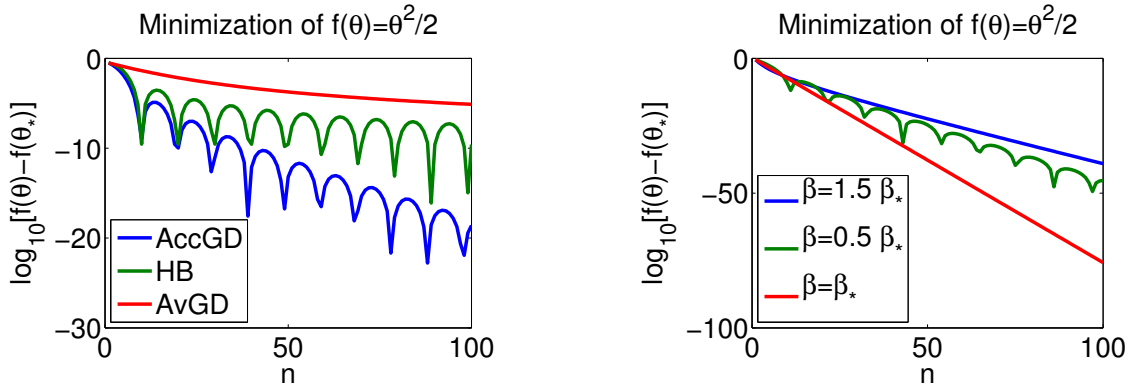


Figure 2.A.1 – Deterministic case for $d = 1$ and $\alpha = 1/10$. Left: classical algorithms, right: different oscillatory behaviors.

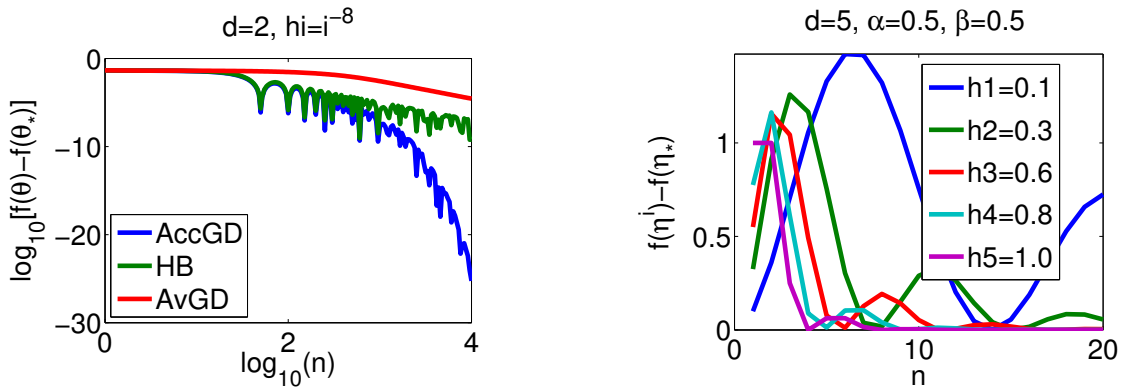


Figure 2.A.2 – Left: Deterministic quadratic optimization for $d = 2$. Right: Function value of the projection of the iterate on the different eigenspaces ($d = 5$).

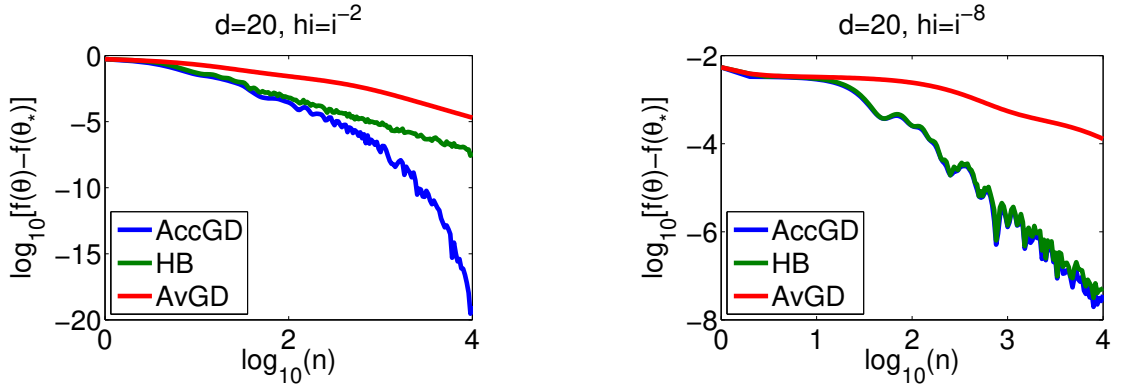


Figure 2.A.3 – Deterministic case for $d = 20$ and $\gamma = 1/10$. Left: $m = 2$. Right: $m = 8$.

(i.e., not appearing as strongly convex since $n\mu$ remains small). The convergence is at the beginning the same for HB and Acc-GD, with oscillation at speed $O(1/n^2)$, since the small eigenvalue prevents Acc-GD from having a linear convergence. Then for large n, the convergence becomes linear for Acc-GD, since μn becomes large. In the right plot, we optimize a quadratic function in dimension $d = 5$ with eigenvalues from 1 to 0.1. We show the function values of the projections of the iterates η_n on the different eigenspaces. We see that high eigenvalues first dominate, but converge quickly to zero, whereas small ones keep oscillating, and converge more slowly.

Comparison for $d = 20$. In Figure 2.A.3, we optimize two 20-dimensional quadratic functions with different eigenvalues with Av-GD, HB and Acc-GD for a fixed step-size $\gamma = 1/10$. In the left plot, the eigenvalues are $1/k^2$ and in the right one, they are $1/k^8$, for $k = 1, \dots, d$. We see that in both cases, Av-GD converges at a rate of $O(1/n)$ and HB at a rate of $O(1/n^2)$. For Acc-GD the convergence is linear when μ is large (left plot) and becomes sublinear at a rate of $O(1/n^2)$ when μ becomes small (right plot).

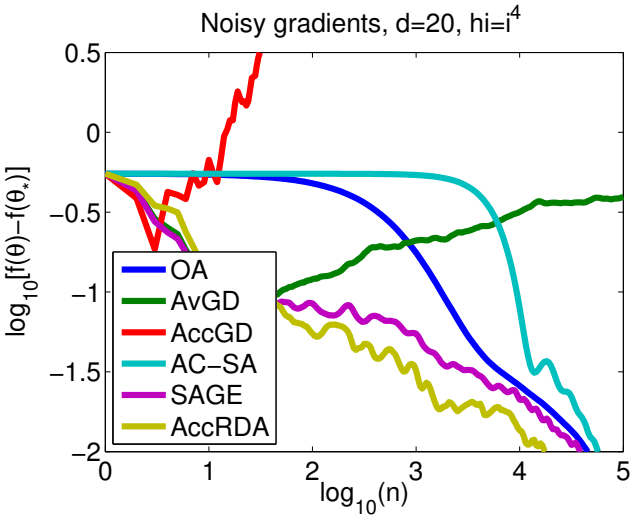


Figure 2.A.4 – Quadratic optimization with additive noise.

2.A.2 Noisy Convergence with Unstructured Additive Noise

We optimize the same quadratic function, but now with noisy gradients. We compare our algorithm to other stochastic accelerated algorithms, that is, AC-SA [Lan, 2012], SAGE [Hu et al., 2009] and Acc-RDA [Xiao, 2010], which are presented in Appendix 2.G. For all these algorithms (and ours) we take the optimal step-sizes defined in these papers. We plot the results averaged over 10 replications.

We consider in Figure 2.A.4 an i.i.d. zero mean noise of variance $C = I$. We see that all the accelerated algorithms achieve the same precision whereas Av-GD with constant step-size does not converge and Acc-Gd diverges. However SAGE and AC-SA are anytime algorithms and are faster at the beginning since their step-sizes are decreasing and not a constant (with respect to n) function of the horizon N.

2.B Proofs of Section 2.2

2.B.1 Proof of Theorem 2

Let $(P_n, Q_n, R_n) \in (\mathbb{R}[X])^3$ for all $n \in \mathbb{N}$ be a sequence of polynomials. We consider the iterates defined for all $n \in \mathbb{N}^*$ by

$$\theta_{n+1} = P_n(H)\theta_n + Q_n(H)\theta_{n-1} + R(H)q,$$

started from $\theta_0 = \theta_1 \in \mathbb{R}^d$. The θ_*-stationarity property gives for $n \in \mathbb{N}^*$:

$$\theta_* = P_n(H)\theta_* + Q_n(H)\theta_* + R_n(H)q.$$

Since $\theta_* = H^{-1}q$ we get for all $q \in \mathbb{R}^d$

$$H^{-1}q = P_n(H)H^{-1}q + Q_n(H)H^{-1}q + R_n(H)q.$$

For all $\tilde{q} \in \mathbb{R}^d$ we apply this relation to vectors $q = H\tilde{q}$:

$$\tilde{q} = P_n(H)\tilde{q} + Q_n(H)\tilde{q} + R_n(H)H\tilde{q} \quad \forall \tilde{q} \in \mathbb{R}^d,$$

and we get

$$I = P_n(H) + Q_n(H) + R_n(H)H \quad \forall n \in \mathbb{N}^*.$$

Therefore there are polynomials $(\bar{P}_n, \bar{Q}_n) \in (\mathbb{R}[X])^2$ and $q_n \in \mathbb{R}$ for all $n \in \mathbb{N}^*$ such that we have for all $n \in \mathbb{N}$:

$$\begin{aligned} P_n(X) &= (1-q_n)I + X\bar{P}_n(X) \\ Q_n(X) &= q_n I + X\bar{Q}_n(X) \\ R_n(X) &= -(\bar{P}_n(X) + \bar{Q}_n(X)). \end{aligned} \tag{2.14}$$

The n-scalability property means that there are polynomials $(P, Q) \in (\mathbb{R}[X])^2$ independent of n such that:

$$\begin{aligned} P_n(X) &= \frac{n}{n+1}P(X), \\ Q_n(X) &= \frac{n-1}{n+1}Q(X). \end{aligned}$$

And in connection with Eq. (2.14) we can rewrite P and Q as:

$$\begin{aligned} P(X) &= \bar{p} + X\bar{P}(X), \\ Q(X) &= \bar{q} + X\bar{Q}(X), \end{aligned}$$

with $(\bar{p}, \bar{q}) \in \mathbb{R}^2$ and $(\bar{P}, \bar{Q}) \in (\mathbb{R}[X])^2$. Thus for all $n \in \mathbb{N}$:

$$q_n = \frac{n-1}{n+1}\bar{q} \tag{2.15}$$

$$\begin{aligned} \bar{Q}_n(X) &= \frac{n-1}{n}Q(X) \\ \frac{n}{n+1}\bar{p} &= (1-q_n) \end{aligned} \tag{2.16}$$

$$\bar{P}_n(X) = \frac{n}{n+1}P(X).$$

Eq. (2.15) and Eq. (2.16) give:

$$\frac{n}{n+1}\bar{p} = \left(1 - \frac{n-1}{n+1}\bar{q}\right).$$

Thus for $n = 1$, we have $\bar{p} = 2$. Then $-\frac{n-1}{n+1}\bar{q} = \frac{2n}{n+1} - 1 = \frac{n-1}{n+1}$ and $\bar{q} = -1$. Therefore

$$P_n(H) = \frac{2n}{n+1}I + \frac{n}{n+1}\bar{P}(H)H$$

$$
\begin{aligned}
Q_n(H) &= -\frac{n-1}{n} I + \bar{Q}(H) H \\
R_n(H) &= -\left(\frac{n\bar{P}(H) + (n-1)\bar{Q}(H)}{n+1} \right).
\end{aligned}
$$

We let $\bar{A} = -(\bar{P} + \bar{Q})$ and $\bar{B} = \bar{Q}$ so that we have:

$$
\begin{aligned}
P_n(H) &= \frac{2n}{n+1} \left(I - \left(\frac{\bar{A}(H) + \bar{B}(H)}{2} \right) H \right) \\
Q_n(H) &= -\frac{n-1}{n} \left(I - \bar{B}(H) H \right) \\
R_n(H) &= \left(\frac{n\bar{A}(H) + \bar{B}(H)}{n+1} \right),
\end{aligned}
$$

and with $\phi_n = \theta_n - \theta_*$ for all $n \in \mathbb{N}$, the algorithm can be written under the form:

$$
\phi_{n+1} = \left[I - \left(\frac{n}{n+1} \bar{A}(H) + \frac{1}{n+1} \bar{B}(H) \right) H \right] \phi_n + \left(1 - \frac{2}{n+1} \right) \left[I - \bar{B}(H) H \right] (\phi_n - \phi_{n-1}).
$$

2.B.2 Av-GD as Two-Steps Algorithm

We show now that when the averaged iterate of Av-GD is seen as the main iterate we have that Av-GD with step-size $\gamma \in \mathbb{R}$ is equivalent to:

$$
\theta_{n+1} = \frac{2n}{n+1} \theta_n - \frac{n-1}{n+1} \theta_{n-1} - \frac{\gamma}{n+1} f'\big(n\theta_n - (n-1)\theta_{n-1} \big).
$$

We remind

$$
\begin{aligned}
\psi_{n+1} &= \psi_n - \gamma f'(\psi_n), \\
\theta_{n+1} &= \theta_n + \frac{1}{n+1} (\psi_{n+1} - \theta_n).
\end{aligned}
$$

Thus, we have:

$$
\begin{aligned}
\theta_{n+1} &= \theta_n + \frac{1}{n+1} (\psi_{n+1} - \theta_n) \\
&= \theta_n + \frac{1}{n+1} (\psi_n - \gamma f'(\psi_n) - \theta_n) \\
&= \theta_n + \frac{1}{n+1} (\theta_n + (n-1)(\theta_n - \theta_{n-1}) - \gamma f'(\theta_n + (n-1)(\theta_n - \theta_{n-1})) - \theta_n) \\
&= \frac{2n}{n+1} \theta_n - \frac{n-1}{n+1} \theta_{n-1} - \frac{\gamma}{n+1} f'(n\theta_n - (n-1)\theta_{n-1}).
\end{aligned}
$$

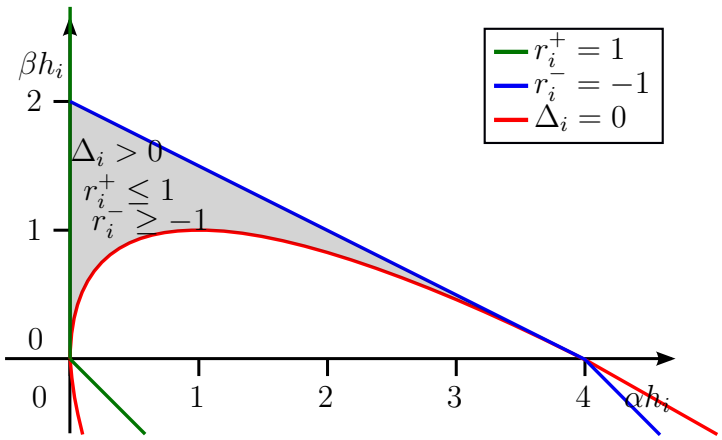


Figure 2.C.1 – Stability in the real case, with all contraints plotted.

2.C Proof of Section 2.3

2.C.1 Proof of Lemma 1

The discriminant Δ_i is strictly positive when $\left(\frac{\alpha+\beta}{2}\right)^2 h_i - \alpha > 0$. This is always true for α strictly negative. For α positive and for $h_i \neq 0$, this is true for $|\frac{\alpha+\beta}{2}| > \sqrt{\alpha/h_i}$. Thus the discriminant Δ_i is strictly positive for

$$\begin{array}{lll} \alpha < 0 & \text{or} & \\ \alpha \geq 0 & \text{and} & \left\{\beta < -\alpha - 2\sqrt{\alpha/h_i} \quad \text{or} \quad \beta > -\alpha + 2\sqrt{\alpha/h_i}\right\}. \end{array}$$

Then we determine when the modulus of the eigenvalues is less than one (which corresponds to $-1 \leq r_i^- \leq r_i^+ \leq 1$).

$$\begin{aligned} r_i^+ \leq 1 \quad &\Leftrightarrow \quad \sqrt{h_i\left(\left(\frac{\alpha+\beta}{2}\right)^2 h_i - \alpha\right)} \leq \left(\frac{\beta+\alpha}{2}\right) h_i \\ &\Leftrightarrow \quad h_i\left(\left(\frac{\beta+\alpha}{2}\right)^2 h_i - \alpha\right) \leq \left[\left(\frac{\beta+\alpha}{2}\right) h_i\right]^2 \quad \text{and} \quad \frac{\alpha+\beta}{2} \geq 0 \\ &\Leftrightarrow \quad h_i\alpha \geq 0 \quad \text{and} \quad \frac{\alpha+\beta}{2} \geq 0 \\ &\Leftrightarrow \quad \alpha \geq 0 \quad \text{and} \quad \alpha+\beta \geq 0. \end{aligned}$$

Moreover, we have :

$$r_i^- \geq -1 \quad \Leftrightarrow \quad \sqrt{h_i\left(\left(\frac{\beta+\alpha}{2}\right)^2 h_i - \alpha\right)} \leq 2 - \left(\frac{\beta+\alpha}{2}\right) h_i$$

$$
\begin{aligned}
&\Leftrightarrow \quad h_i\left(\left(\frac{\beta+\alpha}{2}\right)^2 h_i-\alpha\right) \leq \left[2-\left(\frac{\beta+\alpha}{2}\right)h_i\right]^2 \text{ and } 2-\left(\frac{\beta+\alpha}{2}\right)h_i \geq 0\\
&\Leftrightarrow \quad h_i\left(\left(\frac{\beta+\alpha}{2}\right)^2 h_i-\alpha\right) \leq 4-4\left(\frac{\beta+\alpha}{2}\right)h_i+\left[\left(\frac{\beta+\alpha}{2}\right)h_i\right]^2\\
&\qquad \text{and } \left(\frac{\beta+\alpha}{2}\right) \leq 2/h_i\\
&\Leftrightarrow \quad -h_i\alpha \leq 4-4\left(\frac{\beta+\alpha}{2}\right)h_i \text{ and } \frac{\beta}{2} \leq 2/h_i-\frac{\alpha}{2}\\
&\Leftrightarrow \quad \beta \leq 2/h_i-\alpha/2 \text{ and } \beta \leq 4/h_i-\alpha.
\end{aligned}
$$

Figure 2.C.1 (where we plot all the constraints we have so far) enables to conclude that the discriminant Δ_i is strictly positive and the algorithm is stable when the following three conditions are satisfied:

$$
\begin{aligned}
\alpha &\geq 0\\
\alpha+2\beta &\leq 4/h_i\\
\alpha+\beta &\geq 2\sqrt{\alpha/h_1}.
\end{aligned}
$$

For any of those α et β we will have:

$$\eta_n^i = c_1(r_i^-)^n + c_2(r_i^+)^n.$$

Since $\eta_0^i = 0$, $c_1 + c_2 = 0$ and for $n = 1$, $c_1 = \eta_1^i/(r_i^- - r_i^+)$; we thus have:

$$\eta_n^i = \frac{\eta_1^i}{2}\frac{(r_i^+)^n - (r_i^-)^n}{\sqrt{\Delta_i}}.$$

Thus, we get the final expression:

$$(\phi_n^i)^2 h_i = \frac{(\phi_1^i)^2}{4n^2}\frac{\left\{\left[r_i+\sqrt{\Delta_i}\right]^n - \left[r_i-\sqrt{\Delta_i}\right]^n\right\}^2}{\Delta_i/h_i}.$$

2.C.2 Proof of Lemma 2

The discriminant Δ_i is strictly negative if and only if $\left(\frac{\alpha+\beta}{2}\right)^2 h_i - \alpha < 0$. This implies $|\frac{\alpha+\beta}{2}| < \sqrt{\alpha/h_i}$. The modulus of the eigenvalues is $|r_i^{\pm}|^2 = 1-\beta h_i$. Thus the discriminant Δ_i is strictly negative and the algorithm is stable for

$$
\begin{aligned}
\alpha, \beta &\geq 0\\
\alpha+\beta &< \sqrt{\alpha/h_i},
\end{aligned}
$$

as shown in Figure 2.C.2.

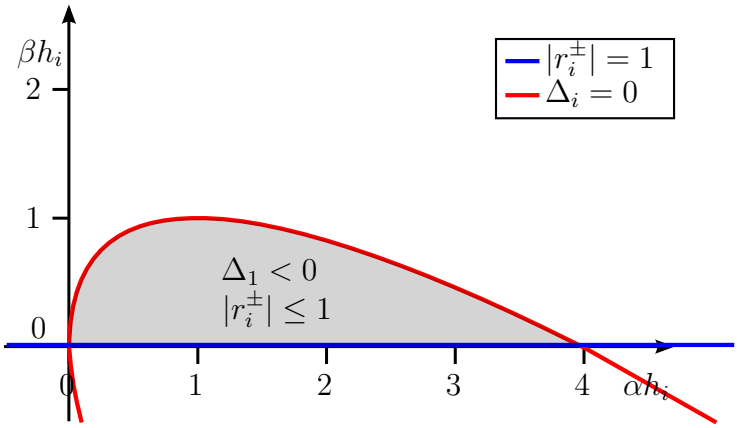


Figure 2.C.2 – Stability in the complex case, with all constraints plotted.

For any of those α et β we have:

$$\eta_n^i = [c_1 \cos(\omega_i n) + c_2 \sin(\omega_i n)]\rho_i^n,$$

with $\rho_i = \sqrt{1-\beta h_i}$, $\sin(\omega_i) = \sqrt{-\Delta_i}/\rho_i$ and $\cos(\omega_i) = r_i/\rho_i$. Since $\eta_0^i = 0$, $c_1 = 0$ and we have for $n = 1$, $c_2 = \eta_1^i/(\sin(\omega_i)\rho_i)$. Therefore

$$\eta_n^i = \eta_1^i \frac{\sin(\omega_i n)}{\sqrt{-\Delta_i}}(1-\beta h_i)^{n/2},$$

and

$$(\phi_n^i)^2 h_i = \frac{(\phi_1^i)^2}{n^2}\frac{\sin^2(\omega_i n)}{\sin^2(\omega_i)/h_i}(1-\beta h_i)^{n-1}.$$

2.C.3 Coalescing Eigenvalues

When $\beta = 2\sqrt{\alpha/h_i} - \alpha$, the discriminant Δ_i is equal to zero and we have a double real eigenvalue:

$$r_i = 1 - \sqrt{\alpha h_i}.$$

Thus the algorithm is stable for $\alpha < \frac{4}{h_i}$. For any of those α et β we have:

$$\eta_n^i = (c_1 + n c_2) r^n.$$

This gives with $\eta_0^i = 0$, $c_1 = 0$ and $c_2 = \eta_1^i/r$. Therefore

$$\eta_n^i = n\eta_i(1-\sqrt{\alpha h_i})^{n-1},$$

and:

$$(\phi_n^i)^2 h_i = h_i(\phi_1^i)^2(1-\sqrt{\alpha h_i})^{2(n-1)}.$$

In the presence of coalescing eigenvalues the convergence is linear if $0 < \alpha < 4/h_i$ and $h_i > 0$, however one might worry about the behavior of $((\phi_n^i)^2 h_i)_n$ when h_i

becomes small. Using the bound $x^2 \exp(-x) \leq 1$ for $x \leq 1$, we have for $\alpha < 4/h_i$:

$$\begin{aligned} h_i(1-\sqrt{\alpha h_i})^{2n} &= h_i \exp(2n \log(|1-\sqrt{\alpha h_i}|)) \\ &\leq hi \exp(-2n \min\{\sqrt{\alpha h_i}, 2-\sqrt{\alpha h_i}\}) \\ &\leq \frac{h_i}{\min\{\sqrt{\alpha h_i}, 2-\sqrt{\alpha h_i}\}^2} \\ &\leq \max\left\{\frac{1}{\alpha}, \frac{h_i}{(2-\sqrt{\alpha h_i})^2}\right\}. \end{aligned}$$

Therefore we always have the following bound for $\alpha < 4/h_i$:

$$(\phi_n^i)^2 h_i \leq \frac{(\phi_1^i)^2}{4n^2} \max\left\{\frac{1}{\alpha}, \frac{h_i}{(2-\sqrt{\alpha h_i})^2}\right\}.$$

Thus for $\alpha h_i \leq 1$ we get:

$$(\phi_n^i)^2 h_i \leq \frac{(\phi_1^i)^2}{4n^2\alpha}.$$

2.D Proof of Theorem 3

2.D.1 Sketch of the Proof

We divide the domain of validity of Theorem 3 in three subdomains as explained in Figure 2.D.4. On the domain described in Figure 2.D.1 we have a first bound on the iterate η_n^i:

Lemma 3. *For $0 \leq \alpha \leq 1/h_i$ and $1-\sqrt{1-\alpha h_i} < \beta h_i < 1+\sqrt{1-\alpha h_i}$, we have:*

$$(\eta_n^i)^2 \leq \frac{(\eta_1^i)^2}{\alpha h_i}.$$

And on the domain described Figure 2.D.2 we also have:

Lemma 4. *For $0 \leq \alpha \leq 1/h_i$ and $\beta \leq \alpha$ we have:*

$$(\eta_n^i)^2 \leq \frac{2(\eta_1^i)^2}{\alpha h_i}.$$

These two lemmas enable us to prove the first bound of Theorem 3 since the domain of this theorem is included in the intersection of the two domains of these lemmas as shown in Figure 2.D.4.

Then we have the following bound on domain described in Figure 2.D.3:

Lemma 5. *For $0 \leq \alpha \leq 2/h_i$ and $0 \leq \beta \leq 2/h_i - \alpha$, we have:*

$$|\eta_n^i| \leq \min\left\{\frac{2\sqrt{2n}}{\sqrt{(\alpha+\beta)h_i}}, \frac{4}{(\alpha+\beta)h_i}\right\}.$$

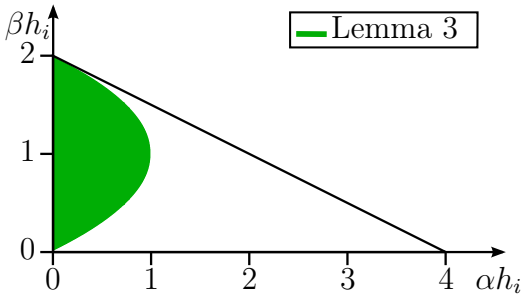


Figure 2.D.1 – Validity of Lemma 3

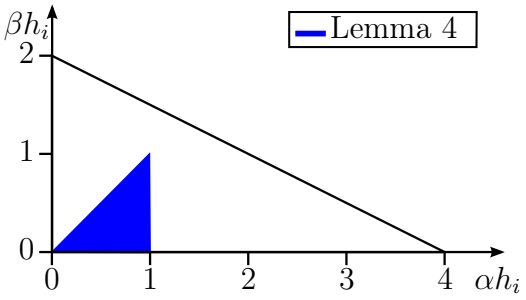


Figure 2.D.2 – Validity of Lemma 4

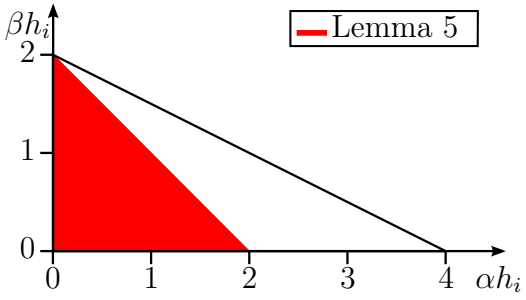


Figure 2.D.3 – Validity of Lemma 5

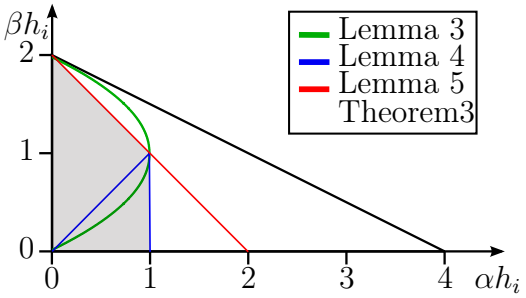


Figure 2.D.4 – Area of Theorem 3

Since the domain of definition of Theorem 3 is included in the domain of definition of Lemma 5 (as shown in Figure 2.D.4), this lemma proves the last two bounds of the theorem.

2.D.2 Outline of the Proofs of the Lemmas

- We find a Lyapunov function G from $\mathbb{R}^2$ to $\mathbb{R}$ such that the sequence $(G(\eta_n^i, \eta_{n-1}^i))$ decrease along the iterates.
- We also prove that $G(\eta_n^i, \eta_{n-1}^i)$ dominates $c\|\eta_n^i\|^2$ when we want to have a bound on $\|\eta_n^i\|^2$ of the form $\frac{1}{c}G(\eta_1^i, \eta_0^i) = \frac{1}{c}G(\theta_0^i - \theta_*^i, 0)$.

For readability, we remove the index i and take $h_i = 1$ without loss of generality.

2.D.3 Proof of Lemma 3

We first consider a quadratic Lyapunov function $\begin{pmatrix} \eta_n \\ \eta_{n-1} \end{pmatrix}^\top G_1 \begin{pmatrix} \eta_n \\ \eta_{n-1} \end{pmatrix}$ with $G_1 = \begin{pmatrix} 1 & \alpha - 1 \\ \alpha - 1 & 1 - \alpha \end{pmatrix}$. We note that G_1 is symmetric positive semi-definite for $\alpha \leq 1$. We recall $F_i = \begin{pmatrix} 2 - (\alpha + \beta) & \beta - 1 \\ 1 & 0 \end{pmatrix}$.

For the result to be true we need for $0 \leq \alpha \leq 1$ and $1-\sqrt{1-\alpha} < \beta < 1+\sqrt{1-\alpha}$ two properties:

$$F_i^\top G_1 F_i \preccurlyeq G_1, \tag{2.17}$$

and

$$\alpha \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \preccurlyeq G_1. \tag{2.18}$$

Proof of Eq. (2.18). We have:

$$G_1 - \alpha \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = (1-\alpha) \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \succcurlyeq 0 \quad \text{for } \alpha \leq 1.$$

Proof of Eq. (2.17). Since $\beta \mapsto F_i(\beta)^\top G_1 F_i(\beta) - G_1$ is convex in β (G_1 is symmetric positive semi-definite), we only have to show Eq. (2.17) for the boundaries of the interval in β. For $x \in \mathbb{R}_+^*$:

$$\begin{pmatrix} x^2-x & x \\ 1 & 0 \end{pmatrix}^\top \begin{pmatrix} 1 & -x^2 \\ -x^2 & x^2 \end{pmatrix} \begin{pmatrix} x^2-x & x \\ 1 & 0 \end{pmatrix} - \begin{pmatrix} 1 & -x^2 \\ -x^2 & x^2 \end{pmatrix} = -(1-x^2)^2 \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \preccurlyeq 0.$$

This especially shows Eq. (2.17) for the boundaries of the interval with $x = \pm\sqrt{1-\alpha}$.

Bound. Thus, because $\eta_0 = 0$, we have

$$\alpha\eta_{n+1}^2 \leq \Theta_n^\top G_1 \Theta_n \leq \Theta_{n-1}^\top G_1 \Theta_{n-1} \leq \Theta_0^\top G_1 \Theta_0 \leq \eta_1^2.$$

This shows that for $0 \leq \alpha \leq 1/h_i$ and $1-\sqrt{1-\alpha h_i} < \beta h_i < 1+\sqrt{1-\alpha h_i}$:

$$(\eta_n^i)^2 \leq \frac{(\eta_1^i)^2}{\alpha h_i}.$$

2.D.4 Proof of Lemma 4

We consider now a second Lyapunov function $G_2(\eta_n, \eta_{n-1}) = (\eta_n - r\eta_{n-1})^2 - \Delta(\eta_{n-1})^2$. We have:

$$\begin{aligned}
G_2(\eta_n, \eta_{n-1}) &= (\eta_n - r\eta_{n-1})^2 - \Delta\eta_{n-1}^2 \\
&= (r\eta_{n-1} - (1-\beta)\eta_{n-2})^2 - \Delta\eta_{n-1}^2 \\
&= (r^2-\Delta)\eta_{n-1}^2 + (1-\beta)^2\eta_{n-2}^2 - 2(1-\beta)r\eta_{n-1}\eta_{n-2} \\
&= ((1-\beta)\eta_{n-1}^2 + (1-\beta)(r^2-\Delta)\eta_{n-2}^2 - 2(1-\beta)r\eta_{n-1}\eta_{n-2} \\
&= (1-\beta)[(\eta_{n-1} - r\eta_{n-2})^2 - \Delta(\eta_{n-2})^2]. \\
&= (1-\beta)G_2(\eta_{n-1}, \eta_{n-2}).
\end{aligned}$$

Where we have used twice $r^2 - \Delta = (1-\beta)$ and $\eta_n = 2r\eta_{n-1} - (1-\beta)\eta_{n-2}$. Moreover $G_2(\eta_n, \eta_{n-1})$ can be rewritten as:

$$G_2(\eta_n, \eta_{n-1}) = (1 - \frac{\alpha+\beta}{2})(\eta_n - \eta_{n-1})^2 + \frac{\alpha-\beta}{2}(\eta_{n-1})^2 + \frac{\alpha+\beta}{2}(\eta_n)^2.$$

Thus for $\alpha + \beta \leq 2$ and $\beta \leq \alpha$ we have:

$$\frac{\alpha}{2}(\eta_n)^2 \leq G_2(\eta_n, \eta_{n-1}) = (1-\beta)^{n-1} G_2(\eta_1, \eta_0) = (1-\beta)^{n-1}(\eta_1)^2.$$

Therefore for $\alpha + \beta \leq 2/h_i$ and $\beta \leq \alpha$, we have:

$$(\eta_n^i)^2 \leq \frac{2(\eta_1^i)^2}{\alpha h_i}.$$

2.D.5 Proof of Lemma 5

We may write η_n as

$$\eta_n = r\eta_{n-1} + (r_+)^n + (r_-)^n.$$

Moreover, we have:

$$|(r_+)^n + (r_-)^n| \leq 2,$$

therefore for $\alpha + \beta \leq 2$,

$$|\eta_n| \leq r|\eta_{n-1}| + 2 \leq 2\frac{1-r^n}{1-r} \leq 2\frac{1-(1-(\frac{\alpha+\beta}{2}))^n}{(\frac{\alpha+\beta}{2})}.$$

Thus

$$|\eta_n| \leq \frac{2}{(\frac{\alpha+\beta}{2})h}.$$

Moreover for all $u \in [0,1]$ and $n \geq 1$ we have $1-(1-u)^n \leq \sqrt{nu}$, since $1-(1-u)^n \leq 1$ and $1-(1-u)^n = u\sum(1-u)^k \leq nu$. Thus

$$|\eta_n| \leq \frac{2\sqrt{n}}{\sqrt{(\frac{\alpha+\beta}{2})}}.$$

Therefore for $0 \leq \alpha \leq 2/h_i$ and $\alpha + \beta \leq 2/h_i$ we have:

$$|\eta_n^i| \leq \min\left\{\frac{2\sqrt{2n}}{\sqrt{(\alpha+\beta)h_i}}, \frac{4}{(\alpha+\beta)h_i}\right\}.$$

2.E Lower Bounds

We have the following lower-bound for the bound shown in Corollary 1, which shows that depending on which of the two terms dominates, we may always find a

sequence of functions that makes it tight.

Proposition 5. *Let $L \geq 0$. For all sequences $0 \leq \alpha_n \leq 1/L$ and $0 \leq \beta_n \leq 2/L - \alpha_n$, such that $\alpha_n + \beta_n = o(n\alpha_n)$ there exists a sequence of one-dimensional quadratic functions $(f_n)_n$ with second-derivative less than L such that:*

$$\lim \alpha_n n^2 (f_n(\theta_n) - f_n(\theta_*)) = \frac{\|\theta_0 - \theta_*\|^2}{2}.$$

For all sequences $0 \leq \alpha_n \leq 1/L$ and $0 \leq \beta_n \leq 2/L - \alpha_n$, such that $n\alpha_n = o(\alpha_n + \beta_n)$, there exists a sequence of one-dimensional quadratic functions $(g_n)_n$ with second-derivative less than L such that:

$$\lim n(\alpha_n + \beta_n)(g_n(\theta_n) - g_n(\theta_*)) = \frac{(1 - \exp(-2))^2 \|\theta_0 - \theta_*\|^2}{4}.$$

Proof of the first lower-bound. For the first lower bound we consider $0 \leq \alpha_n \leq 1/L$ and $0 \leq \beta_n \leq 2/L - \alpha$, such that $\alpha_n + \beta_n = o(n\alpha_n)$. We define $f_n = \pi^2/(4\alpha_n n^2)$ and we consider the sequence of quadratic functions $f_n(\theta) = \frac{f_n \theta^2}{2}$. We consider the iterate $(\eta_n)_n$ defined by our algorithm. We will show that

$$\lim \alpha_n f_n(\eta_n) = \frac{\eta_1^2}{2}.$$

We have, from Lemma 2,

$$f_n(\eta_n) = \frac{\eta_n^2 f_n}{2} = \frac{\eta_1^2 \sin^2(\omega_n n) \rho_n^{2n}}{2\alpha_n (1 - \frac{\pi^2(\alpha_n + \beta_n)^2}{(4\alpha_n n)^2})}.$$

Moreover,

$$\rho_n^{2n} = \left(1 - \frac{\beta_n \pi^2}{4\alpha_n n^2}\right)^n = \exp\left(n \log\left(1 - \frac{\beta_n \pi^2}{4\alpha_n n^2}\right)\right) = 1 + o(1),$$

since $\frac{\beta_n}{\alpha_n n} = o(1)$. Also, $1 - \frac{\pi^2(\alpha_n + \beta_n)^2}{(4\alpha_n n)^2} = 1 + o(1)$, since $\alpha_n + \beta_n = o(n\alpha_n)$. Moreover

$$\sin(\omega_n) = \frac{\sqrt{-\Delta_n}}{\rho_n} = \frac{\sqrt{f_n}\sqrt{\alpha_n - \frac{(\alpha_n + \beta_n)^2}{4} f_n}}{\sqrt{1 - \beta_n f_n}} = \pi/(2n) + o(1/n),$$

thus $\omega_n = \pi/(2n) + o(1/n)$ and $\sin(n\omega_n) = 1 + o(1)$.

Proof of the second lower-bound. We consider now the situation where the second bound is active. Thus we take sequences (α_n) and (β_n), such that $n\alpha_n = o(\alpha_n + \beta_n)$. We define $g_n = \frac{2}{n(\alpha_n + \beta_n)} + \frac{4\alpha_n}{(\alpha_n + \beta_n)^2}$ and consider the sequence of quadratic functions $g_n(\theta) = \frac{g_n \theta^2}{2}$. We will show for the iterate (η_n) defined by our algorithm

that:

$$\lim n(\alpha_n + \beta_n)(g_n(\theta_n) - g_n(\theta_*)) = \frac{(1 - \exp(-2))^2 \|\theta_0 - \theta_*\|^2}{4}.$$

We will use Lemma 1. We first have

$$\Delta_n = \left(\frac{\alpha_n + \beta_n}{2}\right)^2 g_n^2 - \alpha_n g_n = g_n \left(\frac{\alpha_n + \beta_n}{2}\right) \frac{1}{n}.$$

Thus $(n\Delta_n)/g_n = \left(\frac{\alpha_n+\beta_n}{2}\right)$ and

$$\begin{aligned}
\sqrt{\Delta_n} &= \sqrt{\left(\frac{1}{n}\right)^2 + \frac{2\alpha_n}{n(\alpha_n + \beta_n)}} \\
&= \frac{1}{n}\sqrt{1 + \frac{2\alpha_n n}{\alpha_n + \beta_n}} \\
&= \frac{1}{n} + \frac{\alpha_n}{\alpha_n + \beta_n} + o\left(\frac{\alpha_n}{\alpha_n + \beta_n}\right).
\end{aligned}$$

Moreover

$$r_n = 1 - \frac{\alpha_n + \beta_n}{2} g_n = 1 - \frac{1}{n} - \frac{2\alpha_n}{\alpha_n + \beta_n}.$$

Thus

$$r_+ = 1 - \frac{\alpha_n}{\alpha_{n+}\beta_n} + o\left(\frac{\alpha_n}{\alpha_n + \beta_n}\right),$$

and

$$r_+^n = \exp(n \log(r_+)) = \exp\left(-\frac{n\alpha_n}{\alpha_n + \beta_n}\right) + o\left(\frac{n\alpha_n}{\alpha_n + \beta_n}\right) = 1 + o(1).$$

Furthermore

$$r_- = 1 - \frac{2}{n} - \frac{3\alpha_n}{\alpha_{n+}\beta_n} + o\left(\frac{\alpha_n}{\alpha_n + \beta_n}\right),$$

and

$$r_-^n = \exp(n \log(r_+)) = \exp\left(-2 - \frac{3\alpha_n n}{\alpha_n + \beta_n}\right) + o\left(\frac{n\alpha_n}{\alpha_n + \beta_n}\right) = \exp(-2) + o(1).$$

Thus

$$(r_+^n - r_-^n)^2 = (1 - \exp(-2))^2 + o(1).$$

Finally, we have:

$$\begin{aligned}
(\alpha_n + \beta_n) n [g_n(\theta_n) - g_n(\theta_*)] &= \frac{\alpha_n + \beta_n}{2n} \|\theta_0 - \theta_*\|^2 \frac{[r_+^n - r_-^n]^2}{4\Delta_n / g_n} \\
&= \frac{\|\theta_0 - \theta_*\|^2}{4} [r_+^n - r_-^n]^2 \\
&= \frac{\|\theta_0 - \theta_*\|^2}{4} (1 - \exp(-2))^2 + o(1).
\end{aligned}$$

2.F Proofs of Section 2.4

2.F.1 Proofs of Theorem 4 and Theorem 5

We decompose again vectors in an eigenvector basis of H with $\eta_n^i = p_i^\top \eta_n$ and $\varepsilon_n^i = p_i^\top \varepsilon_n$:

$$\eta_{n+1}^i = (1 - \alpha h_i)\eta_n^i + (1 - \beta h_i)(\eta_n^i - \eta_{n-1}^i) + (n\alpha + \beta)\varepsilon_{n+1}^i.$$

We denote by $\xi_{n+1}^i = \begin{pmatrix} [n\alpha + \beta]\varepsilon_{n+1}^i \\ 0 \end{pmatrix}$ and we have the reduced equation:

$$\Theta_{n+1}^i = F_i \Theta_n^i + \xi_{n+1}^i.$$

Unfortunately F_i is not Hermitian and this formulation will not be convenient for calculus. Without loss of generality, we assume $r_i^- \neq r_i^+$ even if it means having $r_i^- - r_i^+$ goes to 0 in the final bound. Let $Q_i = \begin{pmatrix} r_i^- & r_i^+ \\ 1 & 1 \end{pmatrix}$ be the transfer matrix of F_i, i.e., $F_i = Q_i D_i Q_i^{-1}$ with $D_i = \begin{pmatrix} r_i^- & 0 \\ 0 & r_i^+ \end{pmatrix}$ and $Q_i^{-1} = \frac{1}{r_i^- - r_i^+} \begin{pmatrix} 1 & -r_i^+ \\ -1 & r_i^- \end{pmatrix}$. We can reparametrize the problem in the following way:

$$\begin{aligned} Q_i^{-1}\Theta_{n+1}^i &= Q_i^{-1} F_i \Theta_n^i + Q_i^{-1}\xi_{n+1}^i \\ &= Q_i^{-1} F_i Q_i Q_i^{-1} \Theta_n^i + Q_i^{-1}\xi_{n+1}^i \\ &= D_i (Q_i^{-1}\Theta_n^i) + Q_i^{-1}\xi_{n+1}^i. \end{aligned}$$

With $\tilde{\Theta}_n^i = Q_i^{-1}\Theta_n^i$ and $\tilde{\xi}_n^i = Q_i^{-1}\xi_n^i$ we now have:

$$\tilde{\Theta}_{n+1}^i = D_i \tilde{\Theta}_n^i + \tilde{\xi}_{n+1}^i, \tag{2.19}$$

with now D_i Hermitian (even diagonal).

Thus it is easier to tackle using standard techniques for stochastic approximation [see, e.g., Polyak and Juditsky, 1992, Bach and Moulines, 2011]:

$$\tilde{\Theta}_n^i = D_i^{n-1}\tilde{\Theta}_1^i + \sum_{k=2}^{n} D_i^{n-k}\tilde{\xi}_k^i.$$

Let $M_i = \begin{pmatrix} h_i^{1/2} & h_i^{1/2} \\ 0 & 0 \end{pmatrix}$, we then get using standard martingale square moment inequalities, since for $n \neq m$, ε_n^i and ε_m^i are uncorrelated (i.e., $\mathbb{E}[\varepsilon_n^i \varepsilon_m^i] = 0$):

$$\mathbb{E}\|M_i\tilde{\Theta}_n^i\|^2 = \|M_i D_i^{n-1}\tilde{\Theta}_1^i\|^2 + \mathbb{E}\sum_{k=2}^{n} \|M_i D_i^{n-k}\tilde{\xi}_k^i\|^2.$$

This is a bias-variance decomposition; the left term only depends on the initial con-

dition and the right term only depends on the noise process.

We have with $M_i = \begin{pmatrix} h_i^{1/2} & h_i^{1/2} \\ 0 & 0 \end{pmatrix}$, $M_i Q_i^{-1} = \begin{pmatrix} 0 & h_i^{1/2} \\ 0 & 0 \end{pmatrix}$, and $M_i \tilde{\Theta}_n^i = \begin{pmatrix} \sqrt{h_i}\eta_{n-1}^i \\ 0 \end{pmatrix}$. Thus, we have access to the function values through:

$$\|M_i \tilde{\Theta}_n^i\|^2 = h_i (\eta_{n-1}^i)^2.$$

Moreover we have $\Theta_1^i = \begin{pmatrix} \phi_1^i/(r_i^- - r_i^+) \\ -\phi_1^i/(r_i^- - r_i^+) \end{pmatrix}$. Thus

$$\|M_i D_i^{n-1} \tilde{\Theta}_1^i\|^2 = (\phi_1^i)^2 h_i \frac{\left((r_i^+)^{n-1} - (r_i^-)^{n-1}\right)^2}{(r_i^+ - r_i^-)^2}.$$

This is the bias term we have studied in Section 2.3.3 which we bound with Theorem 3. The variance term is controlled by the next proposition.

Proposition 6. *With $\mathbb{E}[(\varepsilon_n^i)^2] = c_i$ for all $n \in \mathbb{N}$, for $\alpha \leq 1/h_i$ and $0 \leq \beta \leq 2/h_i - \alpha$, we have*

$$\frac{1}{(n-1)^2}\mathbb{E}\sum_{k=2}^{n} \|M_i D_i^{n-k} \tilde{\xi}_k^i\|^2 \leq \min\left\{ \frac{2(\alpha(n-1)+\beta)^2}{\alpha\beta(4-(\alpha+2\beta)h_i)(n-1)^2}\frac{c_i}{h_i}, \right.$$
$$\left. \frac{16((n-1)\alpha+\beta)^2}{(n-1)(\alpha+\beta)^2}\frac{c_i}{h_i}, 2\frac{(\alpha(n-1)+\beta)^2}{(n-1)\alpha}c_i, \frac{8(\alpha(n-1)+\beta)^2}{\alpha+\beta}c_i \right\}.$$

The last two bounds prove Theorem 4.

We note that if we restrict β to $\beta \leq 3/(2h_i) - \alpha/2$, then $4 - (\alpha+2\beta)h_i \geq 1$ and the first bound of Proposition 6 is simplified to $\frac{2(\alpha(n-1)+\beta)^2}{\alpha\beta(n-1)^2}\frac{c_i}{h_i}$. This allows to conclude to prove Theorem 5.

2.F.2 Proof of Corollary 3

We let $\nu = \frac{\|\theta_0 - \theta_*\|}{\sqrt{L\,\mathrm{tr}(CH^{-1})}}$ and consider three different regimes depending on ν and L.

If $\nu < 1/L$, we have $\nu/N < 1/L$ and thus $\alpha = \nu/N$ and $\beta = \nu$. Therefore

$$\begin{aligned} \frac{\|\theta_0 - \theta_*\|^2}{N^2\alpha} + \frac{(\alpha N + \beta)^2}{\alpha\beta N^2}\mathrm{tr}(CH^{-1}) &= \frac{\|\theta_0 - \theta_*\|^2}{\nu N} + \frac{4\,\mathrm{tr}(CH^{-1})}{N} \\ &\leq \frac{\sqrt{L\,\mathrm{tr}(CH^{-1})}\|\theta_0 - \theta_*\|}{N} + \frac{4\,\mathrm{tr}(CH^{-1})}{N} \\ &\leq \frac{5\,\mathrm{tr}(CH^{-1})}{N}, \end{aligned}$$

where we have used $\sqrt{L}\|\theta_0 - \theta_*\| < \sqrt{\mathrm{tr}(CH^{-1})}$ since $\nu < 1/L$.

If $\nu > 1/L$ and $\nu < N/L$, we have $\alpha = \nu/N$ and $\beta = 1/L$. Therefore

$$\begin{aligned}\frac{\|\theta_0-\theta_*\|^2}{N^2\alpha}+\frac{(\alpha N+\beta)^2}{\alpha\beta N^2}\operatorname{tr}(CH^{-1}) &\leq \frac{\|\theta_0-\theta_*\|^2}{\nu N}+\frac{4\operatorname{tr}(CH^{-1})}{L\nu N}\\ &\leq \frac{\sqrt{L\operatorname{tr}(CH^{-1})}\|\theta_0-\theta_*\|}{N}+\frac{4\operatorname{tr}(CH^{-1})}{N}\\ &\leq \frac{5\sqrt{L\operatorname{tr}(CH^{-1})}\|\theta_0-\theta_*\|}{N},\end{aligned}$$

where we have used $\sqrt{L}\|\theta_0-\theta_*\| > \sqrt{\operatorname{tr}(CH^{-1})}$ since $\nu > 1/L$.

If $\nu > N/L$, we have $\alpha = 1/L$ and $\beta = 1/L$. Therefore

$$\begin{aligned}\frac{\|\theta_0-\theta_*\|^2}{N^2\alpha}+\frac{(\alpha(N-1)+\beta)^2}{\alpha\beta N^2}\operatorname{tr}(CH^{-1}) &= \frac{L\|\theta_0-\theta_*\|^2}{N^2}+\operatorname{tr}(CH^{-1})\\ &\leq \frac{L\|\theta_0-\theta_*\|^2}{N^2}+\frac{L\|\theta_0-\theta_*\|^2}{N^2}\\ &\leq \frac{2L\|\theta_0-\theta_*\|^2}{N^2},\end{aligned}$$

where we have used that the real bound in Proposition 6 is in fact in $(N-1)\alpha+\beta$, (see Lemma 6) and that $\operatorname{tr}(CH^{-1}) < \frac{L\|\theta_0-\theta_*\|^2}{N^2}$ since $\nu > N/L$.

2.F.3 Proof of Proposition 6

Proof Outline

To prove Proposition 6 we will use Lemmas 6, 7 and 8, that are stated and proved in Section 2.F.3.

We want to bound $\mathbb{E}[\sum_{k=2}^n \|M_i D_i^{n-k}\tilde{\xi}_k^i\|^2]$ and according to Lemma 6, we have an explicit expansion using the roots of the characteristic polynomial:

$$\mathbb{E}\|M_i D_i^{n-k}\tilde{\xi}_k^i\|^2 = h_i((k-1)\alpha+\beta)^2\mathbb{E}[(\varepsilon^i)^2]\frac{[(r_i^-)^{n-k}-(r_i^+)^{n-k}]^2}{(r_i^- - r_i^+)^2}.$$

Thus, by bouding $(k-1)\alpha+\beta$ by $(n-1)\alpha+\beta$, we get

$$\mathbb{E}\sum_{k=2}^n \|M_i D_i^{n-k}\tilde{\xi}_k^i\|^2 \leq h_i((n-1)\alpha+\beta)^2\mathbb{E}[\varepsilon^{i2}]\sum_{k=2}^n \frac{[(r_i^-)^{n-k}-(r_i^+)^{n-k}]^2}{(r_i^- - r_i^+)^2}. \tag{2.20}$$

Then, we have from Lemma 7 the inequality:

$$\sum_{k=0}^{n-2}\frac{[(r_i^-)^k-(r_i^+)^k]^2}{[(r_i^-)-(r_i^+)]^2} \leq \frac{2-\beta h_i}{4\alpha\beta h_i^2(1-(\frac{1}{4}\alpha+\frac{1}{2}\beta)h_i)}.$$

Therefore

$$\mathbb{E}\sum_{k=2}^{n}\|M_i^{1/2}D_i^{n-k}\tilde{\xi}_k^i\|^2 \leq \frac{\mathbb{E}[{\varepsilon^i}^2]}{h_i}\frac{((n-1)\alpha+\beta)^2}{4\alpha\beta}\frac{2-\beta h_i}{(1-(\frac{1}{4}\alpha+\frac{1}{2}\beta)h_i)}.$$

This allows to prove the first part of the bound. The other parts are much simpler and are done in Lemma 8. Thus, adding these bounds gives for $\alpha \leq 1/h_i$ and $0 \leq \beta \leq 2/h_i - \alpha$:

$$\frac{1}{(n-1)^2}\mathbb{E}\sum_{k=2}^{n}\|M_iD_i^{n-k}\tilde{\xi}_k^i\|^2 \leq \min\left\{\frac{2(\alpha(n-1)+\beta)^2}{\alpha\beta(n-1)^2(4-(\alpha+2\beta)h_i)}\frac{c}{h_i},\right.$$
$$\left.\frac{16((n-1)\alpha+\beta)^2}{(n-1)(\alpha+\beta)^2}\frac{c}{h_i}, 2\frac{(\alpha(n-1)+\beta)^2}{(n-1)\alpha}c_i, \frac{8((n-1)\alpha+\beta)^2}{\alpha+\beta}c_i\right\}.$$

Some Technical Lemmas

We first compute an explicit expansion of the noise term as a function of the eigenvalues of the dynamical system.

Lemma 6. *For all $\alpha \leq 1/h_i$ and $0 \leq \beta \leq 2/h_i - \alpha$ we have*

$$\mathbb{E}\|M_iD_i^{n-k}\tilde{\xi}_k^i\|^2 = h_i((k-1)\alpha+\beta)^2\mathbb{E}[(\varepsilon^i)^2]\frac{[(r_i^-)^{n-k}-(r_i^+)^{n-k}]^2}{(r_i^- - r_i^+)^2}.$$

Proof. We first turn the Euclidean norm into a trace, using that $\mathrm{tr}[AB] = \mathrm{tr}[BA]$ for two matrices A and B and that $\mathrm{tr}[x] = x$ for a real x.

$$\mathbb{E}\|M_iD_i^{n-k}\tilde{\xi}_k^i\|^2 = \mathrm{Tr}D_i^{n-k}{M_i}^\top M_iD_i^{n-k}\mathbb{E}[\tilde{\xi}_k^i(\tilde{\xi}_k^i)^\top], \tag{2.21}$$

This enables us to separate the noise term from the rest of the formula. Then we compute the latter from the definition of $\tilde{\xi}_k^i$ in Eq. (2.19) :

$$\mathbb{E}[\tilde{\xi}_k^i(\tilde{\xi}_k^i)^\top] = \frac{((k-1)\alpha+\beta)^2}{(r_i^- - r_i^+)^2}\mathbb{E}[(\varepsilon^i)^2]\begin{pmatrix}1 & -1\\ -1 & 1\end{pmatrix}.$$

And the first part of Eq. (2.21) is equal to:

$$D_i^{n-k}{M_i}^\top M_iD_i^{n-k} = h_i\begin{pmatrix}(r_i^-)^{2(n-k)} & (r_i^-)^{(n-k)} - (r_i^+)^{(n-k)}\\ (r_i^-)^{(n-k)} - (r_i^+)^{(n-k)} & (r_i^+)^{2(n-k)}\end{pmatrix},$$

because $D_i = \begin{pmatrix}r_i^- & 0\\ 0 & r_i^+\end{pmatrix}$ and $M_i = \begin{pmatrix}h_i^{1/2} & h_i^{1/2}\\ 0 & 0\end{pmatrix}$. Therefore:

$$\mathbb{E}\|M_iD_i^{n-k}\tilde{\xi}_k^i\|^2 = h_i\frac{((k-1)\alpha+\beta)^2}{(r_i^- - r_i^+)^2}\mathbb{E}[{\varepsilon^i}^2][(r_i^-)^{n-k}-(r_i^+)^{n-k}]^2.$$

□

In the following lemma, we bound a certain sum of powers of the roots.

Lemma 7. *For all $\alpha \leq 1/h_i$ and $0 \leq \beta \leq 2/h_i - \alpha$ we have*

$$\sum_{k=0}^{n-2} \frac{\left[(r_i^-)^k - (r_i^+)^k\right]^2}{\left[(r_i^-) - (r_i^+)\right]^2} \leq \frac{2 - \beta h_i}{4\alpha\beta h_i^2 (1 - (\frac{1}{4}\alpha + \frac{1}{2}\beta)h_i)}.$$

We first note that when the two roots become close, the denominator and the numerator will go to zero, which prevents from bounding the numerator easily. We also note that this bound is very tight since the difference between the two terms goes to zero when n goes to infinity.

Proof. We first expand the square of the difference of the powers of the roots and compute their sums.

$$\begin{aligned}
\sum_{k=0}^{n-2} \left[(r_i^-)^k - (r_i^+)^k\right]^2 &= \sum_{k=0}^{n-2} \left[{r_i^+}^{2k} + {r_i^-}^{2k} - 2(r_i^+ r_i^-)^k\right] \\
&= \frac{1 - {r_i^+}^{2(n-1)}}{1 - {r_i^+}^2} + \frac{1 - {r_i^-}^{2(n-1)}}{1 - {r_i^-}^2} - 2\frac{1 - (r_i^+ r_i^-)^{n-1}}{1 - (r_i^+ r_i^-)} \\
&= \frac{1}{1 - {r_i^+}^2} + \frac{1}{1 - {r_i^-}^2} - \frac{2}{1 - (r_i^+ r_i^-)} \\
&\quad - \left[\frac{{r_i^+}^{2(n-1)}}{1 - {r_i^+}^2} + \frac{{r_i^-}^{2(n-1)}}{1 - {r_i^-}^2} - 2\frac{(r_i^+ r_i^-)^{(n-1)}}{1 - (r_i^+ r_i^-)}\right] \\
&= \frac{1}{1 - {r_i^+}^2} + \frac{1}{1 - {r_i^-}^2} - \frac{2}{1 - (r_i^+ r_i^-)} - I_{n-1},
\end{aligned}$$

with $I_n = \left[\frac{{r_i^+}^{2n}}{1-{r_i^+}^2} + \frac{{r_i^-}^{2n}}{1-{r_i^-}^2} - 2\frac{(r_i^+ r_i^-)^n}{1-(r_i^+ r_i^-)}\right]$.

This sum is therefore equal to the sum of one term we will compute explicitly and one other term which will go to zero. We have for the first term:

$$\begin{aligned}
\frac{1}{1 - {r_i^+}^2} + \frac{1}{1 - {r_i^-}^2} - \frac{2}{1 - (r_i^+ r_i^-)} &= \frac{(1 - {r_i^-}^2)(1 - (r_i^+ r_i^-)) - (1 - {r_i^-}^2)(1 - {r_i^+}^2)}{(1 - {r_i^+}^2)(1 - {r_i^-}^2)(1 - (r_i^+ r_i^-))} \\
&\quad + \frac{(1 - {r_i^+}^2)(1 - (r_i^+ r_i^-)) - (1 - {r_i^-}^2)(1 - {r_i^+}^2)}{(1 - {r_i^+}^2)(1 - {r_i^-}^2)(1 - (r_i^+ r_i^-))},
\end{aligned}$$

with

$$\begin{aligned}
(1 - {r_i^-}^2)(1 - (r_i^+ r_i^-)) - (1 - {r_i^-}^2)(1 - {r_i^+}^2) &= (1 - {r_i^-}^2)[(1 - (r_i^+ r_i^-)) - (1 - {r_i^+}^2)] \\
&= r_i^+ (1 - {r_i^-}^2)(r_i^+ - r_i^-),
\end{aligned}$$

and

$$(1-r_i^{+2})(1-(r_i^+r_i^-))-(1-r_i^{-2})(1-r_i^{+2})=-r_i^-(1-r_i^{-2})(r_i^+-r_i^-),$$

and with $\square = r_i^+(1-r_i^{-2})(r_i^+-r_i^-)-r_i^-(1-r_i^{-2})(r_i^+-r_i^-)$,

$$\begin{aligned}\square &= (r_i^+-r_i^-)[r_i^+(1-r_i^{-2})-r_i^-(1-r_i^{-2})]\\ &= (r_i^+-r_i^-)[r_i^+-r_i^-+r_i^+r_i^-(r_i^+-r_i^-)]\\ &= (r_i^+-r_i^-)^2[1+r_i^+r_i^-].\end{aligned}$$

Therefore the first term is equal to:

$$\frac{1}{1-r_i^{+2}}+\frac{1}{1-r_i^{-2}}-\frac{2}{1-(r_i^+r_i^-)}=\frac{(r_i^+-r_i^-)^2[1+r_i^+r_i^-]}{(1-r_i^{+2})(1-r_i^{-2})(1-(r_i^+r_i^-))},$$

and the sum can be expanded as:

$$\sum_{k=0}^{n-2}\frac{[(r_i^-)^k-(r_i^+)^k]^2}{[r_i^--r_i^+]^2}=\frac{[1+r_i^+r_i^-]}{(1-r_i^{+2})(1-r_i^{-2})(1-(r_i^+r_i^-))}-J_{n-1},$$

with $J_n=\frac{I_n}{[(r_i^-)-(r_i^+)]^2}$.

Then we simplify the first term of this sum using the explicit values of the roots. We recall $r_i^\pm = r_i \pm \sqrt{\Delta_i} = 1-\frac{\alpha+\beta}{2}h_i \pm \sqrt{\left(\frac{\alpha+\beta}{2}\right)^2 h_i^2-\alpha h_i}$, therefore

$$\begin{aligned}r_i^+r_i^- &= r_i^2-\Delta_i^2\\ &= \left(1-\left(\frac{\alpha+\beta}{2}\right)h_i\right)^2-\left[\left(\frac{\alpha+\beta}{2}\right)h_i\right]^2+\alpha h_i\\ &= 1-\beta h_i,\end{aligned}$$

and

$$\begin{aligned}(1-r_i^{+2})(1-r_i^{-2}) &= [(1-r_i^-)(1-r_i^+)][(1+r_i^+)(1+r_i^+)]\\ &= [(1-r_i+\sqrt{\Delta_i})((1-r_i-\sqrt{\Delta_i})][(1+r_i+\sqrt{\Delta_i})((1+r_i-\sqrt{\Delta_i})]\\ &= [(1-r_i)^2-\Delta_i][(1+r_i)^2-\Delta_i][(1-r_i)^2-\Delta_i]\\ &= 4\alpha h_i\left(1-\left(\frac{1}{4}\alpha+\frac{1}{2}\beta\right)h_i\right).\end{aligned}$$

Thus

$$\sum_{k=0}^{n-2}\frac{[(r_i^-)^k-(r_i^+)^k]^2}{[(r_i^-)-(r_i^+)]^2}=\frac{2-\beta h_i}{4\alpha\beta h_i^2(1-(\frac{1}{4}\alpha+\frac{1}{2}\beta)h_i)}-J_{n-1}.$$

Even if J_n will be asymptotically small, we want a non-asymptotic bound, thus we will show that J_n is always positive.

In the real case $[(r_i^-)-(r_i^+)]^2 \geq 0$ and using $a^2+b^2 \geq 2ab$, for all $(a,b)\in\mathbb{R}^2$, we

have

$$\frac{{r_i^+}^{2n}}{1-{r_i^+}^2}+\frac{{r_i^-}^{2n}}{1-{r_i^-}^2}\geq 2\frac{(r_i^+r_i^-)^n}{\sqrt{(1-{r_i^+}^2)(1-{r_i^-}^2)}},$$

and using ${r_i^+}^2+{r_i^-}^2\geq 2r_i^-r_i^-$ we have

$$\sqrt{(1-{r_i^+}^2)(1-{r_i^-}^2)}\leq 1-(r_i^+r_i^-),$$

since

$$\begin{aligned}(1-{r_i^+}^2)(1-{r_i^-}^2)-[1-(r_i^+r_i^-)]^2 &= 1-{r_i^+}^2-{r_i^-}^2+(r_i^+r_i^-)^2-1+2r_i^+r_i^--(r_i^+r_i^-)^2\\ &= 2r_i^+r_i^--{r_i^+}^2-{r_i^-}^2\\ &\leq 0.\end{aligned}$$

Thus

$$\frac{{r_i^+}^{2n}}{1-{r_i^+}^2}+\frac{{r_i^-}^{2n}}{1-{r_i^-}^2}-2\frac{(r_i^+r_i^-)^n}{1-(r_i^+r_i^-)}\geq 0.$$

and $J_n\geq 0$ in the real case.

In the complex case, $[(r_i^-)-(r_i^+)]^2\leq 0$, and using $z^2+\bar{z}^2\leq 2z\bar{z}$ for all $z\in\mathbb{C}$, we have

$$\frac{{r_i^+}^{2n}}{1-{r_i^+}^2}+\frac{{r_i^-}^{2n}}{1-{r_i^-}^2}\leq 2\frac{(r_i^+r_i^-)^n}{\sqrt{(1-{r_i^+}^2)(1-{r_i^-}^2)}},$$

and using ${r_i^+}^2+{r_i^-}^2\leq 2r_i^-r_i^-$ we have

$$\sqrt{(1-{r_i^+}^2)(1-{r_i^-}^2)}\geq 1-(r_i^+r_i^-).$$

Thus

$$\frac{{r_i^+}^{2n}}{1-{r_i^+}^2}+\frac{{r_i^-}^{2n}}{1-{r_i^-}^2}-2\frac{(r_i^+r_i^-)^n}{1-(r_i^+r_i^-)}\leq 0.$$

and $J_n\geq 0$ in the complexe case.

Therefore we always have:

$$J_n\geq 0,$$

and

$$\sum_{k=0}^{n-2}\frac{[(r_i^-)^k-(r_i^+)^k]^2}{[(r_i^-)-(r_i^+)]^2}\leq\frac{2-\beta h_i}{4\alpha\beta h_i^2(1-(\frac{1}{4}\alpha+\frac{1}{2}\beta)h_i)}.$$

□

However we can also bound roughly Eq. (2.20) using Theorem 3 since we recall we have $\eta_n^i=\frac{[(r_i^-)^n-(r_i^+)^n]^2}{(r_i^--r_i^+)^2}$. This gives us the following lemma which enables to prove the second part of Proposition 6.

Lemma 8. *For all $\alpha \leq 1/h_i$ and $0 \leq \beta \leq 2/h_i - \alpha$ we have*

$$\mathbb{E}\sum_{k=2}^{n} \|M_i D_i^{n-k} \tilde{\xi}_k^i\|^2 \leq \mathbb{E}[(\varepsilon^i)^2](n-1)((n-1)\alpha+\beta)^2 \min\left\{\frac{2}{\alpha}, \frac{8(n-1)}{\alpha+\beta}, \frac{16}{h_i(\alpha+\beta)^2}\right\}.$$

Proof. From Lemma 6, we get

$$\begin{aligned}\mathbb{E}\sum_{k=2}^{n} \|M_i D_i^{n-k} \tilde{\xi}_k^i\|^2 &= h_i \mathbb{E}[(\varepsilon^i)^2] \sum_{k=2}^{n} ((k-1)\alpha+\beta)^2 \frac{[(r_i^-)^{n-k} - (r_i^+)^{n-k}]^2}{(r_i^- - r_i^+)^2} \\ &\leq \mathbb{E}[(\varepsilon^i)^2](n-1)((n-1)\alpha+\beta)^2 \min\left\{\frac{2}{\alpha}, \frac{8(n-1)}{\alpha+\beta}, \frac{16}{h_i(\alpha+\beta)^2}\right\}.\end{aligned}$$

□

2.G Comparison with Additional Other Algorithms

2.G.1 Summary

When the objective function f is quadratic and for correct choices of step-sizes, the AC-SA algorithm of Lan [2012], the SAGE algorithm of Hu et al. [2009] and the Accelerated RDA algorithm of Xiao [2010] are all equivalent to:

$$\theta_{n+1} = [I - \delta_{n+1} H_{n+1}]\theta_n + \frac{n-2}{n+1}[I - \delta_{n+1} H_{n+1}](\theta_n - \theta_{n-1}) + \delta_{n+1}\varepsilon_{n+1},$$

where we use $H_n\theta + \varepsilon_n$ as an unbiased estimate of the gradient and δ_n as step-size which values will be specified later.

Lan [2012] and Hu et al. [2009] only consider bounded cases by projecting their iterates on a bounded space. Xiao [2010] deals with the unbounded case and prove the following convergence result:

Theorem 6. *[Xiao, 2010, Theorem 6]. With $\mathbb{E}[\varepsilon_n \otimes \varepsilon_n] = C$, for step-size $\delta_n \leq \frac{n-1}{n}\gamma$ with $\gamma \leq 1/L$, we have*

$$\mathbb{E}f(\theta_n) - f(\theta_*) \leq \frac{4\|\theta_0 - \theta_*\|^2}{n^2\gamma} + \frac{n\gamma\sigma^2 \operatorname{tr} C}{3}.$$

This result is significantly more general than ours since it is valid for composite optimization and general noise on the gradients.

We now present the different algorithms and show they all share the same form.

2.G.2 AC-SA

Lemma 9. *AC-SA algorithm with step size γ_n and β_n and gradient estimate $H_{n+1}\theta_n + \varepsilon_{n+1}$ is equivalent to:*

$$\theta_{n+1} = (I - \frac{\gamma_n}{\beta_n}H_{n+1})\theta_n + \frac{\beta_{n-1}-1}{\beta_n}(I - \frac{\gamma_n}{\beta_n}H_{n+1})(\theta_n - \theta_{n-1}) + \frac{\gamma_n}{\beta_n}\varepsilon_{n+1}.$$

Proof. We recall the general **AC-SA algorithm**:

— Let the initial points $x_1^{ag} = x_1$, and the step-sizes $\{\beta_n\}_{n\leq 1}$ and $\{\gamma_n\}_{n\leq 1}$ be given.
Set $n = 1$

— **Step 1**. Set $x_n^{md} = \beta_n^{-1}x_n + (1-\beta_n^{-1})x_n^{ag}$,

— **Step 2**. Call the Oracle for computing $G(x_n^{md}, \xi_n)$ where $\mathbb{E}[G(x_n^{md}, \xi_n)] = f'(x_n^{md})$.
Set

$$x_{n+1} = x_n - \gamma_n G(x_n^{md}, \xi_n),$$
$$x_{n+1}^{ag} = \beta_n^{-1}x_{n+1} + (1-\beta_n^{-1})x_n^{ag},$$

— **Step 3**. Set $n \to n+1$ and go to step 1.

When f is quadratic we will have $G(x_n^{md}, \xi_n) = H_{n+1}x_n^{md} - \varepsilon_{n+1}$, thus $x_{n+1} = x_n - \gamma_n H_{n+1}x_n^{md} + \gamma_n\varepsilon_{n+1}$, and:

$$\begin{aligned} x_{n+1}^{ag} &= \beta_n^{-1}x_{n+1} + (1-\beta_n^{-1})x_n^{ag} \\ &= \beta_n^{-1}(x_n - \gamma_n H_{n+1}x_n^{md} + \gamma_n\varepsilon_{n+1}) + (1-\beta_n^{-1})x_n^{ag} \\ &= \beta_n^{-1}(\beta_n x_n^{md} + (1-\beta_n)x_n^{ag} - \gamma_n H_{n+1}x_n^{md} + \gamma_n\varepsilon_{n+1}) + (1-\beta_n^{-1})x_n^{ag} \\ &= x_n^{md} - \frac{\gamma_n}{\beta_n}H_{n+1}x_n^{md} + \frac{\gamma_n}{\beta_n}\varepsilon_{n+1}, \end{aligned}$$

and

$$\begin{aligned} x_n^{md} &= \beta_n^{-1}x_n + (1-\beta_n^{-1})x_n^{ag} \\ &= \beta_n^{-1}\beta_{n-1}x_n^{ag} + \beta_n^{-1}(1-\beta_{n-1})x_{n-1}^{ag} + (1-\beta_n^{-1})x_n^{ag} \\ &= x_n^{ag} + \frac{\beta_{n-1}-1}{\beta_n}[x_n^{ag} - x_{n-1}^{ag}]. \end{aligned}$$

These give the result for $\theta_n = x_n^{ag}$. □

2.G.3 SAGE

Lemma 10. *The algorithm SAGE with step-sizes L_n and α_n is equivalent to:*

$$\theta_{n+1} = (I - 1/L_{n+1}H_{n+1})\theta_n + (1-\alpha_n)\frac{\alpha_{n+1}}{\alpha_n}[I - 1/L_{n+1}H_{n+1}](\theta_n - \theta_{n-1}) + 1/L_{n+1}\varepsilon_{n+1}.$$

Proof. We recall the general **SAGE algorithm**:

— Let the initial points $x_0 = z_0 = 0$, and the step-sizes $\{\beta_n\}_{n\leq 1}$ and $\{L_n\}_{n\leq 1}$ be given.
Set $n = 1$

— **Step 1**. Set $x_n = (1-\alpha_n)y_{n-1} + \alpha_n z_{n-1}$,

— **Step 2**. Call the Oracle for computing $G(x_n, \xi_n)$ where $\mathbb{E}[G(x_n, \xi_n)] = f'(x_n)$. Set

$$y_n = x_n - 1/L_n G(x_n, \xi_n),$$
$$z_n = z_{n-1} - \alpha_n^{-1}(x_n - y_n)$$

— **Step 3**. Set $n \to n+1$ and go to step 1.

We have

$$y_n = (I - 1/L_n H_n)x_n + \gamma_n \varepsilon_n,$$

and

$$\begin{aligned} z_n &= z_{n-1} - \alpha_n^{-1}(x_n - y_n) \\ &= z_{n-1} - \alpha_n^{-1}[(1-\alpha_n)y_{n-1} + \alpha_n z_{n-1} - y_n] \\ &= \alpha_n^{-1} y_n - \alpha_n^{-1}(1-\alpha_n)y_{n-1}. \end{aligned}$$

Thus

$$\begin{aligned} x_n &= (1-\alpha_n)y_{n-1} + \alpha_n z_{n-1} \\ &= (1-\alpha_n)y_{n-1} + \alpha_n[\alpha_{n-1}^{-1} y_{n-1} - \alpha_{n-1}^{-1}(1-\alpha_{n-1})y_{n-2}] \\ &= y_{n-1} + (1-\alpha_{n-1})\frac{\alpha_n}{\alpha_{n-1}}[y_{n-1} - y_{n-2}]. \end{aligned}$$

These give the result for $\theta_n = y_n$. □

2.G.4 Accelerated RDA Method

Lemma 11. *The algorithm AccRDA with step-sizes β and α_n is equivalent to:*

$$\theta_{n+1} = (I - \gamma_{n+1}H_{n+1})\theta_n + (1-\alpha_n)\frac{\alpha_{n+1}}{\alpha_n}[I - \gamma_{n+1}H_{n+1}](\theta_n - \theta_{n-1}) + \gamma_{n+1}\varepsilon_{n+1},$$

with $\gamma_n = \frac{\alpha_n \theta_n}{L+\beta}$.

Proof. We recall the general **Accelerated RDA method**:

— Let the initial points $w_0 = v_0$, $A_0 = 0$, $\tilde{g}_0 = 0$ and the step-sizes $\{\alpha_n\}_{n\leq 1}$ and $\{\beta_n\}_{n\leq 1}$ be given.
Set $n = 1$

— **Step 1**. Set $A_n = A_{n-1} + \alpha_n$ and $\theta_n = \frac{\alpha_n}{A_n}$.

— **Step 2**. Compute the query point $u_n = (1-\theta_n)w_{n-1} + \theta_n v_{n-1}$

— **Step 3**. Call the Oracle for computing $g_n = G(u_n, \xi_n)$ where $\mathbb{E}[G(u_n, \xi_n)] =$

$f'(u_n)$, and update the weighted average $\tilde{g}_n$

$$\tilde{g}_n = (1-\theta_n)\tilde{g}_{n-1} + \theta_n g_n.$$

— **Step 4**. Set $v_n = v_0 - \frac{A_n}{L+\beta_n}\tilde{g}_n$.

— **Step 5**. Set $w_n = (1-\theta_n)w_{n-1} + \theta_n v_n$.

— **Step 6**. Set $n \to n+1$ and go to step 1.

First we have

$$\begin{aligned}
v_n &= v_0 - \frac{A_n}{L+\beta_n}\tilde{g}_n \\
&= v_0 - \frac{A_n}{L+\beta_n}[(1-\theta_n)\tilde{g}_{n-1} + \theta_n g_n] \\
&= v_0 - \frac{A_n}{L+\beta_n}[(1-\theta_n)\tilde{g}_{n-1} + \theta_n(H_{n+1}u_n + \varepsilon_{n+1})] \\
&= v_0 + (1-\theta_n)\frac{A_n(L+\beta_{n-1})}{(L+\beta_n)A_{n-1}}v_{n-1} - \frac{A_n}{L+\beta_n}\theta_n(H_{n+1}u_n + \varepsilon_{n+1})] \\
&= v_0 + (1-\theta_n)\frac{A_n(L+\beta_{n-1})}{(L+\beta_n)A_{n-1}}v_{n-1} - \frac{\alpha_n}{L+\beta_n}(H_{n+1}u_n + \varepsilon_{n+1})].
\end{aligned}$$

With $\beta_n = \beta$ we have $v_n = v_{n-1} - \frac{\alpha_n}{L+\beta}(H_{n+1}u_n + \varepsilon_{n+1})]$ and

$$w_n = (I - \frac{\alpha_n\theta_n}{L+\beta}H_{n+1})u_n + \frac{\alpha_n\theta_n}{L+\beta}\varepsilon_{n+1}.$$

Since $v_{n-1} = \theta_{n-1}^{-1}w_{n-1} - \theta_{n-1}^{-1}(1-\theta_{n-1})w_{n-2}$, then

$$u_n = (1-\theta_n)w_{n-1} + \theta_n(\theta_{n-1}^{-1}w_{n-1} - \theta_{n-1}^{-1}(1-\theta_{n-1})w_{n-2}),$$

and

$$u_n = w_{n-1} + \frac{\alpha_n A_{n-2}}{\alpha_{n-1}A_n}[w_{n-1} - w_{n-2}].$$

□

2.H Lower Bound for Stochastic Optimization for Least-Squares

In this section, we show a lower bound for optimization of quadratic functions with noisy access to gradients. We follow very closely the framework of Agarwal et al. [2012] and use their notations. The only difference with their Theorem 1 in the different choice of two functions f_i^+ and f_i^-, which we choose to be:

$$f_i^{\pm}(x) = c_i(x_i \pm \frac{r}{2})^2,$$

with a non-increasing sequence (c_i) to be chosen later. The function g_α that is optimized is thus:

$$g_\alpha(x) = \frac{1}{d}\sum_{i=1}^{d}\big\{(\frac{1}{2}+\alpha_i\delta)f_i^+(x) + (\frac{1}{2}-\alpha_i\delta)f_i^-(x)\big\}.$$

This function is quadratic and its Hessian has eigenvalues equal to $2c_i/d$. Thus, its largest eigenvalue is $2c_1/d$, which we choose equal to L.

Noisy gradients are obtained by sampling d independent Bernoulli random variables b_i, $i = 1, \ldots, d$, with parameters $(\frac{1}{2}+\alpha_i\delta)$ and using the gradient of the random function $\frac{1}{d}\sum_{i=1}^{d}\big\{b_i f_i^+(x) + (1-b_i)f_i^-(x)\big\}$. The variance of the random gradient is equal to

$$V = \sum_{i=1}^{d}\frac{1}{d^2}\mathrm{var}\Big(b_i\big[c_i(x_i+r/2) - c_i(x_i-r/2)\big]\Big) = \frac{1}{d^2}\sum_{i=1}^{d}c_i^2 r^2(1/4-\delta^2).$$

The function g_α is minimized for $x = -\alpha\delta r$, and the discrepancy measure between two functions g_α and g_β is greater than

$$\begin{aligned}\frac{1}{d}\sum_{i=1}^{d}\left\{\inf_x\big\{f_i^+(x)+f_i^-(x)\big\} - \inf_x f_i^+(x) - \inf_x f_i^-(x)\right\}1_{\alpha_i\neq\beta_i} &\geq \frac{1}{d}\sum_{i=1}^{d}\frac{3c_i r^2\delta^2}{4}1_{\alpha_i\neq\beta_i}\\ &\geq \frac{1}{d}\frac{3c_d r^2\delta^2}{4}\Delta(\alpha,\beta).\end{aligned}$$

Since the vectors $\alpha, \beta \in \{-1,1\}^d$ are so that their Hamming distance $\Delta(\alpha,\beta) \geq d/4$ for $\alpha \neq \beta$, we have a discrepancy measure greater than $\frac{3c_d r^2\delta^2}{16}$. Thus, for a an approximate optimality of $\varepsilon = \frac{c_d r^2\delta^2}{38}$, we have, following the proof of Theorem 1 (equation (29)) from Agarwal et al. [2012], for N iterations of any method that accesses a random gradient, we have:

$$1/3 \geq 1 - 2\frac{16Nd\delta^2 + \log 2}{d\log(2/\sqrt{e})}.$$

Thus, for d large, we get, up to constants, $\delta^2 \geq 1/N$ and thus $\varepsilon \geq \frac{r^2 c_d}{N}$.

For $c_1 = 2Ld$ and $c_i = L\sqrt{d}$ for the remaining ones, we get (up to constants):

$$\varepsilon \geq \frac{V}{L}\frac{\sqrt{d}}{N}.$$

This leads to the desired result for $N \leq d$.

Chapter 3

Optimal Convergence Rates for Least-Squares Regression through Stochastic Approximation

Abstract

We consider the optimization of a quadratic objective function whose gradients are only accessible through a stochastic oracle that returns the gradient at any given point plus a zero-mean finite variance random error. We present the first algorithm that achieves jointly the optimal prediction error rates for least-squares regression, both in terms of forgetting the initial conditions in $O(1/n^2)$, and in terms of dependence on the noise and dimension d of the problem, as $O(d/n)$. Our new algorithm is based on averaged accelerated regularized gradient descent, and may also be analyzed through finer assumptions on initial conditions and the Hessian matrix, leading to dimension-free quantities that may still be small in some distances while the "optimal" terms above are large.

This chapter is extracted from the paper Harder, Better, Faster, Stronger Convergence Rates for Least-Squares Regression, in collaboration with A. Dieuleveut and F. Bach, accepted in the *Journal of Machine Learning Research*.

We do not present here the extension to kernel regression and content ourself with a brief excerpt of Section 3.5 on tighter bounds. Indeed these two contributions will be presented in the thesis of Aymeric Dieuleveut.

3.1 Introduction

Many supervised machine learning problems are naturally cast as the minimization of a smooth function defined on a Euclidean space. This includes least-squares regression, logistic regression [see, e.g., Hastie et al., 2009] or generalized linear models [McCullagh and Nelder, 1989]. While small problems with few or low-dimensional input features may be solved precisely by many potential optimization algorithms (e.g., Newton method), large-scale problems with many high-dimensional features are

typically solved with simple gradient-based iterative techniques whose per-iteration cost is small.

In this chapter, we consider a quadratic objective function f whose gradients are only accessible through a stochastic oracle that returns the gradient at any given point plus a zero-mean finite variance random error. In this stochastic approximation framework [Robbins and Monro, 1951], it is known that two quantities dictate the behavior of various algorithms, namely the covariance matrix V of the noise in the gradients, and the deviation $\theta_0 - \theta_*$ between the initial point of the algorithm θ_0 and any of the global minimizer θ_* of f. This leads to a "bias/variance" decomposition [Bach and Moulines, 2013, Hsu et al., 2014] of the performance of most algorithms as the sum of two terms: (a) the bias term characterizes how fast initial conditions are forgotten and thus is increasing in a well-chosen norm of $\theta_0 - \theta_*$; while (b) the variance term characterizes the effect of the noise in the gradients, independently of the starting point, and with a term that is increasing in the covariance of the noise.

For quadratic functions with (a) a noise covariance matrix V which is proportional (with constant σ^2) to the Hessian of f (a situation which corresponds to least-squares regression) and (b) an initial point characterized by the norm $\|\theta_0 - \theta_*\|^2$, the optimal bias and variance terms are known *separately* from the optimization and statistical theories. On the one hand, the optimal bias dependency after n iterations is proportional to $\frac{L\|\theta_0-\theta_*\|^2}{n^2}$, where L is the largest eigenvalue of the Hessian of f. This rate is achieved by accelerated gradient descent [Nesterov, 1983, 2004], and is known to be optimal if the number of iterations n is less than the dimension d of the underlying predictors, but the algorithm is not robust to random or deterministic noise in the gradients [d'Aspremont, 2008, Schmidt et al., 2011, Devolder et al., 2014]. On the other hand, the optimal variance term is proportional to $\frac{\sigma^2 d}{n}$ [Tsybakov, 2009]; it is known to be achieved by averaged gradient descent [Bach and Moulines, 2013], for which the bias term only achieves $\frac{L\|\theta_0-\theta_*\|^2}{n}$ instead of $\frac{L\|\theta_0-\theta_*\|^2}{n^2}$.

Our first contribution in this chapter is to present a novel algorithm which attains optimal rates for *both the variance and the bias terms*. This algorithm analyzed in Section 3.4 is averaged accelerated gradient descent; beyond obtaining jointly optimal rates, our result shows that averaging is beneficial for accelerated techniques and provides a provable robustness to noise.

While optimal when measuring performance in terms of the dimension d and the initial distance to optimum $\|\theta_0 - \theta_*\|^2$, these rates are not adapted in many situations where either d is larger than the number of iterations n (i.e., the number of observations for regular stochastic gradient descent) or $L\|\theta_0 - \theta_*\|^2$ is much larger than n^2. Our second contribution is to provide in Section 3.5 an analysis of a new algorithm (based on some additional regularization) that can adapt our bounds to finer assumptions on $\theta_0 - \theta_*$ and the Hessian of the problem, leading in particular to dimension-free quantities that can thus be extended to the Hilbert space setting (in particular for non-parametric estimation).

This chapter is organized as follows: in Section 3.2, we present the main problem we tackle, namely least-squares regression, then introduce the two algorithms that we consider in Section 3.2.2, as well as the two types of oracles on the gradient in

	Averaged Algorithm	**Averaged Accelerated Algorithm**
Dimension dependent rates	*Section 3.3*	*Section 3.4*
Additive Noise	Lemma 12$^\diamondsuit$	Theorem 8
Multiplicative Noise	Theorem 7$^\diamondsuit$	$\natural$

Table 3.1 – Organization of Chapter 3. $\diamondsuit$: We extend results from [Bach and Moulines, 2013] to the setting in which extra regularization is added; $\natural$: it is still an open problem to get results in the accelerated setting for a multiplicative noise oracle.

Section 3.2.3. In Section 3.3, we present new results for averaged stochastic gradient descent that set the stage for Section 3.4, where we present our main novel result leading to an accelerated algorithm which is robust to noise. Our tighter analysis of convergence rates based on finer dimension-free quantities is presented in Section 3.5. Organization of the main results is summarized in the Table 3.1 bellow.

3.2 Least-Squares Regression

In this section, we present our least-squares regression framework, which is risk minimization with the square loss, together with the main assumptions regarding our model and our algorithms. These algorithms will rely on stochastic gradient oracles, which will come in two kinds, an additive noise which does not depend on the current iterate, which will correspond in practice to the full knowledge of the covariance matrix, and a "multiplicative/additive" noise, which corresponds to the regular stochastic gradient obtained from a single pair of observations. This second oracle is much harder to analyze.

3.2.1 Statistical Assumptions

We consider the following general setting:

— $\mathcal{H}$ is a d-dimensional Euclidean space with $d \geq 1$.

— The observations $(x_n, y_n) \in \mathcal{H} \times \mathbb{R}$, $n \geq 1$, are independent and identically distributed (i.i.d.), and such that $\mathbb{E}\|x_n\|^2$ and $\mathbb{E}y_n^2$ are finite.

— We consider the *least-squares regression* problem, namely the minimization of the expected loss $f(\theta) = \frac{1}{2}\mathbb{E}(\langle x_n, \theta\rangle - y_n)^2$ which is a quadratic function.

We first introduce an assumption on the distribution of x_n.

Covariance matrix. We denote by $\Sigma = \mathbb{E}(x_n \otimes x_n) \in \mathbb{R}^{d\times d}$ the population covariance matrix, which is the Hessian of f at all points. Without loss of generality, we can assume Σ is invertible by reducing $\mathcal{H}$ to the minimal subspace where all x_n, $n \geq 1$, lie almost surely. This implies that all eigenvalues of Σ are strictly positive

(but they may be arbitrarily small). Following Bach and Moulines [2013], we assume there exists $R > 0$ such that

$$\mathbb{E}\|x_n\|^2 x_n \otimes x_n \preccurlyeq R^2 \Sigma, \tag{$\mathcal{A}_1$}$$

where $A \preccurlyeq B$ means that $B - A$ is positive semi-definite. This assumption implies in particular that (a) $\mathbb{E}\|x_n\|^4$ is finite and (b) $\operatorname{tr}\Sigma = \mathbb{E}\|x_n\|^2 \leq R^2$ since taking the trace of the previous inequality we get $\mathbb{E}\|x_n\|^4 \leq R^2 \mathbb{E}\|x_n\|^2$ and using Cauchy-Schwarz inequality we get $\mathbb{E}\|x_n\|^2 \leq \sqrt{\mathbb{E}\|x_n\|^4} \leq R\sqrt{\mathbb{E}\|x_n\|^2}$.

Assumption $(\mathcal{A}_1)$ is satisfied, for example, for least-square regression with almost surely bounded data, since $\|x_n\|^2 \leq R^2$ almost surely implies $\mathbb{E}\|x_n\|^2 x_n \otimes x_n \preccurlyeq \mathbb{E}\big[R^2 x_n \otimes x_n\big] = R^2\Sigma$. This assumption is also true for data with infinite support and a bounded *kurtosis* for the projection of the covariates x_n on any direction $z \in \mathcal{H}$, e.g, for which there exists $\kappa > 0$, such that:

$$\forall z \in \mathcal{H}, \quad \mathbb{E}\langle z, x_n\rangle^4 \leq \kappa \langle z, \Sigma z\rangle^2. \tag{3.1}$$

Indeed, by Cauchy-Schwarz inequality, Equation (3.1) implies for all $(z,t) \in \mathcal{H}^2$, the following bound $\mathbb{E}\langle z, x_n\rangle^2\langle t, x_n\rangle^2 \leq \kappa\langle z, \Sigma z\rangle\langle t, \Sigma t\rangle$, which in turn implies that for all positive semi-definite symmetric matrices M, N, we have $\mathbb{E}\langle x_n, Mx_n\rangle\langle x_n, Nx_n\rangle \leq \kappa \operatorname{tr}(M\Sigma)\operatorname{tr}(N\Sigma)$. Equation (3.1), which is true for Gaussian vectors with $\kappa = 3$, thus implies $(\mathcal{A}_1)$ for $R^2 = \kappa \operatorname{tr}\Sigma = \kappa\mathbb{E}\|x_n\|^2$.

In the next two paragraphs, we introduce some quantities that will be important in the analysis, in order to get tighter bounds.

Eigenvalue decay. Most convergence bounds depend on the dimension d of $\mathcal{H}$. However it is possible to derive dimension-free and often tighter convergence rates by considering bounds depending on the value $\operatorname{tr}\Sigma^b$ for $b \in [0,1]$. Given b, if we consider the eigenvalues of Σ ordered in decreasing order, which we denote by s_i, then $\operatorname{tr}\Sigma^b = \sum_i s_i^b$, and the eigenvalues decay [1] at least as $\frac{(\operatorname{tr}\Sigma^b)^{1/b}}{i^{1/b}}$. Moreover, it is known that $(\operatorname{tr}\Sigma^b)^{1/b}$ is decreasing in b and thus, the smaller the b, the stronger the assumption. For b going to 0 then $\operatorname{tr}\Sigma^b$ tends to d and we are back in the classical low-dimensional case. When $b = 1$, we simply get $\operatorname{tr}\Sigma = \mathbb{E}\|x_n\|^2$, which will correspond to the weakest assumption in our context.

Optimal predictor. In finite dimension the regression function $f(\theta) = \frac{1}{2}\mathbb{E}(\langle x_n, \theta\rangle - y_n)^2$ always admits a global minimum $\theta_* = \Sigma^\dagger \mathbb{E}(y_n x_n)$. When initializing algorithms at $\theta_0 = 0$ or regularizing by the squared norm, rates of convergence generally depend on $\|\theta_*\|$, a quantity which could be arbitrarily large.

However there exists a systematic upper-bound [2] $\|\Sigma^{\frac{1}{2}}\theta_*\| \leq 2\sqrt{\mathbb{E}y_n^2}$. This leads naturally to the consideration of convergence bounds depending on $\|\Sigma^{r/2}\theta_*\|$ for $r \leq 1$.

1. Indeed for any $i \geq 1$, we have $is_i^b \leq \sum_{t=1}^i s_t^b \leq \operatorname{tr}(\Sigma^b)$.

2. Indeed for all $\theta \in \mathbb{R}^d$ and in particular $\theta = 0$, by Minkowski's inequality, $\|\Sigma^{\frac{1}{2}}\theta_*\| - \sqrt{\mathbb{E}y_n^2} = \sqrt{\mathbb{E}\langle\theta_*, x_n\rangle^2} - \sqrt{\mathbb{E}y_n^2} \leq \sqrt{\mathbb{E}(\langle\theta_*, x_n\rangle - y_n)^2} \leq \sqrt{\mathbb{E}(\langle\theta, x_n\rangle - y_n)^2} \leq \sqrt{\mathbb{E}(y_n)^2}$.

In infinite dimension this will correspond to assuming $\|\Sigma^{r/2}\theta_*\| < \infty$. This new assumption relates the optimal predictor with sources of ill-conditioning (since Σ is the Hessian of the objective function f), the smaller r, the stronger our assumption, with $r = 1$ corresponding to no assumption at all, $r = 0$ to θ_* in $\mathcal{H}$ and $r = -1$ to a convergence of the bias of least-squares regression with averaged stochastic gradient descent in $O\big(\frac{\|\Sigma^{-1/2}\theta_*\|^2}{n^2}\big)$[Dieuleveut and Bach, 2015, Défossez and Bach, 2015]. In this chapter, we will use arbitrary initial points θ_0 and thus our bounds will depend on $\|\Sigma^{r/2}(\theta_0 - \theta_*)\|$.

Finally, we make an assumption on the joint distribution of (x_n, y_n).

Noise. We denote by $\varepsilon_n = y_n - \langle \theta_*, x_n \rangle$ the residual for which we have $\mathbb{E}[\varepsilon_n x_n] = 0$. Although we do not have $\mathbb{E}[\varepsilon_n | x_n] = 0$ in general unless the model is well-specified, we assume the noise to be a structured process such that there exists $\sigma > 0$ with

$$\mathbb{E}[\varepsilon_n^2 x_n \otimes x_n] \preccurlyeq \sigma^2 \Sigma. \tag{$\mathcal{A}_2$}$$

Assumption $(\mathcal{A}_2)$ is satisfied for example for data almost surely bounded or when the model is well-specified, (e.g., $y_n = \langle \theta_*, x_n \rangle + \varepsilon_n$, with $(\varepsilon_n)_{n \in \mathbb{N}}$ i.i.d. of variance σ^2 and independent of x_n).

3.2.2 Averaged Gradient Methods and Acceleration

We focus in this chapter on stochastic gradient methods with and without acceleration for the least-squares function regularized by $\frac{\lambda}{2}\|\theta - \theta_0\|^2$ for $\lambda \in \mathbb{R}^+$. The regularization will be useful when deriving tighter convergence rates in Section 3.5, and it has the additional benefit of making the problem λ-strongly convex. Stochastic gradient descent (referred to from now on as "SGD"), applied to the regularized problem, can be described for $n \geq 1$ as

$$\theta_n = \theta_{n-1} - \gamma f_n'(\theta_{n-1}) - \gamma\lambda(\theta_{n-1} - \theta_0), \tag{3.2}$$

starting from $\theta_0 \in \mathcal{H}$, where $\gamma > 0$ is either called the step-size in optimization or the learning rate in machine learning, and $f_n'(\theta_{n-1})$ is an unbiased estimate of the gradient of f at θ_{n-1}, that is, its conditional expectation given all other sources of randomness is equal to $f'(\theta_{n-1})$.

Accelerated stochastic gradient descent is defined, for the regularized problem, by an iterative system with two parameters (θ_n, ν_n) satisfying for $n \geq 1$

$$\begin{aligned} \theta_n &= \nu_{n-1} - \gamma f_n'(\nu_{n-1}) - \gamma\lambda(\nu_{n-1} - \theta_0) \\ \nu_n &= \theta_n + \delta\big(\theta_n - \theta_{n-1}\big), \end{aligned} \tag{3.3}$$

starting from $\theta_0 = \nu_0 \in \mathcal{H}$, with $\gamma, \delta \in \mathbb{R}^2$ and $f_n'(\theta_{n-1})$ described as before. It may be reformulated as the following second-order recursion

$$\theta_n = (1 - \gamma\lambda)\big(\theta_{n-1} + \delta(\theta_{n-1} - \theta_{n-2})\big) - \gamma f_n'\big(\theta_{n-1} + \delta(\theta_{n-1} - \theta_{n-2})\big) + \gamma\lambda\theta_0.$$

The *momentum* coefficient $\delta \in \mathbb{R}$ is chosen to accelerate the convergence rate [Nesterov, 1983, Beck and Teboulle, 2009] and has its roots in the heavy-ball algorithm from Polyak [1964]. We especially concentrate here, following Polyak and Juditsky [1992], on the average of the sequence

$$\bar{\theta}_n = \frac{1}{n+1} \sum_{i=0}^{n} \theta_n, \tag{3.4}$$

and we note that it can be computed online as $\bar{\theta}_n = \frac{n}{n+1}\bar{\theta}_{n-1} + \frac{1}{n+1}\theta_n$.

The key ingredient in the algorithms presented above is the unbiased estimate on the gradient $f'_n(\theta)$, which can take two forms that we now describe in our setting.

3.2.3 Additive versus Multiplicative Stochastic Oracles on the Gradient

We consider the standard stochastic approximation framework [Kushner and Yin, 2003]. That is, we let $(\mathcal{F}_n)_{n\geq 0}$ be the increasing family of σ-fields that are generated by all variables (x_i, y_i) for $i \leq n$, and such that for each $\theta \in \mathcal{H}$ the random variable $f'_n(\theta)$ is square-integrable and $\mathcal{F}_n$-measurable with $\mathbb{E}[f'_n(\theta)|\mathcal{F}_{n-1}] = f'(\theta)$, for all $n \geq 0$. Consequently it is of the form

$$f'_n(\theta) = f'(\theta) - \xi_n, \tag{$\mathcal{A}_3$}$$

where the noise process ξ_n is $\mathcal{F}_n$-measurable with $\mathbb{E}[\xi_n|\mathcal{F}_{n-1}] = 0$ and $\mathbb{E}[\|\xi_n\|^2]$ is finite. We will consider two different gradient oracles.

Additive noise. The first oracle is the sum of the true gradient $f'(\theta)$ and an independent zero-mean noise that does not depend on θ. This oracle is equal to

$$f'_n(\theta) = \Sigma\theta - y_n x_n. \tag{3.5}$$

Since $f'(\theta) = \Sigma\theta - \mathbb{E}y_n x_n$, the oracle above has a noise vector $\xi_n = y_n x_n - \mathbb{E}y_n x_n$ independent of θ and therefore satisfies Assumption $(\mathcal{A}_3)$. Furthermore we also assume that there exists $\tau \in \mathbb{R}$ such that

$$\mathbb{E}[\xi_n \otimes \xi_n] \preccurlyeq \tau^2 \Sigma, \tag{$\mathcal{A}_4$}$$

that is, the noise has a particular structure adapted to least-squares regression. For optimal results for unstructured noise, with convergence rate for the noise part in $O(1/\sqrt{n})$, see Lan [2012]. Our oracle above with an additive noise which is independent of the current iterate corresponds to the first setting studied in stochastic approximation [Robbins and Monro, 1951, Duflo, 1997, Polyak and Juditsky, 1992]. While used by Bach and Moulines [2013] as an artifact of proof, for least-squares regression, such an additive noise corresponds to the situation where the distribution of x is known so that the population covariance matrix is computable, but the

distribution of the outputs $(y_n)_{n\in\mathbb{N}}$ remains unknown. Thus it may be seen as an intermediate set-up between regression estimation with fixed and random design [see, e.g., Györfi et al., 2006, Section 1.9].

Assumption $(\mathcal{A}_4)$ will be satisfied, for example if the outputs are almost surely bounded because $\mathbb{E}[\xi_n \otimes \xi_n] \preccurlyeq \mathbb{E}[y_n^2 x_n \otimes x_n] \preccurlyeq \tau^2\Sigma$ if $y_n^2 \leq \tau^2$ almost surely. But it will also be for data satisfying Equation (3.1) since we will have

$$\begin{aligned}\mathbb{E}[\xi_n \otimes \xi_n] &\preccurlyeq \mathbb{E}[y_n^2 x_n \otimes x_n] = \mathbb{E}[(\langle\theta_*, x_n\rangle + \varepsilon_n)^2 x_n \otimes x_n] \\ &\preccurlyeq 2\mathbb{E}[\langle\theta_*, x_n\rangle^2 x_n \otimes x_n] + 2\sigma^2\Sigma \\ &\preccurlyeq 2(\kappa\|\Sigma^{1/2}\theta_*\|^2 + \sigma^2)\Sigma \preccurlyeq 2(4\kappa\mathbb{E}[y_n^2] + \sigma^2)\Sigma,\end{aligned}$$

and thus Assumption $(\mathcal{A}_4)$ is satisfied with $\tau^2 = 2(4\kappa\mathbb{E}[y_n^2] + \sigma^2)$.

Stochastic noise ("multiplicative/additive"). This corresponds to:

$$f_n'(\theta) = (\langle x_n, \theta\rangle - y_n)x_n = (\Sigma + \zeta_n)(\theta - \theta_*) - \Xi_n, \tag{3.6}$$

with $\zeta_n = x_n \otimes x_n - \Sigma$ and $\Xi_n = (y_n - \langle x_n, \theta_*\rangle)x_n = \varepsilon_n x_n$. This oracle corresponds to regular SGD, which is often referred to as the least-mean-square (LMS) algorithm for least-squares regression, where the noise comes from sampling a single pair of observations. While still satisfying Assumption $(\mathcal{A}_3)$, it combines an additive noise Ξ_n independent of θ as in Eq. (3.5) and a multiplicative noise ζ_n. This multiplicative noise makes this stochastic oracle harder to analyze which explains why it is often approximated by an additive noise oracle. However it is the most widely used and most practical one. Note that for the oracle in Eq. (3.6), from Equation $(\mathcal{A}_2)$, we have $\mathbb{E}[\Xi_n \otimes \Xi_n] \preccurlyeq \sigma^2\Sigma$. It has a similar form to Assumption $(\mathcal{A}_4)$ which is valid for the additive noise oracle in Eq. (3.5): we use different constants σ^2 and τ^2 to highlight the difference between these two oracles.

3.3 Averaged Stochastic Gradient Descent

In this section, we provide convergence bounds for regularized averaged stochastic gradient descent. The main novelty compared to the work of Bach and Moulines [2013] is (a) the presence of regularization, which will be useful when deriving tighter convergence rates in Section 3.5 and (b) a much simpler proof. We first consider the additive noise in Section 3.3.1 before considering the multiplicative/additive noise in Section 3.3.2.

3.3.1 Additive Noise

We study here the convergence of the averaged SGD recursion defined by Eq. (3.2) under the simple oracle defined in Eq. (3.5). For least-squares regression, it takes the form:

$$\theta_n = \big[I - \gamma\Sigma - \gamma\lambda I\big]\theta_{n-1} + \gamma y_n x_n + \lambda\gamma\theta_0. \tag{3.7}$$

This is an easy adaptation of the work of Bach and Moulines [2013, Lemma 2] for the regularized case.

Lemma 12. *Assume* $(\mathcal{A}_4)$. *Consider the recursion in Eq.* (3.7) *with any regularization parameter* $\lambda \in \mathbb{R}_+$ *and any constant step-size* γ *such that* $\gamma(\Sigma + \lambda I) \preccurlyeq I$. *Then*

$$\mathbb{E}f(\bar{\theta}_n) - f(\theta_*) \leq \left(\lambda + \frac{1}{\gamma n}\right)^2 \|\Sigma^{1/2}(\Sigma + \lambda I)^{-1}(\theta_0 - \theta_*)\|^2 + \frac{\tau^2 \operatorname{tr}\left[\Sigma^2(\Sigma + \lambda I)^{-2}\right]}{n}. \quad (3.8)$$

We can make the following observations:

— The proof (see Appendix 3.A) relies on the fact that $\theta_n - \theta_*$ is obtainable in closed form since the cost function is quadratic and thus the recursions are linear, and follows from Polyak and Juditsky [1992].

— The constraint on the step-size γ is equivalent to $\gamma(L + \lambda) \leqslant 1$ where L is the largest eigenvalue of Σ and we thus recover the usual step-size from deterministic gradient descent [Nesterov, 2004].

— When n tends to infinity, the algorithm converges to the minimum of $f(\theta) + \frac{\lambda}{2}\|\theta - \theta_0\|^2$ and our performance guarantee becomes $\lambda^2\|\Sigma^{1/2}(\Sigma + \lambda I)^{-1}(\theta_0 - \theta_*)\|^2$. This is the standard "bias term" from regularized ridge regression [Hsu et al., 2014] which we naturally recover here. The term $\frac{\tau^2}{n} \operatorname{tr}\left[\Sigma^2(\Sigma + \lambda I)^{-2}\right]$ is usually referred to as the "variance term" [Hsu et al., 2014], and is equal to $\frac{\tau^2}{n}$ times the quantity $\operatorname{tr}\left[\Sigma^2(\Sigma + \lambda I)^{-2}\right]$, which is often called the degrees of freedom of the ridge regression problem [Gu, 2013].

— For finite n, the first term in Eq. (3.8) is the usual bias term which depends on the distance from the initial point θ_0 to the objective point θ_* with an appropriate norm. It includes a regularization-based component which is proportional to λ^2 and optimization-based component which depends on $(\gamma n)^{-2}$. The regularization-based bias appears because the algorithm tends to minimize the regularized function instead of the true function f.

— Given Eq. (3.8), it is natural to set $\lambda\gamma = \frac{1}{n}$, and the two components of the bias term are exactly of the same order leading to $\frac{4}{\gamma^2 n^2}\|\Sigma^{1/2}(\Sigma + \lambda I)^{-1}(\theta_0 - \theta_*)\|^2$. It corresponds up to a constant factor to the bias term of regularized least-squares [Hsu et al., 2014], but it is achieved by an algorithm accessing only n stochastic gradients. Note that when λ or γ depend on n, this term is not necessarily of order $O(n^{-2})$, as the numerator might be arbitrarily large. Note also that here as in the rest of the chapter, we only prove results in the finite horizon setting, meaning that the number of samples is known in advance and the parameters γ, λ may be chosen as functions of n, but remain constant along the iterations (when λ or γ depend on n, our bounds only hold for the last iterate).

— Note that the bias term can also be bounded by $\frac{1}{\gamma n}\|\Sigma^{1/2}(\Sigma + \lambda I)^{-1/2}(\theta_0 - \theta_*)\|^2$ only when $\|\theta_0 - \theta_*\|$ is finite (note the difference in the powers of n and $(\Sigma + \lambda I)^{-1}$). See the proof in Appendix 3.A.2 for details.

— The second term in Eq. (3.8) is the variance term. It depends on the noise in the gradient. When this one is not structured the variance turns to be also bounded by $\gamma \operatorname{tr}\left(\Sigma(\Sigma+\lambda I)^{-1}\mathbb{E}[\xi_n \otimes \xi_n]\right)$ (see Appendix 3.A.3) and we recover for $\gamma = O(1/\sqrt{n})$, the usual rate of $\frac{1}{\sqrt{n}}$ for SGD in the smooth case [Shalev-Shwartz et al., 2009].

— Overall we get the same performance as the empirical risk minimizer with fixed design, but with an algorithm that performs a single pass over the data.

— When $\lambda = 0$ we recover Lemma 2 of Bach and Moulines [2013]. In this case the variance term $\frac{\tau^2 d}{n}$ is optimal over all estimators in $\mathcal{H}$ [Tsybakov, 2009] even without computational limits, in the sense that no estimator that uses the same information can improve upon this rate.

3.3.2 Multiplicative/Additive Noise

When the general stochastic oracle in Eq. (3.6) is considered, the regularized LMS algorithm defined by Eq. (3.2) takes the form:

$$\theta_n = \big[I - \gamma x_n \otimes x_n - \gamma\lambda I\big]\theta_{n-1} + \gamma y_n x_n + \lambda\gamma\theta_0. \tag{3.9}$$

We have a very similar result with an additional corrective term (second line below) compared to Lemma 12.

Theorem 7. *Assume ($\mathcal{A}_{1,2}$). Consider the recursion in Eq. (3.9). For any regularization parameter $\lambda \in \mathbb{R}^+$ and for any constant step-size γ such that $2\gamma(R^2+2\lambda) \leq 1$ we have:*

$$\mathbb{E}f(\bar{\theta}_n) - f(\theta_*) \leqslant 3\Big(2\lambda + \frac{1}{\gamma n}\Big)^2 \|\Sigma^{1/2}(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)\|^2 + \frac{6\sigma^2}{n+1}\operatorname{tr}\left[\Sigma^2(\Sigma+\lambda I)^{-2}\right] + 3\frac{\big\|(\Sigma+\lambda I)^{-1/2}(\theta_0-\theta_*)\big\|^2 \operatorname{tr}(\Sigma(\Sigma+\lambda I)^{-1})}{\gamma^2(n+1)^2}.$$

We can make the following remarks:

— The proof (see Appendix 3.B) relies on a bias-variance decomposition, each term being treated separately. We adapt a proof technique from Bach and Moulines [2013] which considers the difference between the recursions in Eq. (3.9) and in Eq. (3.7).

— As in Lemma 12, the bias term can also be bounded by $\frac{1}{\gamma n}\|\Sigma^{1/2}(\Sigma+\lambda I)^{-1/2}(\theta_0-\theta_*)\|^2$ and the variance term by $\gamma \operatorname{tr}[\Sigma(\Sigma+\lambda I)^{-1}\xi_n \otimes \xi_n]$ (see proof in Appendices 3.B.4 and 3.B.5). This is useful in particular when considering unstructured noise.

— The variance term is the same as in the previous case. However there is a residual term that now appears when we go to the fully stochastic oracle (second line). This term will go to zero when γ tends to zero and can be compared to the corrective term which also appears when Hsu et al. [2014] go from fixed to

random design. Nevertheless our bounds are more concise than theirs, making significantly fewer assumptions and relying on an efficient single-pass algorithm.

— In this setting, the step-size may not exceed $1/(2(R^2+2\lambda))$, whereas with an additive noise in Lemma 12 the condition is $\gamma \leq 1/(L+\lambda)$, a quantity which can be much bigger than $1/(2(R^2+2\lambda))$, as L is the spectral radius of Σ whereas R^2 is of the order of $\operatorname{tr}(\Sigma)$. Note that in practice, computing L is as hard as computing θ_* so that the step-size $\gamma \propto 1/R^2$ is a good practical choice. See Défossez and Bach [2015] for larger allowed step-sizes that require more information.

— For $\lambda = 0$ the error is bounded by $\frac{3(1+d)}{(\gamma n)^2}\|\Sigma^{-1/2}(\theta_0-\theta_*)\|^2 + \frac{6\sigma^2 d}{n+1}$. We recover results from Défossez and Bach [2015] with a non-asymptotic bound but we lose the advantage of having an asymptotic equivalent (i.e., a limit rather than an upper-bound). We note that the assumption $(\mathcal{A}_{1,2})$ are close to the minimal assumptions required to obtain the optimal rate of convergence of $\sigma^2 d/n$ [Lecué and Mendelson, 2016, Oliveira, 2016]

3.4 Accelerated Stochastic Averaged Gradient Descent

We study the convergence under the stochastic oracle from Eq. (3.5) of averaged *accelerated* stochastic gradient descent defined by Eq. (3.3) which can be rewritten for the least-squares function f as a second-order iterative system with constant coefficients:

$$\theta_n = \big[I - \gamma\Sigma - \gamma\lambda I\big]\big[\theta_{n-1} + \delta(\theta_{n-1}-\theta_{n-2})\big] + \gamma y_n x_n + \gamma\lambda\theta_0. \tag{3.10}$$

When using averaging, we refer to this algorithm as "averaged-accelerated-SGD".

Theorem 8. *Assume $(\mathcal{A}_4)$. For any regularization parameter $\lambda \in \mathbb{R}_+$ and for any constant step-size $\gamma(\Sigma+\lambda I) \preccurlyeq I$, we have for any $\delta \in \big[\frac{1-\sqrt{\gamma\lambda}}{1+\sqrt{\gamma\lambda}}, 1\big]$, for the recursion in Eq. (3.10):*

$$\mathbb{E}f(\bar{\theta}_n) - f(\theta_*) \leq 2\Big(\lambda + \frac{36}{\gamma(n+1)^2}\Big)\|\Sigma^{1/2}(\Sigma+\lambda I)^{-1/2}(\theta_0-\theta_*)\|^2 + 8\tau^2\frac{\operatorname{tr}\big[\Sigma^2(\Sigma+\lambda I)^{-2}\big]}{n+1}.$$

The numerical constants are partially artifacts of the proof (see Appendices 3.C and 3.D). Thanks to a wise use of tight inequalities, the bound is independent of δ and valid for all $\lambda \in \mathbb{R}_+$. This results in the simple following corollary for $\lambda = 0$, which corresponds to the particularly simple recursion (with averaging to obtain $\bar{\theta}_n$):

$$\theta_n = \big[I - \gamma\Sigma\big](2\theta_{n-1} - \theta_{n-2}) + \gamma y_n x_n. \tag{3.11}$$

Corollary 4. *Assume ($\mathcal{A}_4$). For any constant step-size $\gamma\Sigma \preccurlyeq I$, we have for $\delta = 1$,*

$$\mathbb{E}f(\bar{\theta}_n) - f(\theta_*) \leq 36\frac{\|\theta_0 - \theta_*\|^2}{\gamma(n+1)^2} + 8\frac{\tau^2 d}{n+1}. \tag{3.12}$$

We can make the following observations:

— The proof technique relies on direct moment computations in each eigensubspace obtained by O'Donoghue and Candès [2013] in the deterministic case. Indeed as Σ is a symmetric matrix, the space can be decomposed on an orthonormal eigenbasis of Σ, and the iterations are decoupled in such an eigenbasis. Although we only provide an upper-bound, this is in fact an equality plus other exponentially small terms as shown in the proof which relies on linear algebra, with difficulties arising from the fact that this second-order system can be expressed as a linear stochastic dynamical system with non-symmetric matrices. We only provide a result for additive noise.

— The first bound $\frac{1}{\gamma n^2}\|\theta_0 - \theta_*\|^2$ in Eq. (3.12) corresponds to the usual accelerated rate. It has been shown by Nesterov [2004] to be the optimal rate of convergence for optimizing a quadratic function with a first-order method that can access only to sequences of gradients when $n \leq d$. We recover by averaging an algorithm dedicated to strongly convex function the traditional convergence rate for non-strongly convex functions. Even if it seems surprising, the algorithm works also for $\lambda = 0$ and $\delta = 1$ (see also simulations in Section 3.6).

— The second bound in Eq. (3.12) also matches the optimal statistical performance $\frac{\tau^2 d}{n}$ described in the observations following Lemma 12. Accordingly this algorithm achieves joint bias/variance optimality (when measured in terms of τ^2 and $\|\theta_0 - \theta_*\|^2$).

— We have the same rate of convergence for the bias when compared to the regular Nesterov acceleration without averaging studied in Chapter 2, which corresponds to choosing $\delta_n = 1 - 2/n$ for all n. However if the problem is μ-strongly convex, this latter was shown to also converge at the linear rate $O\big((1-\gamma\mu)^n\big)$ and thus is adaptive to hidden strong convexity (since the algorithm does not need to know μ to run). This explains that it ends up converging faster for quadratic function since for large n the convergence at rate $1/n^2$ becomes slower than the one at rate $(1-\gamma\mu)^n$ even for very small μ. This is confirmed in our experiments in Section 3.6. Thanks to this adaptivity, we can also show using the same tools and considering its weighted average $\tilde{\theta}_n = \frac{2}{n(n+1)}\sum_{k=0}^n k\theta_k$ that the bias term of $\mathbb{E}f(\tilde{\theta}_n) - f(\theta_*)$ has a convergence rate of order $\big(\lambda^2 + \frac{1}{\gamma^2(n+1)^4}\big)\|\Sigma^{1/2}(\Sigma + \lambda I)^{-1}(\theta_0 - \theta_*)\|^2$ without any change in the variance term. This has to be compared to the bias of averaged SGD $\big(\lambda + \frac{1}{\gamma(n+1)^2}\big)\|\Sigma^{1/2}(\Sigma + \lambda I)^{-1}(\theta_0 - \theta_*)\|^2$ in Section 3.3 and may lead to faster convergence for the bias in presence of hidden strong convexity.

— Overall, the bias term is improved whereas the variance term is not degraded and acceleration is thus robust to noise in the gradients. Thereby, while second-order finite difference methods for optimizing quadratic functions in the singular

case, such as conjugate gradient [Polyak, 1987, Section 6.1] are notoriously highly sensitive to noise, we are able to propose a version which is robust to stochastic noise.

— Note that when there is no assumption on the covariance of the noise we still have the variance bounded by $\frac{\gamma n}{2} \operatorname{tr}\left[\Sigma(\Sigma+\lambda I)^{-1}V\right]$; setting $\gamma = 1/n^{3/2}$ and $\lambda = 0$ leads to the bound $\frac{\|\theta_0-\theta_*\|^2}{\sqrt{n}} + \frac{\operatorname{tr} V}{\sqrt{n}}$. We recover the usual rate for accelerated stochastic gradient in the non-strongly convex case [Xiao, 2010]. When the values of the bias and the variance are known, we can achieve the optimal trade-off of Lan [2012] $\frac{R^2\|\theta_0-\theta_*\|^2}{n^2} + \frac{\|\theta_0-\theta_*\|\sqrt{\operatorname{tr} V}}{\sqrt{n}}$ for $\gamma = \min\left\{1/R^2, \frac{\|\theta_0-\theta_*\|}{\sqrt{\operatorname{tr} V} n^{3/2}}\right\}$.

3.5 Tighter Dimension-Independent Convergence Rates

We have seen in Corollary 4 that the averaged accelerated gradient algorithm achieves the lower bounds $\tau^2 d/n$ and $\frac{L}{n^2}\|\theta_0-\theta_*\|^2$ for the prediction error. However the algorithm behaves often better than in the worst-case scenarios corresponding to the lower bounds. Indeed the algorithm still predicts well when the dimension d is much larger than n or when the norm of the optimal predictor $\|\theta_*\|^2$ is huge. Actually gradients algorithms are adaptive to the difficulty of the problem as presented in the following corollary for the averaged *accelerated* algorithm.

Corollary 5. *Assume ($\mathcal{A}_4$), for any constant step-size $\gamma(\Sigma+\lambda I) \preccurlyeq I$, we have for $\lambda = \frac{1}{\gamma(n+1)^2}$ and $\delta \in \left[1-\frac{2}{n+2}, 1\right]$, for the recursion in Eq. (3.10):*

$$\mathbb{E} f(\bar{\theta}_n) - f(\theta_*) \leq \min_{r\in[0,1],\ b\in[0,1]} \left[74\ \frac{\|\Sigma^{r/2}(\theta_0-\theta_*)\|^2}{\gamma^{1-r}(n+1)^{2(1-r)}} + 8\frac{\tau^2\gamma^b \operatorname{tr}(\Sigma^b)}{(n+1)^{1-2b}}\right].$$

We can make the following observations:

— This corollary is a direct consequence of Theorem 9 by Dieuleveut et al. [2016] considered for $\lambda = (\gamma n^2)^{-1}$. It extends previously known bounds in the kernel least-mean-squares setting [Dieuleveut and Bach, 2015] with the use of an extra regularization.

— The algorithm is independent of r and b, thus all the bounds for different values of (r, b) are valid. This is a strong property of the algorithm, which is indeed adaptative to the regularity and the effective dimension of the problem (once γ is chosen). In situations in which either d is larger than n or $L\|\theta_0-\theta_*\|^2$ is larger than n^2, the algorithm can still enjoy good convergence properties, by adapting to the best values of b and r.

— For $b = 0$ we recover the variance term of Corollary 4, but for $b > 0$ and fast decays of eigenvalues of Σ, the bound may be much smaller; note that we lose in the dependency in n, but typically, for large d, this can be advantageous.

— For $r = 0$ we recover the bias term of Corollary 4 and for $r = 1$ (no assumption at all) the bias is bounded by $\|\Sigma^{1/2}\theta_*\|^2 \leq 4R^2$, which is not going to zero. The smaller r is, the stronger the decrease of the bias with respect to n is (which is coherent with the fact that we have a stronger assumption).

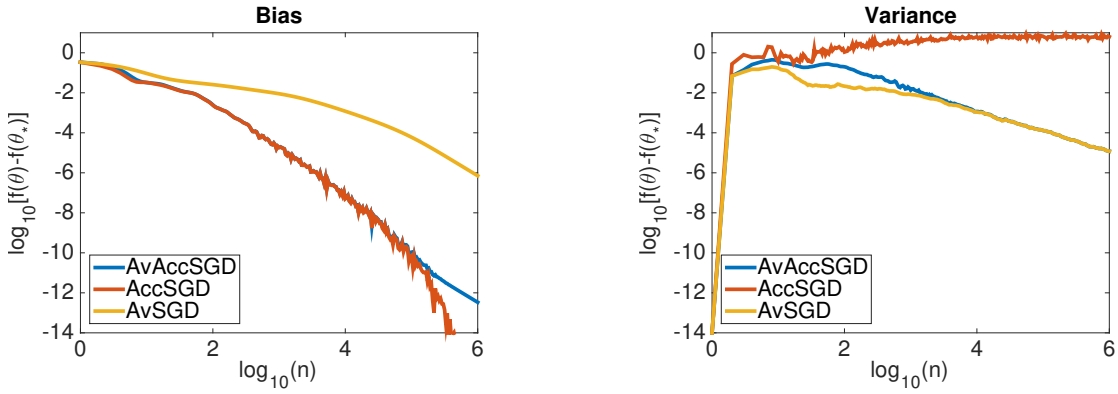


Figure 3-1 – Synthetic problem ($d = 25$) and $\gamma = 1/R^2$. Left: Bias. Right: Variance.

In the companion paper of Dieuleveut et al. [2016], we show that in the setting of non parametric learning in kernel spaces, these bounds lead to the optimal statistical rate of convergence.

3.6 Experiments

We illustrate now our theoretical results on synthetic examples. For $d = 25$ we consider normally distributed inputs x_n with random covariance matrix Σ which has eigenvalues $1/i^3$, for $i = 1, \dots, d$, and random optimum θ_* and starting point θ_0 such that $\|\theta_0 - \theta_*\| = 1$. The outputs y_n are generated from a linear function with homoscedastic noise with unit signal to noise-ratio ($\sigma^2 = 1$), we take $R^2 = \operatorname{tr} \Sigma$ the average radius of the data and a step-size $\gamma = 1/R^2$ and $\lambda = 0$. The additive noise oracle is used. We show results averaged over 10 replications.

We compare the performance of averaged SGD (AvSGD), AccSGD (usual Nesterov acceleration for convex functions) and our novel averaged accelerated SGD from Section 3.4 AvAccSGD (which is not the averaging of AccSGD because the momentum term is proportional to $1 - 3/n$ for AccSGD instead of being equal to 1 for AvAccSGD), on two different problems: one deterministic ($\|\theta_0 - \theta_*\| = 1$, $\sigma^2 = 0$) which will illustrate how the bias term behaves, and one purely stochastic ($\|\theta_0 - \theta_*\| = 0$, $\sigma^2 = 1$) which will illustrate how the variance term behaves. For the bias (left plot of Figure 3-1), AvSGD converges at speed $O(1/n)$, while AvAccSGD and AccSGD converge both at speed $O(1/n^2)$. However, as mentioned in the observations following Corollary 4, AccSGD takes advantage of the hidden strong convexity of the least-squares function and starts converging linearly at the end. For the variance (right plot of Figure 3-1), AccSGD is not converging to the optimum and keeps oscillating whereas AvSGD and AvAccSGD both converge to the optimum at a speed $O(1/n)$. However AvSGD remains slightly faster in the beginning.

Note that for small n, or when the bias $L\|\theta_0 - \theta_*\|^2/n^2$ is much bigger than the variance $\sigma^2 d/n$, the bias may have a stronger effect, although asymptotically, the variance always dominates. It is thus essential to have an algorithm which is optimal in both regimes; this is achieved by AvAccSGD.

3.7 Conclusion

In this chapter, we showed that stochastic *averaged* accelerated gradient descent was robust to structured noise in the gradients present in least-squares regression. Beyond being the first algorithm which is jointly optimal in terms of both bias and finite-dimensional variance, it is also adapted to finer assumptions such as fast decays of the covariance matrices or optimal predictors with large norms.

Our current analysis is performed for least-squares regression. While it could be directly extended to smooth losses through efficient online Newton methods [Bach and Moulines, 2013], an extension to all smooth or self-concordant-like functions [Bach, 2014] would widen its applicability. Moreover, our accelerated gradient analysis is performed for additive noise (i.e., for least-squares regression, with knowledge of the population covariance matrix) and it would be interesting to study the robustness of our results in the contexts of least-mean squares and online learning. Our analysis relies on single observations per iteration and could be made finer by using mini-batches [Cotter et al., 2011, Dekel et al., 2012], which should not change the variance term but could impact the bias term. Finally, it would be appealing to consider proximal extensions of this algorithm. In Chapter 4 we take a first step towards this goal by studying a proximal extension of averaged gradient descent.

Appendix

3.A Proofs of Section 3.3

3.A.1 Proof of Lemma 12

We proof here Lemma 12 which is the extension of Lemma 2 of Bach and Moulines [2013] for the regularized case. The proof technique relies on the fact that recursions in Eq. (3.7) are linear since the cost function is quadratic which allows us to obtain $\theta_n - \theta_*$ in closed form.

For any regularization parameter $\lambda \in \mathbb{R}_+$ and any constant step-size $\gamma(\Sigma+\lambda I) \preccurlyeq I$ we may rewrite the regularized stochastic gradient recursion in Eq. (3.7) as:

$$\theta_n - \theta_* = \big[I - \gamma\Sigma - \gamma\lambda I\big](\theta_{n-1} - \theta_*) + \gamma\xi_n + \lambda\gamma(\theta_0 - \theta_*).$$

We thus get for $n \geq 1$ the expansion

$$\begin{aligned}
\theta_n - \theta_* &= (I - \gamma\Sigma - \gamma\lambda I)^n(\theta_0 - \theta_*) + \gamma\sum_{k=1}^{n}(I - \gamma\Sigma - \gamma\lambda I)^{n-k}\xi_k \\
&\quad +\gamma\lambda\sum_{k=1}^{n}(I - \gamma\Sigma - \gamma\lambda I)^{n-k}(\theta_0 - \theta_*) \\
&= (I - \gamma\Sigma - \gamma\lambda I)^n(\theta_0 - \theta_*) + \gamma\sum_{k=1}^{n}(I - \gamma\Sigma - \gamma\lambda I)^{n-k}\xi_k \\
&\quad +\lambda\big[I - (I - \gamma\Sigma - \gamma\lambda I)^n\big](\Sigma + \lambda I)^{-1}(\theta_0 - \theta_*) \\
&= (I - \gamma\Sigma - \gamma\lambda I)^n[I - \lambda(\Sigma + \lambda I)^{-1}](\theta_0 - \theta_*) + \gamma\sum_{k=1}^{n}(I - \gamma\Sigma - \gamma\lambda I)^{n-k}\xi_k \\
&\quad +\lambda(\Sigma + \lambda I)^{-1}(\theta_0 - \theta_*).
\end{aligned}$$

We then have using the definition of the average

$$\begin{aligned}
n(\bar{\theta}_{n-1} - \theta_*) &= \sum_{j=0}^{n-1}(\theta_j - \theta_*) \\
&= \sum_{j=0}^{n-1}(I - \gamma\Sigma - \gamma\lambda I)^j[I - \lambda(\Sigma + \lambda I)^{-1}](\theta_0 - \theta_*) \\
&\quad +\gamma\sum_{j=0}^{n-1}\sum_{k=1}^{j}(I - \gamma\Sigma - \gamma\lambda I)^{n-k}\xi_k + n\lambda(\Sigma + \lambda I)^{-1}(\theta_0 - \theta_*).
\end{aligned}$$

For which we will compute the two sums separately

$$\sum_{j=0}^{n-1}(I-\gamma\Sigma-\gamma\lambda I)^j[I-\lambda(\Sigma+\lambda I)^{-1}](\theta_0-\theta_*)$$
$$=\frac{1}{\gamma}\big[I-(I-\gamma\Sigma-\gamma\lambda I)^n\big](\Sigma+\lambda I)^{-1}[I-\lambda(\Sigma+\lambda I)^{-1}](\theta_0-\theta_*),$$

and

$$\begin{aligned}\gamma\sum_{j=0}^{n-1}\sum_{k=1}^{j}(I-\gamma\Sigma-\gamma\lambda I)^{j-k}\xi_k &= \gamma\sum_{k=1}^{n-1}\Big(\sum_{j=k}^{n-1}(I-\gamma\Sigma-\gamma\lambda I)^{j-k}\Big)\xi_k\\ &= \gamma\sum_{k=1}^{n-1}\Big(\sum_{j=0}^{n-1-k}(I-\gamma\Sigma-\gamma\lambda I)^{j}\Big)\xi_k\\ &= \sum_{k=1}^{n-1}\big[I-(I-\gamma\Sigma-\gamma\lambda I)^{n-k}\big](\Sigma+\lambda I)^{-1}\xi_k.\end{aligned}$$

Gathering the three terms together, we thus have

$$\begin{aligned}n(\bar\theta_{n-1}-\theta_*) &= \frac{1}{\gamma}\big[I-(I-\gamma\Sigma-\gamma\lambda I)^n\big](\Sigma+\lambda I)^{-1}[I-\lambda(\Sigma+\lambda I)^{-1}](\theta_0-\theta_*)\\ &\quad+\sum_{k=1}^{n-1}\big[I-(I-\gamma\Sigma-\gamma\lambda I)^{n-k}\big](\Sigma+\lambda I)^{-1}\xi_k+n\lambda(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)\\ &= \Big[\frac{1}{\gamma}\big[I-(I-\gamma\Sigma-\gamma\lambda I)^n\big][I-\lambda(\Sigma+\lambda I)^{-1}]+n\lambda I\Big](\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)\\ &\quad+\sum_{k=1}^{n-1}\big[I-(I-\gamma\Sigma-\gamma\lambda I)^{n-k}\big](\Sigma+\lambda I)^{-1}\xi_k.\end{aligned}$$

Using standard martingale square moment inequalities which amount to consider ξ_i, $i=1,\dots,n$ independent, the variance of the sum is the sum of variances and we have for $V=\mathbb{E}\xi_n\otimes\xi_n$

$$\begin{aligned}&n^2\mathbb{E}\|\Sigma^{1/2}(\bar\theta_{n-1}-\theta_*)\|^2=\textstyle\sum_{k=1}^{n-1}\operatorname{tr}\big[I-(I-\gamma\Sigma-\gamma\lambda I)^{n-k}\big]^2\Sigma(\Sigma+\lambda I)^{-2}V\\ &+\Big\|\Big[\frac{1}{\gamma}\big[I-(I-\gamma\Sigma-\gamma\lambda I)^n\big][I-\lambda(\Sigma+\lambda I)^{-1}]+n\lambda I\Big]\Sigma^{1/2}(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)\Big\|^2. \quad (3.13)\end{aligned}$$

Since all the matrices in this equality are symmetric positive-definite we are allowed to bound

$$\begin{aligned}\Big[\frac{1}{\gamma}\big[I-(I-\gamma\Sigma-\gamma\lambda I)^n\big][I-\lambda(\Sigma+\lambda I)^{-1}]+n\lambda I\Big] &\preccurlyeq \Big(\frac{1}{\gamma}+n\lambda\Big)I \quad (3.14)\\ \big[I-(I-\gamma\Sigma-\gamma\lambda I)^{n-k}\big]^2 &\preccurlyeq I.\end{aligned}$$

This concludes proof of the Lemma 12

$$\mathbb{E}\|\Sigma^{1/2}(\bar{\theta}_{n-1}-\theta_*)\|^2 \leq \Big(\frac{1}{n\gamma}+\lambda\Big)^2\|\Sigma^{1/2}(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)\|^2 + \frac{1}{n}\operatorname{tr}\Sigma(\Sigma+\lambda I)^{-2}V. \quad (3.15)$$

3.A.2 Proof When Only $\|\theta_0-\theta_*\|$ Is Finite

Unfortunately $\|\Sigma^{-1}(\theta_0-\theta_*)\|$ may not be finite. However we can use that for all $u\in[0,1]$ we have $\frac{1-(1-u)^n}{nu}\leq 1$ since $\frac{1-(1-u)^n}{u}=\sum_{k=0}^n(1-u)^k\leq n$ and it yields

$$\begin{aligned}
&\Big[\frac{1}{\gamma}\big[I-(I-\gamma\Sigma-\gamma\lambda I)^n\big][I-\lambda(\Sigma+\lambda I)^{-1}]+n\lambda I\Big][\Sigma+\lambda I]^{-1}\\
&\preccurlyeq \Big[\frac{1}{\gamma}\big[I-(I-\gamma\Sigma-\gamma\lambda I)^n\big]+n\lambda I\Big][\Sigma+\lambda I]^{-1}\\
&\preccurlyeq \Big[\frac{1}{\gamma}\big[I-(I-\gamma\Sigma-\gamma\lambda I)^n\big][\Sigma+\lambda I]^{-1}+n\lambda[\Sigma+\lambda I]^{-1}\Big]\\
&\preccurlyeq I+nI.
\end{aligned}$$

Combining with Eq. (3.14) we have

$$\Big\|\Big[\frac{1}{\gamma}\big[I-(I-\gamma\Sigma-\gamma\lambda I)^n\big][I-\lambda(\Sigma+\lambda I)^{-1}]+n\lambda I\Big]\Sigma^{1/2}(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)\Big\|^2 \leq (n+1)\Big(\frac{1}{\gamma}+n\lambda\Big)\|\Sigma^{1/2}(\Sigma+\lambda I)^{-1/2}(\theta_0-\theta_*)\|^2,$$

which implies that

$$\mathbb{E}\|\Sigma^{1/2}(\bar{\theta}_{n-1}-\theta_*)\|^2 \leq 2\Big(\frac{1}{n\gamma}+\lambda\Big)\|\Sigma^{1/2}(\Sigma+\lambda I)^{-1/2}(\theta_0-\theta_*)\|^2 + \frac{1}{n}\operatorname{tr}\Sigma(\Sigma+\lambda I)^{-2}V, \quad (3.16)$$

which is interesting when only $\|\theta_0-\theta_*\|$ is finite.

3.A.3 Proof When The noise Is Not Structured

The bound in Eq. (3.15) becomes less interesting when the noise is not structured. However using the same technique we have that $\big[I-(I-\gamma\Sigma-\gamma\lambda I)^{n-k}\big]^2(\Sigma+\lambda I)^{-1}\preccurlyeq (n-k)\gamma I$ and we get the following upper-bound on the variance

$$\begin{aligned}
\sum_{k=1}^n \operatorname{tr}\big[I-(I-\gamma\Sigma-\gamma\lambda I)^{n-k}\big]^2\Sigma(\Sigma+\lambda I)^{-2}V &\leq \gamma\sum_{k=1}^n(n-k)\operatorname{tr}\Sigma(\Sigma+\lambda I)^{-1}V\\
&\leq \gamma\frac{n(n+1)}{2}\operatorname{tr}\Sigma(\Sigma+\lambda I)^{-1}V.
\end{aligned}$$

Therefore we get

$$\mathbb{E}\|\Sigma^{1/2}(\bar{\theta}_{n-1}-\theta_*)\|^2 \leq \Big(\frac{1}{n\gamma}+\lambda\Big)^2\|\Sigma^{1/2}(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)\|^2 + \gamma\,\mathrm{tr}\,\Sigma(\Sigma+\lambda I)^{-1}V, \quad (3.17)$$

which is meaningful when the noise is not structured.

3.B Proof of Theorem 7

In this section, we will prove Theorem 7. The proof relies on a decomposition of the error as the sum of three main terms which will be studied separately. We state decomposition in Section 3.B.1 then prove upper bounds for the different terms in Sections 3.B.2 and 3.B.3.

3.B.1 Expansion of the Recursion

We may rewrite the regularized stochastic gradient recursion as:

$$\begin{aligned}\theta_n &= \big[I-\gamma x_n\otimes x_n-\gamma\lambda I\big]\theta_{n-1}+\gamma\varepsilon_n x_n+\gamma\langle x_n,\theta_*\rangle x_n+\lambda\gamma\theta_0\\ \theta_n-\theta_* &= \big[I-\gamma x_n\otimes x_n-\gamma\lambda I\big](\theta_{n-1}-\theta_*)+\gamma\varepsilon_n x_n+\lambda\gamma(\theta_0-\theta_*).\end{aligned}$$

For $i\geqslant k$, let

$$M(i,k)=\big[I-\gamma x_i\otimes x_i-\gamma\lambda I\big]\cdots\big[I-\gamma x_k\otimes x_k-\gamma\lambda I\big]$$

be an operator from $\mathcal{H}$ to $\mathcal{H}$. We have the expansion

$$\theta_n-\theta_*=M(n,1)(\theta_0-\theta_*)+\gamma\sum_{k=1}^n M(n,k+1)\varepsilon_k x_k+\gamma\sum_{k=1}^n M(n,k+1)\lambda(\theta_0-\theta_*).$$

Our goal is to study these three terms separately and bound $\|\Sigma^{1/2}(\bar{\theta}_n-\theta_*)\|$ for each of them.

3.B.2 Regularization-Based Bias Term

This is the term: $\theta_n-\theta_*=\gamma\sum_{k=1}^n M(n,k+1)\lambda(\theta_0-\theta_*)$, which corresponds to the recursion

$$\theta_n-\theta_*=\big(I-\gamma x_n\otimes x_n-\gamma\lambda I\big)(\theta_{n-1}-\theta_*)+\lambda\gamma(\theta_0-\theta_*), \quad (3.18)$$

initialized with $\theta_0=\theta_*$, and no noise.

Following the proof technique of Bach and Moulines [2013], we are going to consider a related recursion by replacing in Equation (3.18) the operator $x_n\otimes x_n$ by its

expectation Σ. Thus, we consider η_n defined as

$$\eta_n - \theta_* = \gamma \sum_{k=1}^{n} (I - \gamma\Sigma - \lambda\gamma I)^{n-k} \lambda(\theta_0 - \theta_*),$$

which satisfies the recursion (with initialization $\eta_0 = \theta_*$) and

$$\eta_n - \theta_* = \big[I - \gamma\Sigma - \lambda\gamma I\big](\eta_{n-1} - \theta_*) + \lambda\gamma(\theta_0 - \theta_*).$$

In order to bound $\|\Sigma^{1/2}(\theta_n - \theta_*)\|$, we will independently bound $\|\Sigma^{1/2}(\eta_n - \theta_*)\|$ and $\|\Sigma^{1/2}(\theta_n - \eta_n)\|$ using Minkowski's inequality.

Bounding $\|\Sigma^{1/2}(\theta_n - \eta_n)\|$. We have $\theta_0 - \eta_0 = 0$, and

$$\theta_n - \eta_n = \big[I - \gamma x_n \otimes x_n - \lambda\gamma I\big](\theta_{n-1} - \eta_{n-1}) + \gamma\big[\Sigma - x_n \otimes x_n\big](\eta_{n-1} - \theta_*).$$

We can now bound the recursion for $\theta_n - \eta_n$ as follows, using standard online learning proofs [Nemirovski et al., 2009]:

$$\begin{aligned}
\|\theta_n - \eta_n\|^2 \leqslant\ & \|\theta_{n-1} - \eta_{n-1}\|^2 - 2\gamma\big\langle \theta_{n-1} - \eta_{n-1}, (x_n \otimes x_n + \lambda I)(\theta_{n-1} - \eta_{n-1})\big\rangle \\
& + 2\gamma\big\langle \theta_{n-1} - \eta_{n-1}, \big[\Sigma - x_n \otimes x_n\big](\eta_{n-1} - \theta_*)\big\rangle \\
& + \gamma^2 \big\| \big[x_n \otimes x_n + \lambda I\big](\theta_{n-1} - \eta_{n-1}) - \big[\Sigma - x_n \otimes x_n\big](\eta_{n-1} - \theta_*)\big\|^2.
\end{aligned}$$

By taking conditional expectations given $\mathcal{F}_{n-1}$, we get, using first the fact that $\mathbb{E}(\Sigma - x_n \otimes x_n | \mathcal{F}_{n-1}) = 0$ and the inequality $(a+b)^2 \leq 2(a^2+b^2)$, then developing and using $\mathbb{E}[(x_n \otimes x_n)^2] \leq R^2\Sigma$, which is assumption $\mathcal{A}_1$.

$$\begin{aligned}
\mathbb{E}\big(\|\theta_n - \eta_n\|^2 | \mathcal{F}_{n-1}\big) \leqslant\ & \|\theta_{n-1} - \eta_{n-1}\|^2 - 2\gamma\big\langle \theta_{n-1} - \eta_{n-1}, (\Sigma + \lambda I)(\theta_{n-1} - \eta_{n-1})\big\rangle \\
& + 2\gamma^2 \mathbb{E}\big(\big\| \big[x_n \otimes x_n + \lambda I\big](\theta_{n-1} - \eta_{n-1})\big\|^2 | \mathcal{F}_{n-1}\big) \\
& + 2\gamma^2 \mathbb{E}\big(\big\| \big[\Sigma - x_n \otimes x_n\big](\eta_{n-1} - \theta_*)\big\|^2 | \mathcal{F}_{n-1}\big) \\
\leqslant\ & \|\theta_{n-1} - \eta_{n-1}\|^2 - 2\gamma\big\langle \theta_{n-1} - \eta_{n-1}, (\Sigma + \lambda I)(\theta_{n-1} - \eta_{n-1})\big\rangle \\
& + 2\gamma^2 \big\langle \theta_{n-1} - \eta_{n-1}, (R^2\Sigma + \lambda^2 I + 2\lambda\Sigma)(\theta_{n-1} - \eta_{n-1})\big\rangle \\
& + 2\gamma^2 R^2 \langle \eta_{n-1} - \theta_*, \Sigma\rangle \\
\leqslant\ & \|\theta_{n-1} - \eta_{n-1}\|^2 + 2\gamma^2 R^2 \langle \eta_{n-1} - \theta_*, \Sigma(\eta_{n-1} - \theta_*)\rangle\rangle \\
& - 2\gamma\big[1 - \gamma(R^2 + 2\lambda)\big]\big\langle \theta_{n-1} - \eta_{n-1}, \Sigma(\theta_{n-1} - \eta_{n-1})
\end{aligned}$$

This leads by taking full expectations and moving terms to

$$\begin{aligned}
\mathbb{E}\big\langle \theta_{n-1} - \eta_{n-1}, \Sigma(\theta_{n-1} - \eta_{n-1})\big\rangle \leqslant\ & \frac{1}{2\gamma\big[1 - \gamma(R^2+2\lambda)\big]}\big[\mathbb{E}\|\theta_{n-1} - \eta_{n-1}\|^2 - \mathbb{E}\|\theta_n - \eta_n\|^2\big] \\
& + \frac{\gamma R^2}{1 - \gamma(R^2 + 2\lambda)} \langle \eta_{n-1} - \theta_*, \Sigma(\eta_{n-1} - \theta_*)\rangle.
\end{aligned}$$

Thus, if $\gamma(R^2+2\lambda) \leqslant \frac{1}{2}$

$$\begin{aligned}\mathbb{E}\big\langle \theta_{n-1}-\eta_{n-1}, \Sigma(\theta_{n-1}-\eta_{n-1})\big\rangle &\leqslant \frac{1}{\gamma}\big[\mathbb{E}\|\theta_{n-1}-\eta_{n-1}\|^2 - \mathbb{E}\|\theta_n-\eta_n\|^2\big] \\ &\quad +2\gamma R^2 \mathbb{E}\langle \eta_{n-1}-\theta_*, \Sigma(\eta_{n-1}-\theta_*)\rangle.\end{aligned}$$

This leads to, summing and using initial conditions $\theta_0-\eta_0=0$, then using convexity to upper bound $\big\langle \bar{\theta}_n-\bar{\eta}_n, \Sigma(\bar{\theta}_n-\bar{\eta}_n)\big\rangle \leq \frac{1}{n+1}\sum_{k=0}^n \big\langle \theta_k-\eta_k, \Sigma(\theta_k-\eta_k)\big\rangle$,

$$\mathbb{E}\big\langle \bar{\theta}_n-\bar{\eta}_n, \Sigma(\bar{\theta}_n-\bar{\eta}_n)\big\rangle \leqslant \frac{2\gamma R^2}{n+1}\sum_{k=0}^n \langle \eta_k-\theta_*, \Sigma(\eta_k-\theta_*)\rangle.$$

Bounding $\|\Sigma^{1/2}(\eta_n-\theta_*)\|$. Moreover we have:

$$\begin{aligned}\eta_n-\theta_* &= \lambda(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*) - (I-\gamma\Sigma-\lambda\gamma I)^n\big[\lambda(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)\big] \\ \bar{\eta}_n-\theta_* &= \lambda(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*) - \frac{1}{n+1}\sum_{k=0}^n (I-\gamma\Sigma-\lambda\gamma I)^k\big[\lambda(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)\big] \\ &= \lambda(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*) \\ &\quad -\frac{1}{n+1}\gamma^{-1}(\Sigma+\lambda I)^{-1}\big[I-(I-\gamma\Sigma-\lambda\gamma I)^{n+1}\big]\big[\lambda(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)\big].\end{aligned}$$

This leads using Minkowski inequality to

$$\begin{aligned}\big(\mathbb{E}\|\Sigma^{1/2}(\eta_n-\theta_*)\|^2\big)^{1/2} &\leqslant \|\lambda\Sigma^{1/2}(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)\| \\ \big(\mathbb{E}\|\Sigma^{1/2}(\bar{\eta}_n-\theta_*)\|^2\big)^{1/2} &\leqslant \|\lambda\Sigma^{1/2}(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)\|.\end{aligned}$$

Thus this part is such that

$$\begin{aligned}\sqrt{\mathbb{E}\|\Sigma^{1/2}(\bar{\theta}_n-\theta_*)\|^2} &\leqslant \|\lambda\Sigma^{1/2}(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)\| + \sqrt{2\gamma R^2\|\lambda\Sigma^{1/2}(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)\|^2} \\ &\leqslant \|\lambda\Sigma^{1/2}(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)\|\big(1+\sqrt{2\gamma R^2}\big),\end{aligned}$$

that gives the first bound on the regularization-based bias

$$\mathbb{E}\|\Sigma^{1/2}(\bar{\theta}_n-\theta_*)\|^2 \leqslant \|\lambda\Sigma^{1/2}(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)\|^2\big(1+\sqrt{2\gamma R^2}\big)^2. \tag{3.19}$$

3.B.3 Expansion without the Regularization Term

We will follow here the outline of the proof of Györfi and Walk [1996] which considers a full expansion of the function value $\|\Sigma^{1/2}(\bar{\theta}_n-\theta_*)\|^2$. This corresponds to

$$\theta_n-\theta_* = M(n,1)(\theta_0-\theta_*) + \gamma\sum_{k=1}^n M(n,k+1)\varepsilon_k x_k.$$

We have

$$\mathbb{E}\sum_{i=0}^{n}\sum_{j=0}^{n}\langle\theta_i-\theta_*,\Sigma(\theta_j-\theta_*)\rangle = \mathbb{E}\sum_{i=0}^{n}\|\Sigma^{1/2}(\theta_i-\theta_*)\|^2+2\mathbb{E}\sum_{i=0}^{n-1}\sum_{j=i+1}^{n}\langle\theta_i-\theta_*,\Sigma(\theta_j-\theta_*)\rangle.$$

Moreover,

$$\begin{aligned}
&\mathbb{E}\sum_{i=0}^{n-1}\sum_{j=i+1}^{n}\langle\theta_i-\theta_*,\Sigma(\theta_j-\theta_*)\rangle\\
=\ &\mathbb{E}\sum_{i=0}^{n-1}\sum_{j=i+1}^{n}\left\langle\theta_i-\theta_*,\Sigma\Big[M(j,i+1)(\theta_i-\theta_*)+\sum_{k=i+1}^{j}M(j,k+1)\gamma\varepsilon_k x_k\Big]\right\rangle\\
=\ &\mathbb{E}\sum_{i=0}^{n-1}\sum_{j=i+1}^{n}\langle\theta_i-\theta_*,\Sigma M(j,i+1)(\theta_i-\theta_*)\rangle \text{ because } \varepsilon_k x_k \text{ and } \theta_i \text{ are independent,}\\
=\ &\mathbb{E}\sum_{i=0}^{n-1}\sum_{j=i+1}^{n}\langle\theta_i-\theta_*,\Sigma(I-\gamma\Sigma-\gamma\lambda I)^{j-i}(\theta_i-\theta_*)\rangle\\
&\qquad\text{because } M(j,i+1) \text{ and } \theta_i \text{ are independent,}\\
=\ &\mathbb{E}\sum_{i=0}^{n-1}\Big\langle\theta_i-\theta_*,\gamma^{-1}\Sigma(\Sigma+\lambda I)^{-1}\big[(I-\gamma\Sigma-\gamma\lambda I)-(I-\gamma\Sigma-\gamma\lambda I)^{n-i+1}\big](\theta_i-\theta_*)\Big\rangle\\
\leqslant\ &\mathbb{E}\sum_{i=0}^{n}\Big\langle\theta_i-\theta_*,\gamma^{-1}\Sigma(\Sigma+\lambda I)^{-1}(I-\gamma\Sigma-\gamma\lambda I)(\theta_i-\theta_*)\Big\rangle \text{ using } (\Sigma+\lambda I)\preccurlyeq I,\\
=\ &\gamma^{-1}\mathbb{E}\sum_{i=0}^{n}\langle\theta_i-\theta_*,\Sigma(\Sigma+\lambda I)^{-1}(\theta_i-\theta_*)\rangle-\mathbb{E}\sum_{i=0}^{n}\langle\theta_i-\theta_*,\Sigma(\theta_i-\theta_*)\rangle.
\end{aligned}$$

We thus simply need to bound $\gamma^{-1}\mathbb{E}\sum_{i=0}^{n}\langle\theta_i-\theta_*,\Sigma(\Sigma+\lambda I)^{-1}(\theta_i-\theta_*)\rangle$, to get a bound on $n^2\mathbb{E}\|\Sigma^{1/2}(\bar{\theta}_n-\theta_*)\|^2$.

Recursion on operators. We have:

$$\begin{aligned}
\mathbb{E}\big[M(i,k)\Sigma(\Sigma+\lambda I)^{-1}M(i,k)^*\big] &= \mathbb{E}\Big[M(i,k+1)\big[I-\gamma x_k\otimes x_k-\gamma\lambda I\big]\Sigma(\Sigma+\lambda I)^{-1}\\
&\qquad\big[I-\gamma x_k\otimes x_k-\gamma\lambda I\big]M(i,k+1)^*\Big]\\
&= \mathbb{E}\Big[M(i,k+1)\Big(\Sigma(\Sigma+\lambda I)^{-1}-2\gamma\Sigma+\gamma^2\big[x_k\otimes x_k\\
&\qquad+\lambda I\big]\Sigma(\Sigma+\lambda I)^{-1}\big[x_k\otimes x_k+\lambda I\big]\Big)M(i,k+1)^*\Big]\\
&\preccurlyeq \mathbb{E}\Big[M(i,k+1)\big[\Sigma(\Sigma+\lambda I)^{-1}-2\gamma\Sigma\\
&\qquad+\gamma^2(R^2+2\lambda)\Sigma\big]M(i,k+1)^*\Big]\\
&= \mathbb{E}\Big[M(i,k+1)\Sigma(\Sigma+\lambda I)^{-1}M(i,k+1)^*\Big]
\end{aligned}$$

$$-\gamma(2-\gamma(R^2+2\lambda))\mathbb{E}\Big[M(i,k+1)\Sigma M(i,k+1)^*\Big],$$

which leads to

$$\mathbb{E}\Big[M(i,k+1)\Sigma M(i,k+1)^*\Big] \preccurlyeq \frac{1}{\gamma(2-\gamma(R^2+2\lambda))} \Big(E\Big[M(i,k+1)\Sigma(\Sigma+\lambda I)^{-1}M(i,k+1)^*\Big] - E\Big[M(i,k)\Sigma(\Sigma+\lambda I)^{-1}M(i,k)^*\Big]\Big). \tag{3.20}$$

Using the operator T on matrices defined below, this corresponds to showing

$$(I-\gamma T)\big[\Sigma(\Sigma+\lambda I)\big] \preccurlyeq \Sigma(\Sigma+\lambda I)-\gamma\Sigma.$$

Noise term. For $\theta_0-\theta_*=0$, we have:

$$\begin{aligned}
&\mathbb{E}\langle\theta_i-\theta_*,\Sigma(\Sigma+\lambda I)^{-1}(\theta_i-\theta_*)\rangle\\
=\;&\gamma^2\mathbb{E}\sum_{k=1}^{i}\sum_{j=1}^{i}\varepsilon_j x_j^* M(i,j+1)^*\Sigma(\Sigma+\lambda I)^{-1}M(i,k+1)\varepsilon_k x_k \text{ by expanding all terms,}\\
=\;&\gamma^2\mathbb{E}\sum_{k=1}^{i}\varepsilon_k x_k^* M(i,k+1)^*\Sigma(\Sigma+\lambda I)^{-1}M(i,k+1)\varepsilon_k x_k \text{ using independence,}\\
=\;&\gamma^2\operatorname{tr}\left(\sum_{k=1}^{i}\mathbb{E}\varepsilon_k^2 x_k x_k^*\mathbb{E}M(i,k+1)^*\Sigma(\Sigma+\lambda I)^{-1}M(i,k+1)\right)\\
\leqslant\;&\gamma^2\sigma^2\operatorname{tr}\left(\sum_{k=1}^{i}\mathbb{E}M(i,k+1)\Sigma M(i,k+1)^*\Sigma(\Sigma+\lambda I)^{-1}\right)\\
&\qquad\text{using our assumption regarding the noise.}
\end{aligned}$$

Using the recurrence between operators

$$\begin{aligned}
&\mathbb{E}\langle\theta_i-\theta_*,\Sigma(\Sigma+\lambda I)^{-1}(\theta_i-\theta_*)\rangle\\
\leqslant\;&\frac{\gamma\sigma^2}{2-\gamma(R^2+2\lambda)}\operatorname{tr}\sum_{k=1}^{i}\left(E\Big[M(i,k+1)\Sigma(\Sigma+\lambda I)^{-1}M(i,k+1)^*\Sigma(\Sigma+\lambda I)^{-1}\Big]\right.\\
&\left.-E\Big[M(i,k)\Sigma(\Sigma+\lambda I)^{-1}M(i,k)^*\Sigma(\Sigma+\lambda I)^{-1}\Big]\right)\\
\leqslant\;&\frac{\gamma\sigma^2}{2-\gamma(R^2+2\lambda)}\operatorname{tr}\left(E\Big[M(i,i+1)\Sigma(\Sigma+\lambda I)^{-1}M(i,i+1)^*\Sigma(\Sigma+\lambda I)^{-1}\Big]\right.\\
&\left.-E\Big[M(i,1)\Sigma(\Sigma+\lambda I)^{-1}M(i,1)^*\Sigma(\Sigma+\lambda I)^{-1}\Big]\right) \text{ by summing,}\\
\leqslant\;&\frac{\gamma\sigma^2}{2-\gamma(R^2+2\lambda)}\operatorname{tr}\Sigma^2(\Sigma+\lambda I)^{-2}.
\end{aligned}$$

This implies that for the noise process

$$\mathbb{E}\|\Sigma^{1/2}(\bar{\theta}_n - \theta_*)\|^2 \leqslant \left(\frac{\sigma^2}{n+1}\operatorname{tr}\left[\Sigma^2(\Sigma+\lambda I)^{-2}\right]\right)\frac{1}{1-\gamma(R^2/2+\lambda)}.$$

Note that when γ tends to zero, we recover the optimal variance term.

Noiseless term. Without noise, we then need to bound:

$$\gamma^{-1}\mathbb{E}\sum_{i=0}^{n}\langle\theta_i - \theta_*, \Sigma(\Sigma+\lambda I)^{-1}(\theta_i-\theta_*)\rangle,$$

with $\theta_i - \theta_* = M(i,1)(\theta_0-\theta_*)$, that is

$$\gamma^{-1}\mathbb{E}\sum_{i=0}^{n}\operatorname{tr}\Big[M(i,1)^*\Sigma(\Sigma+\lambda I)^{-1}M(i,1)(\theta_0-\theta_*)(\theta_0-\theta_*)^*\Big].$$

We follow here the proof of Défossez and Bach [2015] and consider the operator T from symmetric matrices to symmetric matrices defined as

$$TA = (\Sigma+\lambda I)A + A(\Sigma+\lambda I) - \gamma E\big[(x_n\otimes x_n+\lambda I)A(x_n\otimes x_n+\lambda I)\big].$$

of the form $TA = (\Sigma+\lambda I)A + (\Sigma+\lambda I)A - \gamma SA$.

The operator S is self-adjoint and positive. Moreover:

$$\begin{aligned}
\langle A, SA\rangle &= \mathbb{E}\operatorname{tr}\left[A(x_n\otimes x_n+\lambda I)A(x_n\otimes x_n+\lambda I)\right]\\
&= \operatorname{tr}\left[2A^2\lambda\Sigma+\lambda^2A^2\right] + \mathbb{E}\operatorname{tr}\left[\langle x_n, Ax_n\rangle^2\right]\\
&\leqslant \operatorname{tr}\left[2A^2\lambda\Sigma+\lambda^2A^2\right] + \mathbb{E}\operatorname{tr}\left[\|x_n\|^2x_n\otimes x_n, A^2\right]\text{(Cauchy-Schwarz inequality)}\\
&\leqslant \operatorname{tr}\left[2A^2\lambda\Sigma+\lambda^2A^2\right] + R^2\operatorname{tr}\Sigma A^2\\
&\leqslant (R^2+2\lambda)\operatorname{tr}\left[\Sigma+\lambda I\right]A^2.
\end{aligned}$$

We have for any symmetric matrix A:

$$\mathbb{E}M(i,1)^*AM(i,1) = (I-\gamma T)^iA.$$

Thus,

$$\mathbb{E}\sum_{i=0}^{n}\operatorname{tr}\Big[M(i,1)^*\Sigma(\Sigma+\lambda I)^{-1}M(i,1)(\theta_0-\theta_*)(\theta_0-\theta_*)^*\Big] = \mathbb{E}\sum_{i=0}^{n}\langle\langle(I-\gamma T)^iA, E_0\rangle\rangle$$

with $E_0 = (\theta_0-\theta_*)(\theta_0-\theta_*)^*$ and $A = \Sigma(\Sigma+\lambda I)^{-1}$. This leads to

$$\gamma^{-1}\mathbb{E}\langle\langle\gamma^{-1}T^{-1}(I-(I-\gamma T)^{n+1})A, E_0\rangle\rangle,$$

where $\langle\langle\cdot,\cdot\rangle\rangle$ denote the dot-product between self-adjoint operators.

The sum is less than its limit for $n \to \infty$, and thus, we can get rid of the term $(I - \gamma T)^{n+1}$, and we need to bound

$$\gamma^{-2}\langle\langle M, E_0\rangle\rangle = \gamma^{-2}\langle\langle T^{-1}(\Sigma(\Sigma+\lambda I)^{-1}), E_0\rangle\rangle,$$

with $M := T^{-1}\big[\Sigma(\Sigma+\lambda I)^{-1}\big]$, i.e., such that

$$\begin{aligned}\Sigma(\Sigma+\lambda I)^{-1} &= (\Sigma+\lambda I)M + M(\Sigma+\lambda I) - \gamma\mathbb{E}(x_n\otimes x_n + \lambda I)M(x_n\otimes x_n+\lambda I)\\ &= (\Sigma+\lambda I)M + M(\Sigma+\lambda I) - \gamma SM. \qquad (3.21)\end{aligned}$$

So that :

$$\begin{aligned}M &= \big[(\Sigma+\lambda I)\otimes I + I\otimes(\Sigma+\lambda I)\big]^{-1}\big[\Sigma(\Sigma+\lambda I)^{-1}\big] + \gamma\big[(\Sigma+\lambda I)\otimes I + I\otimes(\Sigma+\lambda I)\big]^{-1}SM\\ &= \frac{1}{2}\Sigma(\Sigma+\lambda I)^{-2} + \gamma\big[(\Sigma+\lambda I)\otimes I + I\otimes(\Sigma+\lambda I)\big]^{-1}SM.\end{aligned}$$

The operator $(\Sigma+\lambda I)\otimes I + I\otimes(\Sigma+\lambda I)$ is self adjoint, and so is its inverse, thus:

$$\begin{aligned}\langle\langle M, E_0\rangle\rangle &= \langle\langle \frac{1}{2}\Sigma(\Sigma+\lambda I)^{-2} + \gamma^3\big[(\Sigma+\lambda I)\otimes I + I\otimes(\Sigma+\lambda I)\big]^{-1}SM, E_0\rangle\rangle\\ &= \frac{1}{2}\langle\langle \Sigma(\Sigma+\lambda I)^{-2}, E_0\rangle\rangle + \gamma\langle\langle SM, \big[(\Sigma+\lambda I)\otimes I + I\otimes(\Sigma+\lambda I)\big]^{-1}E_0\rangle\rangle\\ &= \frac{1}{2}\operatorname{tr}(\Sigma(\Sigma+\lambda I)^{-2}E_0) + \gamma\langle\langle SM, \big[(\Sigma+\lambda I)\otimes I + I\otimes(\Sigma+\lambda I)\big]^{-1}E_0\rangle\rangle.\end{aligned}$$

Moreover,

$$\begin{aligned}E_0 &= (\theta_0-\theta_*)(\theta_0-\theta_*)^*\\ &= (\Sigma+\lambda I)^{1/2}(\Sigma+\lambda I)^{-1/2}(\theta_0-\theta_*)(\theta_0-\theta_*)^*(\Sigma+\lambda I)^{-1/2}(\Sigma+\lambda I)^{+1/2}\\ &\preccurlyeq [(\theta_0-\theta_*)^*(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)]\ \ (\Sigma+\lambda I),\end{aligned}$$

as $(\Sigma+\lambda I)^{-1/2}(\theta_0-\theta_*)(\theta_0-\theta_*)^*(\Sigma+\lambda I)^{-1/2} \preccurlyeq (\theta_0-\theta_*)^*(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)I$. Thus, as $\big[(\Sigma+\lambda I)\otimes I + I\otimes(\Sigma+\lambda I)\big]^{-1}$ is an non-decreasing operator on $(S_n(\mathbb{R}), \preccurlyeq)$ (see technical Lemma 16 in Appendix 3.D):

$$\begin{aligned}&\big[(\Sigma+\lambda I)\otimes I + I\otimes(\Sigma+\lambda I)\big]^{-1}E_0\\ \preccurlyeq\ &\big[(\Sigma+\lambda I)\otimes I + I\otimes(\Sigma+\lambda I)\big]^{-1}\big([(\theta_0-\theta_*)^*(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)](\Sigma+\lambda I)\big)\\ =\ &\frac{(\theta_0-\theta_*)^*(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)}{2}I.\end{aligned}$$

Thus as SM is positive :

$$\gamma^{-2}\langle\langle M, E_0\rangle\rangle \ \le\ \frac{1}{2\gamma^2}\operatorname{tr}(\Sigma(\Sigma+\lambda I)^{-2}E_0) + \frac{(\theta_0-\theta_*)^*(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)}{2\gamma}\operatorname{tr}(SM).$$

Moreover we can upper bound $\mathrm{tr}(SM)$: using Equation (3.21) we have

$$\mathrm{tr}(\Sigma(\Sigma+\lambda I)^{-1}) = 2\,\mathrm{tr}(\Sigma+\lambda I)M - \gamma\,\mathrm{tr}\,\mathbb{E}(x_n\otimes x_n+\lambda I)M(x_n\otimes x_n+\lambda I)$$

then, using Assumption $(\mathcal{A}_1)$:

$$\mathrm{tr}\,\mathbb{E}(x_n\otimes x_n+\lambda I)M(x_n\otimes x_n+\lambda I) \leqslant R^2\,\mathrm{tr}M\Sigma+2\,\mathrm{tr}M\Sigma\lambda+\lambda^2\,\mathrm{tr}\,M \leqslant (R^2+2\lambda)\,\mathrm{tr}M(\Sigma+\lambda I).$$

This implies

$$\begin{aligned}\mathrm{tr}\left[\Sigma(\Sigma+\lambda I)^{-1}\right] &\geqslant \big(\frac{2}{R^2+2\lambda}-\gamma\big)\,\mathrm{tr}\,\mathbb{E}(x_n\otimes x_n+\lambda I)M(x_n\otimes x_n+\lambda I),\\ &\geqslant \frac{1}{R^2+2\lambda}\,\mathrm{tr}\,\mathbb{E}(x_n\otimes x_n+\lambda I)M(x_n\otimes x_n+\lambda I) \text{ since } \gamma(R^2+2\lambda)\leqslant 1,\\ &\geqslant \frac{1}{R^2+2\lambda}\,\mathrm{tr}\,SM.\end{aligned}$$

Thus finally:

$$\begin{aligned}\gamma^{-2}\langle\langle M, E_0\rangle\rangle \leq \frac{1}{2\gamma^2}\,\mathrm{tr}\,E_0\Sigma(\Sigma+\lambda I)^{-2}&\\ + \frac{(\theta_0-\theta_*)^*(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)}{2\gamma}(R^2+2\lambda)\,\mathrm{tr}(\Sigma(\Sigma+\lambda I)^{-1}),&\end{aligned}$$

which leads to the desired error term.

3.B.4 Proof When Only $\|\theta_0-\theta_*\|$ Is Finite

When $\lambda = 0$, without noise, we then need to bound:

$$\gamma^{-1}\mathbb{E}\sum_{i=0}^{n}\langle\theta_i-\theta_*,\theta_i-\theta_*\rangle,$$

with $\theta_i-\theta_* = M(i,1)(\theta_0-\theta_*)$, that is

$$\gamma^{-1}\mathbb{E}\sum_{i=0}^{n}\mathrm{tr}\left[M(i,1)^*M(i,1)(\theta_0-\theta_*)(\theta_0-\theta_*)^*\right].$$

By definition of $M(i,1)$ we have that $\mathbb{E}M(i,1)^*M(i,1)\preccurlyeq I$ leading to

$$\gamma^{-1}\mathbb{E}\sum_{i=0}^{n}\langle\theta_i-\theta_*,\theta_i-\theta_*\rangle \leq \frac{(n+1)\|\theta_0-\theta_*\|^2}{\gamma}.$$

For the regularization-based bias we also have

$$\|\lambda\Sigma^{1/2}(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)\|^2 \leq \lambda\|\Sigma^{1/2}(\Sigma+\lambda I)^{-1/2}(\theta_0-\theta_*)\|^2.$$

3.B.5 Proof When the Noise Is Not Structured

For $\|\theta_0 - \theta_*\| = 0$ we have $\theta_n - \theta_* = \gamma \sum_{k=1}^n M(n, k+1)\varepsilon_k x_k$ which leads to

$$\mathbb{E}\|\Sigma^{1/2}(\theta_n - \theta_*)\|^2 = \gamma^2 \sum_{k=1}^n \operatorname{tr} \mathbb{E} M(n, k+1)^* \Sigma M(n, k+1) V,$$

where $V = \mathbb{E}\varepsilon_k^2 x_k x_k^*$. And using the recursion on operators in Eq. (3.20) by changing order of elements we have

$$\mathbb{E}\Big[M(n, k+1)^* \Sigma M(n, k+1)\Big] \preccurlyeq \frac{1}{\gamma(2 - \gamma(R^2 + 2\lambda))}$$
$$\Big(E\Big[M(n, k+1)^* \Sigma (\Sigma + \lambda I)^{-1} M(n, k+1)\Big] - E\Big[M(n, k)^* \Sigma (\Sigma + \lambda I)^{-1} M(n, k)\Big]\Big).$$

And by adding the terms

$$\mathbb{E}\|\Sigma^{1/2}(\theta_n - \theta_*)\|^2 \preccurlyeq \frac{\gamma^2}{\gamma(2 - \gamma(R^2 + 2\lambda))} \operatorname{tr} \Sigma(\Sigma + \lambda I)^{-1} V,$$

We conclude by convexity

$$\mathbb{E}\|\Sigma^{1/2}(\bar{\theta}_n - \theta_*)\|^2 \preccurlyeq \frac{\gamma^2}{\gamma(2 - \gamma(R^2 + 2\lambda))} \operatorname{tr} \Sigma(\Sigma + \lambda I)^{-1} V.$$

3.C Convergence of Accelerated Averaged Stochastic Gradient Descent

We now prove Theorem 8. We thus consider iterates satisfying Eq. (3.10), under Assumptions $(\mathcal{A}_3)$, $(\mathcal{A}_4)$. We consider a fixed step size γ such that $\gamma(\Sigma + \lambda I) \preccurlyeq I$. Seeing Eq. (3.10) as a linear second order for θ_n, we will derive from exact calculations a decomposition of the errors a sum of three terms that will be studied independently. The proof is organized as follows: in Section 3.C.1, we state the formulation as a second order linear system and derive the three main terms that have to be studied (see Lemma 13). Section 3.C.2 studies asymptotic behaviors of the three terms, ignoring some exponentially decreasing terms, in order to give insight of how they behave. This section is not necessary for the proof, indeed a direct and exact calculation in the eigenbasis of Σ, following O'Donoghue and Candès [2013], is provided in Section 3.C.3. Results are summed up in Section 3.C.4.

3.C.1 General Expansion

We study the regularized stochastic accelerated gradient descent recursion defined for $n \geq 1$ by

$$\theta_n \;=\; \nu_{n-1} - \gamma f'(\nu_{n-1}) - \gamma\lambda(\nu_n - \theta_0) + \gamma\xi_n$$

$$\nu_n \;=\; \theta_n + \delta(\theta_n - \theta_{n-1}),$$

starting from $\theta_0 = \nu_0 \in \mathcal{H}$. We may rewrite it for a quadratic function $f : \theta \mapsto \frac{1}{2}\langle \theta - \theta_*, \Sigma(\theta - \theta_*)\rangle$ for $n \geq 2$ as

$$\theta_n = \big[I - \gamma\Sigma - \gamma\lambda I\big]\Big[\theta_{n-1} + \delta(\theta_{n-1} - \theta_{n-2})\Big] + \gamma\xi_n + \gamma\lambda\theta_0 + \gamma\Sigma\theta_*,$$

with $\theta_0 \in \mathcal{H}$ and $\theta_1 = \big[I - \gamma\Sigma - \gamma\lambda I\big]\theta_0 + \gamma\xi_1 + \gamma\lambda\theta_0 + \gamma\Sigma\theta_*$.

And by centering around the optimum, we get:

$$\theta_n - \theta_* = \big[I - \gamma\Sigma - \gamma\lambda I\big]\Big[\theta_{n-1} - \theta_* + \delta(\theta_{n-1} - \theta_* - \theta_{n-2} + \theta_*)\Big] + \gamma\xi_n + \lambda\gamma(\theta_0 - \theta_*).$$

Thus this is a second order iterative system which is standard to cast in a linear form

$$\Theta_n = F\Theta_{n-1} + \gamma\Xi_n + \gamma\lambda\Theta_\lambda, \tag{3.22}$$

with $T = I - \gamma\Sigma - \gamma\lambda I$, $F = \begin{pmatrix} (1+\delta)T & -\delta T \\ I & 0 \end{pmatrix}$, $\Theta_n = \begin{pmatrix} \theta_n - \theta_* \\ \theta_{n-1} - \theta_* \end{pmatrix}$, $\Theta_0 = \begin{pmatrix} \theta_0 - \theta_* \\ \theta_0 - \theta_* \end{pmatrix}$, $\Xi_n = \begin{pmatrix} \xi_n \\ 0 \end{pmatrix}$ and $\Theta_\lambda = \begin{pmatrix} \theta_0 - \theta_* \\ 0 \end{pmatrix}$.

We are interested in the behavior of the average $\bar{\Theta}_n = \frac{1}{n+1}\sum_{k=0}^n \Theta_k$ for which we have the following general convergence result:

Lemma 13. *For all $\lambda \in \mathbb{R}_+$ and γ such that $\gamma(\Sigma + \lambda I) \preccurlyeq I$ and any matrix C the average of the iterates Θ_n defined by Eq. (3.22) satisfy for $P_k \overset{(def)}{=} C^{1/2}(I - F^k)(I - F)^{-1}$, with $\tilde{\Theta}_0 = \Theta_0 - \gamma\lambda(I - F)^{-1}\Theta_\lambda$,*

$$\begin{aligned}\mathbb{E}\langle \bar{\Theta}_n, C\bar{\Theta}_n\rangle \;\leq\; & 2\,(\gamma\lambda)^2\,\|C^{1/2}(I - F)^{-1}\Theta_\lambda\|^2 + \frac{2}{(n+1)^2}\|P_{n+1}\tilde{\Theta}_0\|^2 \\ & + \frac{\gamma^2}{(n+1)^2}\sum_{j=1}^n \operatorname{tr} P_j V P_j^\top.\end{aligned}$$

The error thus decomposes as the sum of three main terms:

— the two first ones are bias terms, one arising from the regularization (the first one), and one arising computation (the second one),

— a variance term. which is the last one.

We remark that as we have assumed that Σ is invertible, the matrix $I - F$ can be shown to be invertible for all the considered δ.

The regularization-based term will be studied directly whereas the two others will be studied in two stages. First a heuristic will lead to an asymptotic bound then an exact computation will give a non-asymptotic bound. Then using $C = H = \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix}$ would give a convergence result on the function value and $C = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}$ a result on the iterate. The end of the section is devoted to the proof of this lemma.

Proof. The sequence Θ_n satisfies a linear recursion, from which we get, for all $n \geq 1$:

$$\begin{aligned}\Theta_n &= F^n\Theta_0 + \gamma\sum_{k=1}^{n} F^{n-k}\Xi_k + \gamma\lambda\sum_{k=1}^{n} F^{n-k}\Theta_\lambda \\ &= F^n\Theta_0 + \gamma\sum_{k=1}^{n} F^{n-k}\Xi_k + \gamma\lambda(I-F^n)(I-F)^{-1}\Theta_\lambda.\end{aligned}$$

We study the averaged sequence: $\bar{\Theta}_n = \frac{1}{n+1}\sum_{k=0}^{n}\Theta_k$. We get, using the identity $\sum_{k=0}^{n-1} F^k = (I-F^n)(I-F)^{-1}$,

$$\bar{\Theta}_n = \frac{1}{n+1}\sum_{k=0}^{n} F^k\Theta_0 + \frac{\gamma}{n+1}\sum_{k=1}^{n}\sum_{j=1}^{k} F^{k-j}\Xi_j + \frac{\gamma\lambda}{n+1}\sum_{k=1}^{n}(I-F^k)(I-F)^{-1}\Theta_\lambda.$$

With

$$\tilde{\Theta}_0 = \Theta_0 - \gamma\lambda(I-F)^{-1}\Theta_\lambda,$$

and $\sum_{k=1}^{n}(I-F^k) = \sum_{k=0}^{n}(I-F^k) = [n+1-(I-F^{n+1})(I-F)^{-1}]$.

Using summation formulas for geometric series, we derive:

$$\begin{aligned}\bar{\Theta}_n &= \frac{1}{n+1}(I-F^{n+1})(I-F)^{-1}\tilde{\Theta}_0 + \frac{\gamma}{n+1}\sum_{k=1}^{n}\sum_{j=1}^{k} F^{k-j}\Xi_j + \gamma\lambda(I-F)^{-1}\Theta_\lambda \\ &= \frac{1}{n+1}(I-F^{n+1})(I-F)^{-1}\tilde{\Theta}_0 + \frac{\gamma}{n+1}\sum_{j=1}^{n}\Big(\sum_{k=j}^{n} F^{k-j}\Big)\Xi_j + \gamma\lambda(I-F)^{-1}\Theta_\lambda \\ &= \frac{1}{n+1}(I-F^{n+1})(I-F)^{-1}\tilde{\Theta}_0 + \frac{\gamma}{n+1}\sum_{j=1}^{n}\Big(\sum_{k=0}^{n-j} F^{k}\Big)\Xi_j + \gamma\lambda(I-F)^{-1}\Theta_\lambda \\ &= \frac{1}{n+1}(I-F^{n+1})(I-F)^{-1}\tilde{\Theta}_0 + \frac{\gamma}{n+1}\sum_{j=1}^{n}(I-F^{n+1-j})(I-F)^{-1}\Xi_j + \gamma\lambda(I-F)^{-1}\Theta_\lambda \\ &= \frac{1}{n+1}(I-F^{n+1})(I-F)^{-1}\tilde{\Theta}_0 + \frac{\gamma}{n+1}\sum_{j=1}^{n}(I-F^{j})(I-F)^{-1}\Xi_{n+1-j} + \gamma\lambda(I-F)^{-1}\Theta_\lambda.\end{aligned}$$

Using martingale square moment inequalities which amount to consider $\Xi_i, i = 1, ..., n$ independent, so that the variance of the sum is the sum of variances, and denoting by $V = \mathbb{E}[\Xi_n \otimes \Xi_n]$ we have for any positive semi-definite C,

$$\begin{aligned}\mathbb{E}\langle\bar{\Theta}_n, C\bar{\Theta}_n\rangle &= \left\|C^{1/2}\left(\frac{1}{n+1}(I-F^{n+1})(I-F)^{-1}\tilde{\Theta}_0 + \gamma\lambda(I-F)^{-1}\Theta_\lambda\right)\right\|^2 \\ &\quad + \frac{\gamma^2}{(n+1)^2}\sum_{j=1}^{n}\mathrm{tr}(I-F^j)(I-F)^{-1}V(I-F^\top)^{-1}(I-F^j)^\top C,\end{aligned}$$

where $C^{1/2}$ denotes a symmetric square root of C. Define $P_k \overset{(def)}{=} C^{1/2}(I-F^k)(I-F)^{-1}$, we have, Using Minkowski's inequality and inequality $(a+b)^2 \leq 2(a^2+b^2)$ for any $a, b \in \mathbb{R}$,

$$\begin{aligned}\mathbb{E}\langle \bar{\Theta}_n, C\bar{\Theta}_n\rangle &= \left\| \frac{1}{n+1} P_{n+1}\tilde{\Theta}_0 + \gamma\lambda C^{1/2}(I-F)^{-1}\Theta_\lambda \right\|^2 + \frac{\gamma^2}{(n+1)^2}\sum_{j=1}^{n} \operatorname{tr} P_j V P_j^\top \\ &\leq 2\left(\gamma\lambda\right)^2 \|C^{1/2}(I-F)^{-1}\Theta_\lambda\|^2 + \frac{2\|P_{n+1}\tilde{\Theta}_0\|^2}{(n+1)^2} + \frac{\gamma^2}{(n+1)^2}\sum_{j=1}^{n} \operatorname{tr} P_j V P_j^\top .\end{aligned}$$

This concludes proof of Lemma 13. □

3.C.2 Asymptotic Expansion

To give the main terms that we expect, we first provide an asymptotic analysis, which shall only be understood as an insight and is not necessary for the proof. Operator F will have only eigenvalues smaller than 1, thus $|||F^j|||$ will decrease exponentially to 0 as $j \to \infty$ (even if $|||F|||$ [3] might be bigger than 1). The asymptotic analysis relies on ignoring all terms in which F^j appears. We thus approximately have:

$$\begin{aligned}\mathbb{E}\langle \bar{\Theta}_n, C\bar{\Theta}_n\rangle \;\leq\; & 2\left(\gamma\lambda\right)^2 \|C^{1/2}(I-F)^{-1}\Theta_\lambda\|^2 + 2\left\| C^{1/2}\frac{1}{n+1}(I-F^{n+1})(I-F)^{-1}\tilde{\Theta}_0 \right\|^2 \\ & + \frac{\gamma^2}{(n+1)^2}\sum_{j=1}^{n} \operatorname{tr}(I-F^j)(I-F)^{-1}V(I-F^\top)^{-1}(I-F^j)^\top C \\ \approx\; & 2\left(\gamma\lambda\right)^2 \|C^{1/2}(I-F)^{-1}\Theta_\lambda\|^2 + 2\left\| C^{1/2}\frac{1}{n+1}(I-F)^{-1}\tilde{\Theta}_0 \right\|^2 \\ & + \frac{\gamma^2}{(n+1)^2}\sum_{j=1}^{n} \operatorname{tr}(I-F)^{-1}V(I-F^\top)^{-1}C,\end{aligned}$$

where, as it has been explained $\approx$ stands for an equality up to terms that will decay exponentially. However, these terms have to be studied very carefully, what will be done in the Section 3.C.3.

Using the matrix inversion lemma we have for $C = \begin{pmatrix} c & 0 \\ 0 & 0 \end{pmatrix}$,

$$\begin{aligned} I-F &= \begin{pmatrix} (1+\delta)(\gamma\Sigma+\gamma\lambda I) - \delta I & \delta(I-(\gamma\Sigma+\gamma\lambda I)) \\ -I & I \end{pmatrix} \\ (I-F)^{-1} &= \begin{pmatrix} (\gamma\Sigma+\gamma\lambda I)^{-1} & \delta\big(I-(\gamma\Sigma+\gamma\lambda I)^{-1}\big) \\ (\gamma\Sigma+\gamma\lambda I)^{-1} & (1+\delta)I - \delta(\gamma\Sigma+\gamma\lambda I)^{-1} \end{pmatrix} \\ C^{1/2}(I-F)^{-1} &= \begin{pmatrix} c^{1/2}(\gamma\Sigma+\gamma\lambda I)^{-1} & \delta c^{1/2}\big(I-(\gamma\Sigma+\gamma\lambda I)^{-1}\big) \\ 0 & 0 \end{pmatrix}. \end{aligned} \tag{3.23}$$

3. $|||F|||$ denotes the operator norm of F, i.e., $\sup_{\|x\|\leq 1}\|Fx\|$.

Regularization based term. This gives for the regularization based term

$$\left\|C^{1/2}(I-F)^{-1}\Theta_\lambda\right\|^2 = \left\| \begin{pmatrix} c^{1/2}(\gamma\Sigma+\gamma\lambda I)^{-1} & \delta c^{1/2}\big(I-(\gamma\Sigma+\gamma\lambda I)^{-1}\big) \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \theta_0-\theta_* \\ 0 \end{pmatrix} \right\|^2$$
$$= \left(\frac{1}{\gamma}\right)^2 \|(c^{1/2}(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*))\|^2. \tag{3.24}$$

The computation of this term is exact (not asymptotic).

Bias term. For the bias term we have

$$\begin{aligned}
\tilde\Theta_0 &= \Theta_0 - \gamma\lambda(I-F)^{-1}\Theta_\lambda \\
&= \begin{pmatrix} \theta_0-\theta_* \\ \theta_0-\theta_* \end{pmatrix} - \gamma\lambda \begin{pmatrix} (\gamma\Sigma+\gamma\lambda I)^{-1} & \delta\big(I-(\gamma\Sigma+\gamma\lambda I)^{-1}\big) \\ (\gamma\Sigma+\gamma\lambda I)^{-1} & (1+\delta)I-\delta(\gamma\Sigma+\gamma\lambda I)^{-1} \end{pmatrix} \begin{pmatrix} \theta_0-\theta_* \\ 0 \end{pmatrix} \\
&= \begin{pmatrix} \theta_0-\theta_* \\ \theta_0-\theta_* \end{pmatrix} - \gamma\lambda \begin{pmatrix} (\gamma\Sigma+\gamma\lambda I)^{-1}(\theta_0-\theta_*) \\ (\gamma\Sigma+\gamma\lambda I)^{-1}(\theta_0-\theta_*) \end{pmatrix} \\
&= \begin{pmatrix} [I-\lambda(\Sigma+\lambda I)^{-1}](\theta_0-\theta_*) \\ [I-\lambda(\Sigma+\lambda I)^{-1}](\theta_0-\theta_*) \end{pmatrix}.
\end{aligned}$$

Thus this gives for the dominant term

$$\left\|C^{1/2}(I-F)^{-1}\tilde\Theta_0\right\|^2 = \left\| \begin{pmatrix} c^{1/2}(\gamma\Sigma+\gamma\lambda I)^{-1} & \delta c^{1/2}\big(I-(\gamma\Sigma+\gamma\lambda I)^{-1}\big) \\ 0 & 0 \end{pmatrix} \tilde\Theta_0 \right\|^2$$
$$= \|(c^{1/2}[(1-\delta)(\gamma\Sigma+\gamma\lambda I)^{-1}+\delta I][I-\lambda(\Sigma+\lambda I)^{-1}](\theta_0-\theta_*)\|^2.$$

And if c commutes with Σ we have the bound for $\delta \in [\frac{1-\sqrt{\gamma\lambda}}{1+\sqrt{\gamma\lambda}}, 1]$

$$\begin{aligned}
\left\|C^{1/2}(I-F)^{-1}\tilde\Theta_0\right\|^2 &\le (\frac{(1-\delta)}{\gamma\lambda}+\delta)\|(c^{1/2}[I-\lambda(\Sigma+\lambda I)^{-1}](\theta_0-\theta_*)\|^2 \\
&\le (\frac{2}{\sqrt{\gamma\lambda}}+1)\|(c^{1/2}[I-\lambda(\Sigma+\lambda I)^{-1}](\theta_0-\theta_*)\|^2.
\end{aligned}$$

Variance term. And for the variance term with $V = \begin{pmatrix} v & 0 \\ 0 & 0 \end{pmatrix}$, we have $C^{1/2}(I-F)^{-1}V^{1/2} = \begin{pmatrix} c^{1/2}(\gamma\Sigma+\gamma\lambda I)^{-1}v^{1/2} & 0 \\ 0 & 0 \end{pmatrix}$, and

$$\operatorname{tr} C^{1/2}(I-F)^{-1}V(I-F^\top)^{-1}C^{1/2} = \operatorname{tr} c(\gamma\Sigma+\gamma\lambda I)^{-1}v(\gamma\Sigma+\gamma\lambda I)^{-1}.$$

This gives the three dominant terms. However in order to control the remainders we have to compute the eigenvalues more carefully, as done in the next section.

3.C.3 Direct Computation without the Regularization Based Term

We derive now direct computation both the bias and variance terms. This is not required for the regularization based term whose previous expression in Eq. (3.24) is already non-asymptotic. Following O'Donoghue and Candès [2013] we consider an eigen-decomposition of the matrix F, in order to study independently the recursion on eigenspaces. We assume Σ has eigenvalues (s_i) and we decompose vectors in an eigenvector basis of Σ we denote by (p_i), with $\theta_n^i = p_i^\top \theta_n$ and $\xi_n^i = p_i^\top \xi_n$ and we have the reduced equation:

$$\Theta_{n+1}^i = F_i \Theta_n^i + \gamma \Xi_{n+1}^i.$$

with $\Theta_0^i = \tilde{\Theta}_0^i$, $F_i = \begin{pmatrix} (1+\delta)T_i & -\delta T_i \\ 1 & 0 \end{pmatrix}$, with $T_i = 1 - \gamma s_i - \gamma\lambda$.

Computing initial point $\tilde{\Theta}_0^i$. $\tilde{\Theta}_0^i = \Theta_0^i - \gamma\lambda(I - F_i)^{-1}\Theta_\lambda^i$, with $\Theta_0^i = \begin{pmatrix} \theta_0^i - \theta_*^i \\ \theta_0^i - \theta_*^i \end{pmatrix}$, $\Theta_\lambda^i = \begin{pmatrix} \theta_0^i - \theta_*^i \\ 0 \end{pmatrix}$ and $(I - F_i)^{-1}$ given in Eq. (3.23). Thus

$$\begin{aligned} \tilde{\Theta}_0^i &= \begin{pmatrix} \theta_0^i - \theta_*^i \\ \theta_0^i - \theta_*^i \end{pmatrix} - \frac{\gamma\lambda}{(\gamma s_i + \gamma\lambda)} \begin{pmatrix} 1 & \delta((\gamma s_i + \gamma\lambda) - 1) \\ 1 & (1+\delta)(\gamma s_i + \gamma\lambda) - \delta \end{pmatrix} \begin{pmatrix} \theta_0^i - \theta_*^i \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} (1 - \frac{\lambda}{\lambda + s_i})(\theta_0^i - \theta_*^i) \\ (1 - \frac{\lambda}{\lambda + s_i})(\theta_0^i - \theta_*^i) \end{pmatrix}. \end{aligned} \tag{3.25}$$

Study of spectrum of F_i. Depending on δ, F_i may have two distinct complex eigenvalues of same modulus, only one (double) eigenvalue, or two real eigenvalues. We only consider the two former cases, which we detail below.

Indeed, the characteristic polynomial

$$\chi_{F_i}(X) \overset{def}{=} \det(XI - F_i) = X^2 - (1+\delta)(1 - \gamma(s_i + \lambda))X + \delta(1 - \gamma(s_i + \lambda))$$

has discriminant $\Delta_i = (1-\gamma(s_i+\lambda))((1+\delta)^2(1-\gamma(s_i+\lambda)) - 4\delta)$ which is non positive as far as $\delta \in [\delta_-; \delta_+]$, with $\delta_- = \frac{1-\sqrt{\gamma(s_i+\lambda)}}{1+\sqrt{\gamma(s_i+\lambda)}}$, $\delta_+ = \frac{1+\sqrt{\gamma(s_i+\lambda)}}{1-\sqrt{\gamma(s_i+\lambda)}}$.

Two Distinct Eigenvalues

We first assume that F_i has two distinct complex eigenvalues

$$r_\pm = \frac{(1+\delta)(1 - \gamma(s_i + \lambda)) \pm \sqrt{-1}\sqrt{-\Delta_i}}{2}$$

which are conjugate. Thus the roots are of the form $\rho_i e^{\pm i\omega_i}$ with $\rho_i = \sqrt{\delta(1-\gamma(s_i+\lambda))}$, $\cos(\omega_i) = \frac{(1+\delta)(1-\gamma(s_i+\lambda))}{2\rho_i}$, $\omega_i \in [-\pi/2; \pi/2]$ and $\sin(\omega_i) = \frac{\sqrt{-\Delta_i}}{2\rho_i}$.

Let $Q_i = \begin{pmatrix} r_i^- & r_i^+ \\ 1 & 1 \end{pmatrix}$ be the transfer matrix into an eigenbasis of F_i, i.e., $F_i = Q_i D_i Q_i^{-1}$ with $D_i = \begin{pmatrix} r_i^- & 0 \\ 0 & r_i^+ \end{pmatrix}$ and $Q_i^{-1} = \frac{1}{r_i^- - r_i^+} \begin{pmatrix} 1 & -r_i^+ \\ -1 & r_i^- \end{pmatrix}$.

Computing $P_{i,k}$. We first compute the matrix $P_{i,k}$: With

$$C_i^{1/2} = \begin{pmatrix} \sqrt{c_i} & 0 \\ 0 & 0 \end{pmatrix}, C_i^{1/2} Q_i = \begin{pmatrix} r_i^- \sqrt{c_i} & r_i^+ \sqrt{c_i} \\ 0 & 0 \end{pmatrix}$$

we have

$$C_i^{1/2} Q_i (I - D_i^k)(I - D_i)^{-1} = \sqrt{c_i} \begin{pmatrix} \frac{1-(r_i^-)^k}{1-r_i^-} r_i^- & \frac{1-(r_i^+)^k}{1-r_i^+} r_i^+ \\ 0 & 0 \end{pmatrix},$$

and, when developing and regrouping terms which depend on k, we get :

$$\begin{aligned} P_{i,k} &= C_i^{1/2} Q_i (I - D_i^k)(I - D_i)^{-1} Q_i^{-1} \\ &= \frac{\sqrt{c_i}}{r_i^- - r_i^+} \begin{pmatrix} \frac{1-(r_i^-)^k}{1-r_i^-} r_i^- - \frac{1-(r_i^+)^k}{1-r_i^+} r_i^+ & \frac{1-(r_i^+)^k}{1-r_i^+} r_i^- r_i^+ - \frac{1-(r_i^-)^k}{1-r_i^-} r_i^+ r_i^- \\ 0 & 0 \end{pmatrix} \\ &= \sqrt{c_i} \begin{pmatrix} \frac{1}{(1-r_i^-)(1-r_i^+)} & \frac{-r_i^+ r_i^-}{(1-r_i^-)(1-r_i^+)} \\ 0 & 0 \end{pmatrix} \\ &\quad - \frac{\sqrt{c_i}}{r_i^- - r_i^+} \begin{pmatrix} \frac{(r_i^-)^{k+1}}{1-r_i^-} - \frac{(r_i^+)^{k+1}}{1-r_i^+} & \frac{(r_i^+)^{k+1}}{1-r_i^+} r_i^- - \frac{(r_i^-)^{k+1}}{1-r_i^-} r_i^+ \\ 0 & 0 \end{pmatrix}. \end{aligned}$$

We also have $P_{i,k} = C_i^{1/2} Q_i (I - D_i^k)(I - D_i)^{-1} Q_i^{-1} = \sum_{j=0}^{k-1} R_{i,j}$ with

$$\begin{aligned} R_{i,j} &= C_i^{1/2} Q_i D_i^j Q_i^{-1} \\ &= \sqrt{c_i} \begin{pmatrix} (r_i^-)^{j+1} & (r_i^+)^{j+1} \\ 0 & 0 \end{pmatrix} Q_i^{-1} \\ &= \frac{\sqrt{s_i}}{r_i^- - r_i^+} \begin{pmatrix} (r_i^-)^{j+1} - (r_i^+)^{j+1} & -r_i^+ (r_i^-)^{j+1} + r_i^- (r_i^+)^{j+1} \\ 0 & 0 \end{pmatrix}, \end{aligned}$$

but computing error terms based in $R_{i,j}$ before summing these errors gives a looser error bound than a tight calculation using $P_{i,k}$. More precisely, if we use $P_{i,k}\Theta_0^i = \sum_{j=0}^{k-1} R_{i,j}\Theta_0^i$ to upper bound $\|P_{i,k}\Theta_0^i\| \leq \sum_{j=0}^{k-1} \|R_{i,j}\Theta_0^i\|$, we end up with a worse bound.

Bias term. Thus, for the bias term:

$$P_{i,k}\Theta_0^i = \sqrt{c_i}\theta_0^i \frac{1 - r_i^+ r_i^-}{(1 - r_i^-)(1 - r_i^+)} - \frac{\sqrt{c_i}\theta_0^i}{r_i^- - r_i^+} \begin{pmatrix} \left[(r_i^-)^{k+1} \frac{1-r_i^+}{1-r_i^-} - (r_i^+)^{k+1} \frac{1-r_i^-}{1-r_i^+} \right] \\ 0 \end{pmatrix}$$

$$= \frac{\sqrt{c_i}\theta_0^i}{\sqrt{(1-r_i^-)(1-r_i^+)}} \begin{pmatrix} \frac{\left[(1-r_i^+r_i^-)-\rho_i^k A_1\right]}{\sqrt{(1-r_i^-)(1-r_i^+)}} \\ 0 \end{pmatrix},$$

where

$$\rho_i^k A_1 = \frac{(r_i^-)^{k+1}(1-r_i^+)^2 - (r_i^+)^{k+1}(1-r_i^-)^2}{r_i^- - r_i^+}.$$

This can be bound with the following lemma

Lemma 14. *For all $\rho \in (0,1)$ and $\omega \in [-\pi/2;\pi/2]$ and $r^\pm = \rho(\cos(\omega) \pm \sqrt{-1}\sin(\omega))$ we have:*

$$\left|\frac{1-r^+r^- - \rho^k|A_1|}{|1-r^+|}\right| \leq 3 + 3\rho^k \leq 6 \tag{3.26}$$

We note that the exact constant seems empirically to be 2. This lemma is proved as Lemma 17 in Appendix 3.D. This gives for the bias term

$$\begin{aligned} \|P_{i,k}\Theta_0^i\| &= \frac{\sqrt{c_i}(\theta_0^i)}{\sqrt{(1-r_i^-)(1-r_i^+)}}\Big[\frac{1}{\sqrt{(1-r_i^-)(1-r_i^+)}}\left((1-r_i^+r_i^-) - \rho_i^k A_1\right)\Big] \\ &\leq 6\frac{\sqrt{c_i}(\theta_0^i)}{\sqrt{\gamma(s_i+\lambda)}}, \end{aligned}$$

since:

$$\begin{aligned} (1-r_i^-)(1-r_i^+) &= 1 - 2\,\mathfrak{Re}\,(r_i^+) + |r_i^+|^2 \\ &= 1 - (1+\delta)(1-\gamma(s_i+\lambda)) + \delta(1-\gamma(s_i+\lambda)) \\ &= \gamma(s_i+\lambda). \end{aligned}$$

We also have a looser bound using $P_{i,k}\Theta_0^i = \sum_{j=0}^{k-1} R_{i,j}\Theta_0^i$.

$$\begin{aligned} R_{i,j}\Theta_0^i &= \frac{\sqrt{c_i}\theta_0^i}{r_i^- - r_i^+}\left((1-r_i^+)(r_i^-)^{j+1} - (1-r_i^-)(r_i^+)^{j+1}\right) \\ &= \sqrt{c_i}\theta_0^i\left(\frac{(r_i^-)^{j+1} - (r_i^+)^{j+1}}{r_i^- - r_i^+} - \frac{r_i^+(r_i^-)^{j+1} - r_i^-(r_i^+)^{j+1}}{r_i^- - r_i^+}\right) \text{(De Moivre's formula)} \\ &= \sqrt{c_i}\theta_0^i\left(\frac{\rho_i^{j+1}\sin(\omega_i(j+1))}{\rho_i\sin(\omega_i)} - \frac{\rho_i e^{i\omega_i}\rho_i^{j+1}e^{-i\omega_i(j+1)} - \rho_i e^{-i\omega_i}\rho_i^{j+1}e^{+i\omega_i(j+1)}}{\rho_i e^{-i\omega_i} - \rho_i e^{i\omega_i}}\right) \\ &= \sqrt{c_i}\theta_0^i\left(\frac{\rho_i^{j+1}\sin(\omega_i(j+1))}{\rho_i\sin(\omega_i)} - \rho_i^{j+1}\frac{e^{-i\omega_i j} - e^{+i\omega_i j}}{e^{-i\omega_i} - e^{i\omega_i}}\right) \\ &= \sqrt{c_i}\theta_0^i\left(\frac{\rho_i^{j}\sin(\omega_i(j+1))}{\sin(\omega_i)} - \rho_i^{j+1}\frac{\sin(\omega_i j)}{\sin(\omega_i)}\right) \\ &\leq (1+e^{-1})\sqrt{c_i}\theta_0^i \quad \text{using Lemma 18 (see proof in Appendix 3.D),} \end{aligned}$$

which also gives for the bias term

$$\|P_{i,k}\Theta_0^i\| \leq (1+e^{-1})\sqrt{c_i}\theta_0^i k.$$

Thus we have the final bound:

$$\|P_{i,k}\Theta_0^i\|^2 \leq \min\left\{36\frac{c_i(\theta_0^i)^2}{\gamma(s_i+\lambda)}, 6n(1+e^{-1})\frac{c_i(\theta_0^i)^2}{\sqrt{\gamma(s_i+\lambda)}}, n^2(1+e^{-1})^2 c_i(\theta_0^i)^2\right\}. \quad (3.27)$$

Variance term. As for the variance term, with $V_i = \begin{pmatrix} v_i & 0 \\ 0 & 0 \end{pmatrix}$, we have $\operatorname{tr} P_{i,k}V_iP_{i,k} = \left\|P_{i,k}\begin{pmatrix}\sqrt{v_i}\\0\end{pmatrix}\right\|^2$.

$$\begin{aligned}\left\|P_{i,k}\begin{pmatrix}\sqrt{v_i}\\0\end{pmatrix}\right\| &= \frac{\sqrt{v_ic_i}}{(1-r_i^-)(1-r_i^+)}\left[1+\frac{(r_i^-)^{k+1}(1-r_i^+)-(r_i^+)^{k+1}(1-r_i^-)}{r_i^+-r_i^-}\right]\\ &= \frac{\sqrt{v_ic_i}}{\gamma(s_i+\lambda)}\left[1-\rho_i^k B_{i,k}\right],\end{aligned}$$

where

$$\rho_i^k B_{i,k} = -\frac{(r_i^-)^{k+1}(1-r_i^+)-(r_i^+)^{k+1}(1-r_i^-)}{r_i^+-r_i^-},$$

which we can bound using the following Lemma:

Lemma 15. *For all $\rho \in (0,1)$ and $\omega \in [-\pi/2;\pi/2]$ and $r^\pm = \rho(\cos(\omega) \pm \sqrt{-1}\sin(\omega))$ we have:*

$$\left|\rho^k B_k\right| \leq 1.75.$$

Where we note that the exact majoration seems to be 1.3. This Lemma is proved as Lemma 19 in Appendix 3.D.

We can also have a looser bound using $P_{i,k}\begin{pmatrix}v_i^{1/2}\\0\end{pmatrix} = \sum_{j=0}^{k-1} R_{i,j}\begin{pmatrix}v_i^{1/2}\\0\end{pmatrix}$ and

$$\begin{aligned}R_{i,j}\begin{pmatrix}v_i^{1/2}\\0\end{pmatrix} &= \frac{\sqrt{c_iv_i}}{r_i^- - r_i^+}\left((r_i^-)^{j+1}-(r_i^+)^{j+1}\right)\\ &= \sqrt{c_iv_i}\ \frac{\rho_i^{j+1}\sin(\omega_i(j+1))}{\rho_i\sin(\omega_i)}\\ &\leq (j+1)\sqrt{c_iv_i},\ \text{using the inequality } |\sin(k\omega_i)| \leq k|\sin(\omega_i)|\end{aligned}$$

and $\left\|P_{i,k}\begin{pmatrix}v_i^{1/2}\\0\end{pmatrix}\right\| \leq \frac{\sqrt{c_iv_i}(k+1)k}{2}$.
This gives for the Variance term

$$\sum_{k=1}^n \operatorname{tr} P_{i,k}V_iP_{i,k} \leq v_ic_i\sum_{k=1}^n \min\left\{\frac{\left[1-\rho_i^kB_{1,k}\right]^2}{\gamma^2(s_i+\lambda)^2}, \frac{\left[1-\rho_i^kB_{1,k}\right]k(k+1)}{2\gamma(s_i+\lambda)}, \frac{k^2(k+1)^2}{4}\right\}$$

$$\leq \quad v_i c_i \min\left\{\frac{8n}{\gamma^2(s_i+\lambda)^2}, \frac{(n+1)^3}{2\gamma(s_i+\lambda)}, \frac{(n+1)^5}{20}\right\}. \tag{3.28}$$

One Coalescent Eigenvalue

We now turn to the case where F has two coalescent eigenvalues, which happens when the discriminant $\Delta = 0$. We assume that F_i has one coalescent eigenvalue $r_i = \frac{(1+\delta)(1-\gamma(s_i+\lambda))}{2}$. Then, with $\delta = \frac{1-\sqrt{\gamma(s_i+\lambda)}}{1+\sqrt{\gamma(s_i+\lambda)}}$, $r_i = \frac{(1+\delta)(1-\gamma(s_i+\lambda))}{2} = 1-\sqrt{\gamma(s_i+\lambda)}$. Then F_i can be trigonalized as $F_i = Q_i D_i Q_i^{-1}$ with $Q_i = \begin{pmatrix} r_i & 1 \\ 1 & 0 \end{pmatrix}$, $D_i = \begin{pmatrix} r_i & 1 \\ 0 & r_i \end{pmatrix}$ and $Q_i^{-1} = \begin{pmatrix} 0 & 1 \\ 1 & -r_i \end{pmatrix}$. We note that for all $k \geq 0$, then $D_i^k = r_i^{k-1}\begin{pmatrix} r_i & k \\ 0 & r_i \end{pmatrix}$.

Computing $P_{i,k}$. We first compute $P_{i,k}$:

$$(I_2 - D_i)^{-1} = \begin{pmatrix} \frac{1}{1-r_i} & \frac{1}{(1-r_i)^2} \\ 0 & \frac{1}{1-r_i} \end{pmatrix}$$

and

$$(I_2 - D_i^k)(I_2 - D_i)^{-1} = \begin{pmatrix} \frac{1-r_i^k}{1-r_i} & \frac{1-r_i^k}{(1-r_i)^2} - \frac{kr_i^{k-1}}{1-r_i} \\ 0 & \frac{1-r_i^k}{1-r_i} \end{pmatrix}.$$

Thus with $C_i^{1/2}Q_i = \begin{pmatrix} \sqrt{c_i}r_i & \sqrt{c_i} \\ 0 & 0 \end{pmatrix}$ we have

$$C_i^{1/2}Q_i(I_2 - D_i^k)(I_2 - D_i)^{-1} = \sqrt{c_i}\begin{pmatrix} \frac{1-r_i^k}{1-r_i}r_i & \frac{1-r_i^k}{(1-r_i)^2} - \frac{kr_i^k}{1-r_i} \\ 0 & 0 \end{pmatrix}.$$

And, computing as previously the matrices products, we derive:

$$\begin{aligned} P_{i,k} &= C_i^{1/2}Q_i(I_2 - D_i^k)(I_2 - D_i)^{-1}Q_i^{-1} \\ &= \sqrt{c_i}\begin{pmatrix} \frac{1-r_i^k}{(1-r_i)^2} - \frac{kr_i^k}{1-r_i} & \frac{1-r_i^k}{1-r_i}r_i - (\frac{1-r_i^k}{(1-r_i)^2} - \frac{kr_i^k}{1-r_i})r_i \\ 0 & 0 \end{pmatrix} \\ &= \sqrt{c_i}\begin{pmatrix} \frac{1-r_i^k}{(1-r_i)^2} - \frac{kr_i^k}{1-r_i} & \frac{1-r_i^k}{(1-r_i)^2}(r_i)^2 + \frac{kr_i^{k+1}}{1-r_i} \\ 0 & 0 \end{pmatrix} \\ &= \frac{\sqrt{c_i}}{1-r_i}\begin{pmatrix} \frac{1-r_i^k}{1-r_i} - kr_i^k & -\frac{1-r_i^k}{1-r_i}(r_i)^2 + kr_i^{k+1} \\ 0 & 0 \end{pmatrix}. \end{aligned}$$

Bias term. We thus have:

$$P_{i,k}\Theta_0^i = \frac{\sqrt{c_i}}{1-r_i}\begin{pmatrix} \frac{1-r_i^k}{1-r_i} - kr_i^k & -\frac{1-r_i^k}{1-r_i}(r_i)^2 + kr_i^{k+1} \\ 0 & 0 \end{pmatrix}\begin{pmatrix} \theta_0^i \\ \theta_0^i \end{pmatrix}$$

$$= \theta_0^i \sqrt{c_i} \begin{pmatrix} (1-r_i^k)\frac{1+r_i}{1-r_i} - kr_i^k \\ 0 \end{pmatrix},$$

and this gives for the bias term:

$$\begin{aligned} \|P_{i,k}\Theta_0^i\|^2 &= (\theta_0^i)^2 c_i \Big[(1-r_i^k)\frac{1+r_i}{1-r_i} - kr_i^k\Big]^2 \\ &= (\theta_0^i)^2 c_i \Big[\frac{1+r_i}{1-r_i} - \Big(k + \frac{1+r_i}{1-r_i}\Big) r_i^k\Big]^2, \end{aligned}$$

developing the product, then using formulas for r_i,

$$\begin{aligned} \|P_{i,k}\Theta_0^i\|^2 &= (\theta_0^i)^2 c_i \Big[\frac{2-\sqrt{\gamma(s_i+\lambda)}}{\sqrt{\gamma(s_i+\lambda)}} - \Big(k + \frac{2-\sqrt{\gamma(s_i+\lambda)}}{\sqrt{\gamma(c_i+\lambda)}}\Big)(1-\sqrt{\gamma(s_i+\lambda)})^k\Big]^2 \\ &= \frac{(\theta_0^i)^2 c_i}{\gamma(s_i+\lambda)} \Big[2-\sqrt{\gamma(s_i+\lambda)} - \big(k\sqrt{\gamma(s_i+\lambda)} + 2 - \sqrt{\gamma(s_i+\lambda)}\big)(1-\sqrt{\gamma(s_i+\lambda)})^k\Big]^2 \\ &= \frac{(\theta_0^i)^2 c_i}{\gamma(s_i+\lambda)} \Big[2 - \sqrt{\gamma(s_i+\lambda)} - \big(2 + (k-1)\sqrt{\gamma(s_i+\lambda)}\big)(1-\sqrt{\gamma(s_i+\lambda)})^k\Big]^2 \\ &\leq 4\frac{(\theta_0^i)^2 c_i}{\gamma(s_i+\lambda)}, \text{ using Lemma 20 in Appendix 3.D.} \end{aligned} \tag{3.29}$$

Variance term. With $V = \begin{pmatrix} v_i & 0 \\ 0 & 0 \end{pmatrix}$,

$$\begin{aligned} &\operatorname{tr} P_{i,k} V P_{i,k} \\ &= \frac{s_i}{(1-r_i)^2} \begin{pmatrix} \frac{1-r_i^k}{1-r_i} - kr_i^k & -\frac{1-r_i^k}{1-r_i}(r_i)^2 + kr_i^{k+1} \\ 0 & 0 \end{pmatrix} \begin{pmatrix} v_i & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \frac{1-r_i^k}{1-r_i} - kr_i^k & -\frac{1-r_i^k}{1-r_i}(r_i)^2 + kr_i^{k+1} \\ 0 & 0 \end{pmatrix}^\top \\ &= \frac{s_i v_i}{(1-r_i)^2} \Big[\frac{1-r_i^k}{1-r_i} - kr_i^k\Big]^2 \\ &= \frac{v_i h_i}{\gamma(s_i+\lambda)} \Big[\frac{1-r_i^k}{1-r_i} - kr_i^k\Big]^2 \\ &= \frac{v_i h_i}{\gamma(s_i+\lambda)(1-r_i)^2} \Big[1 - r_i^k - (1-r_i)kr_i^k\Big]^2 \\ &= \frac{v_i h_i}{\gamma^2(s_i+\lambda)^2} \Big[1 - (1 + k\sqrt{\gamma(s_i+\lambda)})(1-\sqrt{\gamma(s_i+\lambda)})^k\Big]^2, \end{aligned}$$

and

$$\begin{aligned} \sum_{k=1}^n \operatorname{tr} P_{i,k} V P_{i,k} &= \frac{v_i s_i}{\gamma^2(s_i+\lambda)^2} \sum_{k=1}^n \Big[1 - (1 + k\sqrt{\gamma(s_i+\lambda)})(1-\sqrt{\gamma(s_i+\lambda)})^k\Big]^2 \\ &\leq n\frac{v_i s_i}{\gamma^2(s_i+\lambda)^2} \text{ using Lemma 20 in Appendix 3.D.} \end{aligned} \tag{3.30}$$

Alternative bounds for the bias and the variance term, as in Equations(3.24), (3.27) may be derived as well. Combining all these results, we are now able to state Theorem 8.

3.C.4 Conclusion

Combining results from Lemma 13, and Equations (3.24), (3.27), (3.28), with $c = \Sigma$, and using the following simple facts:

— For the least squares regression function, with $c = \Sigma$, $\mathbb{E}\langle \bar{\Theta}_n, C\bar{\Theta}_n\rangle = \mathbb{E}f(\bar{\theta}_n) - f(\theta_*)$.

— Under assumption $\mathcal{A}_3$, $\mathcal{A}_4$, we have $V \preccurlyeq \tau^2\Sigma$.

— The squared norm of a vector is the sum of its squared components on the orthonormal eigenbasis. For example $\|P_{n+1}\Theta_0\|^2 = \sum_{i=1}^d \|P_{i,n+1}\Theta_0^i\|^2$.

— For any regularization parameter $\lambda \in \mathbb{R}_+$ and for any constant step-size $\gamma(\Sigma + \lambda I) \preccurlyeq I$, for any $\delta \in \left[\frac{1-\sqrt{\gamma\lambda}}{1+\sqrt{\gamma\lambda}}, 1\right]$, matrix F will have only two distinct complex eigenvalues or two coalescent eigenvalues.

Proposition 7. *Under ($\mathcal{A}_{4,5}$), for any regularization parameter $\lambda \in \mathbb{R}_+$ and for any constant step-size $\gamma(\Sigma + \lambda I) \preccurlyeq I$ we have for any $\delta \in \left[\frac{1-\sqrt{\gamma\lambda}}{1+\sqrt{\gamma\lambda}}, 1\right]$, for the recursion in Eq. (3.10):*

$$\begin{aligned}
\mathbb{E}f(\bar{\theta}_n) - f(\theta_*) \;\leq\; & 2\lambda\|\lambda^{1/2}\Sigma^{1/2}(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)\|^2 \\
& + \sum_{i=1}^d \frac{2}{(n+1)^2}\min\left\{36\frac{c_i(\tilde{\theta}_0^i)^2}{\gamma(s_i+\lambda)}, 6n(1+e^{-1})\frac{c_i(\tilde{\theta}_0^i)^2}{\sqrt{\gamma(s_i+\lambda)}}, \right. \\
& \qquad \left. n^2(1+e^{-1})^2 c_i(\tilde{\theta}_0^i)^2\right\} \\
& + \sum_{i=1}^d \frac{\gamma^2}{(n+1)^2} v_i c_i \min\left\{\frac{8n}{\gamma^2(s_i+\lambda)^2}, \frac{(n+1)^3}{2\gamma(s_i+\lambda)}, \frac{(n+1)^5}{20}\right\}.
\end{aligned}$$

This implies, using the Equation (3.25) for the initial point, using $c_i = \sigma_i$ and regrouping sums as traces or norms:

$$\begin{aligned}
\mathbb{E}f(\bar{\theta}_n) - f(\theta_*) \leq\; & 2\lambda\|\lambda^{1/2}\Sigma^{1/2}(\Sigma+\lambda I)^{-1}(\theta_0-\theta_*)\|^2 \\
& + 2\min\left\{\frac{36\|\Sigma^{1/2}(\Sigma+\lambda I)^{-1/2}(\theta_0-\theta_*)\|^2}{\gamma(n+1)^2}, (1+e^{-1})^2\|\Sigma^{1/2}(\theta_0-\theta_*)\|^2\right\} \\
& + \min\left\{\frac{8\operatorname{tr}(V\Sigma(\Sigma+\lambda I)^{-2})}{n+1}, n\gamma\operatorname{tr}(V\Sigma(\Sigma+\lambda I)^{-1})\right\},
\end{aligned}$$

which gives exactly Theorem 8 using $V \preccurlyeq \tau^2\Sigma$ in the Variance term, and $\lambda^{1/2}(\Sigma + \lambda I)^{-1/2} \preccurlyeq I$ in the first term.

3.D Technical Lemmas

The following sequence of Lemmas appear in the proof. They are mostly independent and rely on simple calculations.

Lemma 16. *The operator* $\big[(\Sigma+\lambda I)\otimes I+I\otimes(\Sigma+\lambda I)\big]^{-1}$ *is a non-decreasing operator on* $(S_n, \preccurlyeq)$.

Proof. Lemma means that for two matrices $M, N \in S_n(\mathbb{R})$ such that $M \preccurlyeq N$, then

$$\big[(\Sigma+\lambda I)\otimes I+I\otimes(\Sigma+\lambda I)\big]^{-1}M \preccurlyeq \big[(\Sigma+\lambda I)\otimes I+I\otimes(\Sigma+\lambda I)\big]^{-1}N.$$

It is equivalent to show that for any symmetric positive matrix $A \in S_n^+$,

$$\big[(\Sigma+\lambda I)\otimes I+I\otimes(\Sigma+\lambda I)\big]^{-1}A \in S_n^+(\mathbb{R}).$$

We consider a matrix $A \in S_n^+(\mathbb{R})$. A can be decomposed as a sum of (at most) n rank one matrices $A = \sum_{i=1}^n \omega_i\omega_i^\top$, with $\omega_i \in \mathbb{R}^n$. We thus just have to prove that for some $\omega \in \mathbb{R}^n$, $\big[(\Sigma+\lambda I)\otimes I+I\otimes(\Sigma+\lambda I)\big]^{-1}\omega\omega^\top \in S_n^+(\mathbb{R})$.

Let $\Sigma = \sum_{i\geqslant 0}\mu_i e_i\otimes e_i$ is the eigenvalue decomposition of Σ, then

$$\big[(\Sigma+\lambda I)\otimes I+I\otimes(\Sigma+\lambda I)\big]^{-1}\omega\omega^\top = \sum_{i,j\geqslant 0}\frac{\langle\omega,e_i\rangle\langle\omega,e_j\rangle}{\mu_i+\mu_j+2\lambda}e_i\otimes e_j.$$

Thus, in the orthonormal basis of eigenvectors, this is thus Hadamard product between

$$\sum_{i,j\geqslant 0}\langle\omega,e_i\rangle\langle\omega,e_j\rangle e_i\otimes e_j = \omega\omega^\top$$

and the matrix $C = \Big(\big(\frac{1}{\mu_i+\mu_j+2\lambda}\big)_{i,j\geqslant 0}\Big)$. Matrix C is a Cauchy matrix and is thus positive. Moreover the Hadamard product of two positive matrices is positive, which concludes the proof. □

Remark: surprisingly, the inverse operator $(\Sigma+\lambda I)\otimes I+I\otimes(\Sigma+\lambda I)$ is not non-decreasing. Indeed, $\preccurlyeq$ is not a total order on S_n so we may have that an operator is non-decreasing and its inverse is not.

Lemma 17. *For all* $\rho \in (0,1)$ *and* $\omega \in [-\pi/2;\pi/2]$ *and* $r^\pm = \rho(\cos(\omega)\pm\sqrt{-1}\sin(\omega))$ *we have:*

$$\left|\frac{1-r^+r^--\rho^k|A_1|}{|1-r^+|}\right| \leq \min\{1+\rho+e^{-1}+4\rho^k, 2+\rho+\sqrt{5}\rho^{k+1}\} \leq 6. \tag{3.31}$$

Proof. We note that $\rho_i^k A_1$ is a real number as is is a quotient of pure complex numbers, which come from the difference between a complex and its conjugate. We first write A_1 as a combination of sine and cosine functions:

$$\rho_i^k A_1 \;=\; \frac{(r_i^-)^{k+1}(1-r_i^+)^2-(r_i^+)^{k+1}(1-r_i^-)^2}{r_i^- - r_i^+}$$

$$= -\frac{(r_i^-)^{k+1} - (r_i^+)^{k+1} - 2r_i^- r_i^+((r_i^-)^k - (r_i^+)^k) + (r_i^- r_i^+)(r_i^-)^{k-1} - (r_i^+)^{k-1})}{\rho_i \sin\omega_i}$$
$$= -\frac{\rho_i^{k+1}\sin((k+1)\omega_i) - 2\rho_i^{k+2}\sin(k\omega_i) + \rho_i^{k+3}\sin((k-1)\omega_i)}{\rho_i \sin\omega_i}.$$

This quantity can be simplified when $\rho \to 1$ or $\omega \to 0$. We thus modify the expression of A_1 to make these dependencies clearer:

$$-A_1 = \frac{\sin((k+1)\omega_i) - 2\rho_i\sin(k\omega_i) + \rho_i^2\sin((k-1)\omega_i)}{\sin\omega_i}$$
$$= \frac{(\cos(\omega)-\rho)(\sin(k\omega)-\rho\sin((k-1)\omega)) + \cos(k\omega)\sin(\omega) - \rho\cos((k-1)\omega)\sin(\omega)}{\sin\omega_i}$$
developing $\sin(a+b) = \sin(a)\cos(b) + \cos(a)\sin(b)$ and regrouping terms,
$$= \frac{(\cos(\omega)-\rho)^2\sin((k-1)\omega) + (\cos(\omega)-\rho)\sin(\omega)\cos((k-1)\omega)}{\sin\omega_i}$$
$$+\frac{\cos(k\omega)\sin(\omega) - \rho\cos((k-1)\omega)\sin(\omega)}{\sin\omega_i}$$
$$= \frac{(\cos(\omega)-\rho)^2\sin((k-1)\omega)}{\sin\omega_i} + (\cos(\omega)-\rho)\cos((k-1)\omega) + \cos(k\omega) - \rho\cos((k-1)\omega)$$
simplifying expression, then developing the cosine,
$$= \frac{(\cos(\omega)-\rho)^2\sin((k-1)\omega)}{\sin\omega_i} + 2(\cos(\omega)-\rho)\cos((k-1)\omega) + \sin(\omega)\sin((k-1)\omega) \quad (3.32)$$

So that in that final expression all the terms behave relatively simply when $\rho \to 1$ or $\omega \to 0$. We want to upper bound:

$$\left|\frac{1 - r^+r^- - \rho^k|A_1|}{|1-r^+|}\right|.$$

We thus consider separately the first and second term.

$$\frac{1 - r_i^+ r_i^-}{|1-r_i^+|} = \frac{1-\rho^2}{|1-r_i^+|} \le 1+\rho \qquad \text{(exact if } \omega = 0\text{)}.$$

Then, using Equation (3.32):

$$\frac{-\rho_i^k|A_1|}{|1-r_i^+|} = \rho^k \frac{\frac{(\cos(\omega)-\rho)^2\sin((k-1)\omega)}{\sin\omega_i} + 2(\cos(\omega)-\rho)\cos((k-1)\omega) + \sin(\omega)\sin((k-1)\omega)}{\sqrt{(1-\rho\cos\omega)^2 + \rho^2\sin^2(\omega)}}.$$

Considering separately the three terms in the numerator, using numerous times that for any $a, b \in [0;1]$, $|a-b| \le 1-ab$:

$$\left|\frac{\rho^k\frac{(\cos(\omega)-\rho)^2\sin((k-1)\omega)}{\sin\omega_i}}{\sqrt{(1-\rho\cos\omega)^2 + \rho^2\sin^2(\omega)}}\right| \le \rho^k\frac{(\cos(\omega)-\rho)\sin((k-1)\omega)}{\sin\omega_i}$$

$$
\begin{aligned}
&\text{as } |(\cos(\omega)-\rho)| \le 1-\rho\cos(\omega),\\
\le\ & \rho^k \frac{(\cos(\omega)-1)\sin((k-1)\omega)}{\sin\omega_i} + \rho^k \frac{(1-\rho)\sin((k-1)\omega)}{\sin\omega_i}\\
&\text{writing } \cos(\omega)-\rho = \cos(\omega)-1+1-\rho\\
\le\ & \rho^k(1-\rho)(k-1) + \rho^k \frac{(\cos(\omega)-1)\sin((k-1)\omega)}{\sin\omega_i}\\
&\text{as } |\sin((k-1)\omega)| \le |(k-1)\sin(\omega)|,\\
\le\ & \rho^k(1-\rho)k-(1-\rho)\rho^k+\rho^k \frac{(\cos(\omega)-1)\sin((k-1)\omega)}{\sin\omega_i}\\
&\text{writing } \cos(\omega)-1 = 2\sin^2(\omega/2),\\
\le\ & \rho^k(1+(1-\rho))^k - \rho^k - (1-\rho)\rho^k + \rho^k \frac{2\sin^2(\omega/2)}{\sin\omega_i}\\
&\text{using } 1+(1-\rho)k \le (1+(1-\rho))^k,\\
\le\ & \rho^k(1+(1-\rho))^k - \rho^k - (1-\rho)\rho^k + \rho^k\tan(\omega/2)\\
&\text{and as } \tan(\omega/2) \le 1 \text{ for } |\omega| \le \pi/2,\\
\le\ & 1-(1-\rho)\rho^k\\
&\text{using } \rho^k(1+(1-\rho))^k = (1-(1-\rho)^2)^k \le 1,
\end{aligned}
$$

And for the second and third term:

$$
\begin{aligned}
2\left|\rho^k \frac{(\cos(\omega)-\rho)\cos((k-1)\omega)}{\sqrt{(1-\rho\cos\omega)^2+\rho^2\sin^2(\omega)}}\right| &\le 2\rho^k,\\
\left|\rho^k \frac{+\sin(\omega)\sin((k-1)\omega)}{\sqrt{(1-\rho\cos\omega)^2+\rho^2\sin^2(\omega)}}\right| &\le \rho^k.
\end{aligned}
$$

Thus:

$$
\left|\frac{1-r_i^+r_i^- - \rho_i^k|A_1|}{|1-r_i^+|}\right| \le 1+\rho+1+3\rho^k.
$$

We also have

$$
\begin{aligned}
\left|\frac{\rho^k\frac{(\cos(\omega)-\rho)^2\sin((k-1)\omega)}{\sin\omega_i}}{\sqrt{(1-\rho\cos\omega)^2+\rho^2\sin^2(\omega)}}\right| &\le \rho^k\frac{(\cos(\omega)-\rho)\sin((k-1)\omega)}{\sin\omega_i}\\
&\le \rho^k(1-\rho)(k-1) + \rho^k\frac{(\cos(\omega)-1)\sin((k-1)\omega)}{\sin\omega_i}\\
&\le (1-\frac{1}{k+1})^{k+1} - (1-\rho)\rho^k + \rho^k\frac{(\cos(\omega)-1)\sin((k-1)\omega)}{\sin\omega_i}\\
&\le e^{-1} - (1-\rho)\rho^k + \rho^k\frac{\sin^2(\omega/2)}{\sin\omega_i}.
\end{aligned}
$$

Using that

$$\begin{aligned} k \sup_{x\in[0;1]} x^k(1-x) &= k\frac{1}{k+1}(1-\frac{1}{k+1})^k \\ &= (1-\frac{1}{k+1})^{k+1} \\ &= \exp((k+1)\ln((1-\frac{1}{k+1})) \leq e^{-1}, \end{aligned} \tag{3.33}$$

we get

$$\left|\frac{1-r_i^+r_i^- - \rho_i^k|A_1|}{|1-r_i^+|}\right| \leq 1+\rho+e^{-1}+4\rho^k$$

We can also change $3\rho^k$ into $\sqrt{5}\rho^k$ We have used that $|(\rho-\cos(\omega))| \leq (1-\rho\cos(\omega))$. □

Lemma 18. *For any $\rho_i \in (0;1)$, for any $\omega_i \in [-\pi/2;\pi/2]$*

$$\frac{\rho_i^j \sin(\omega_i(j+1))}{\sin(\omega_i)} - \rho_i^{j+1}\frac{\sin(\omega_i j)}{\sin(\omega_i)} \leq 1+e^{-1}.$$

Proof.

$$\begin{aligned} \frac{\rho_i^j \sin(\omega_i(j+1))}{\sin(\omega_i)} - \rho_i^{j+1}\frac{\sin(\omega_i j)}{\sin(\omega_i)} &= \rho_i^j\left(\frac{\sin(\omega_i(j+1)) - \rho_i\sin(\omega_i j)}{\sin(\omega_i)}\right) \\ &= \rho_i^j\left(\frac{(\cos(\omega_i)-\rho_i)\sin(\omega_i j)}{\sin(\omega_i)} + \cos(j\omega_i)\right) \\ &\leq \rho_i^j\left((1-\rho_i)j+1\right) \\ &\leq 1+e^{-1} \text{ using (3.33).} \end{aligned}$$

□

Lemma 19. *For all $\rho \in (0,1)$ and $\omega \in [-\pi/2;\pi/2]$ and $r^\pm = \rho(\cos(\omega) \pm \sqrt{-1}\sin(\omega))$ we have:*

$$\left|\rho_i^k B_{1,k}\right| \leq 1.75 \tag{3.34}$$

Proof. Once again, as the considered quantity is real, we first express it as a combination of sine and cosine functions. We then use some simple trigonometric trics to upper bound the quantity.

$$\begin{aligned} \rho_i^k B_{1,k} &= -\frac{(r_i^-)^{k+1}(1-r_i^+) - (r_i^+)^{k+1}(1-r_i^-)}{r_i^+ - r_i^-} \\ &= -\frac{2\mathfrak{Im}\left[(r_i^-)^{k+1}(1-r_i^+)\right]}{\sqrt{-\Delta_i}} \text{ as it is the difference between a complex} \end{aligned}$$

$$
\begin{aligned}
& \text{and its conjugate,} \\
= & -\frac{\mathfrak{Im}\big[\rho_i^k e^{-(k+1)i\omega_i}(1-\rho_i\cos(\omega_i)-i\rho_i\sin(\omega_i))\big]}{\sin\omega_i\rho_i} \text{ developing the product,} \\
= & \rho_i^k\frac{\cos((k+1)\omega_i)\sin(\omega_i)\rho_i+\sin((k+1)\omega_i)(1-\rho_i\cos(\omega_i))}{\sin\omega_i\rho_i} \\
= & \rho_i^k\Big[\rho_i\cos((k+1)\omega_i)+(1-\rho_i\cos(\omega_i))\frac{\sin((k+1)\omega_i)}{\sin\omega_i}\Big] \text{ and simplifying.}
\end{aligned}
$$

Let's turn our interest to the second part of the quantity. We denote by

$$
\square = \left|\rho_i^k(1-\rho_i\cos(\omega_i))\frac{\sin((k+1)\omega_i)}{\sin\omega_i}\right|,
$$

and

$$
\begin{aligned}
\square &= \left|\rho_i^k(1-\rho_i+\rho_i(1-\cos(\omega_i)))\frac{\sin((k+1)\omega_i)}{\sin\omega_i}\right| \\
&\quad \text{introducing an artificial } +\rho_i-\rho_i, \\
&\leq \rho_i^k\left|(1-\rho_i)\frac{\sin((k+1)\omega_i)}{\sin\omega_i}\right|+\rho_i^k\left|\rho_i(1-\cos(\omega_i))\frac{\sin((k+1)\omega_i)}{\sin\omega_i}\right| \\
&\quad \text{by triangular inequality,} \\
&\leq \rho_i^k\left|(1-\rho_i)(k+1)\right|+\rho_i^k\left|\rho_i\sin^2(\frac{\omega}{2})\frac{1}{2\cos(\frac{\omega}{2})\sin(\frac{\omega}{2})}\right| \\
&\quad \text{using } 1-\cos(\omega_i)=2\sin^2(\frac{\omega}{2}) \\
&\leq \rho_i^k(1-\rho_i)k+\rho_i^k(1-\rho)+\rho_i^k\left|\rho_i\sin^2(\frac{\omega}{2})\frac{1}{2\cos(\frac{\omega}{2})\sin(\frac{\omega}{2})}\right| \\
&\leq (1-(1-\rho_i))^k(1+(1-\rho_i))^k-\rho_i^k+\frac{1}{2(k+1)}+\rho_i^k\left|\frac{\rho_i}{2}\tan(\frac{\omega}{2})\right| \\
&\leq (1-(1-\rho_i)^2)^k+\frac{1}{4}+\frac{1}{2}\leq 1+\frac{1}{4}+\frac{1}{2}-\rho_i^k.
\end{aligned}
$$

Thus

$$
\left|\rho_i^k B_{1,k}\right| = \rho_i^k+1+\frac{1}{4}+\frac{1}{2}-\rho_i^k\leq 1+\frac{1}{4}+\frac{1}{2}=1.75.
$$

□

Lemma 20. *For any $s_i,\gamma,\lambda\in\mathbb{R}_+^3$ such that $\gamma(s_i+\lambda)\leq 1$, for any $k\in\mathbb{N}$, we have the two following highly related identities:*

$$
\begin{aligned}
0\leq 2-\sqrt{\gamma(s_i+\lambda)}-\big(2+(k-1)\sqrt{\gamma(s_i+\lambda)}\big)(1-\sqrt{\gamma(s_i+\lambda)})^k &\leq 2 \\
0\leq 1-(1+k\sqrt{\gamma(s_i+\lambda)})(1-\sqrt{\gamma(s_i+\lambda)})^k &\leq 1.
\end{aligned}
$$

Proof. Proof relies on the trick, for any $\alpha\in\mathbb{R},n\in\mathbb{N}$: $1+n\alpha\leq(1+\alpha)^n$. For the

first one:

$$\begin{aligned}
&\sqrt{\gamma(s_i+\lambda)} + \big(2 + (k-1)\sqrt{\gamma(s_i+\lambda)}\big)(1-\sqrt{\gamma(s_i+\lambda)})^k = \\
=\ &\sqrt{\gamma(s_i+\lambda)} + (1-\sqrt{\gamma(s_i+\lambda)})^k + \big(1 + (k-1)\sqrt{\gamma(s_i+\lambda)}\big)(1-\sqrt{\gamma(s_i+\lambda)})^k \\
\leq\ &\sqrt{\gamma(s_i+\lambda)} + (1-\sqrt{\gamma(s_i+\lambda)}) + \big(1 + (k-1)\sqrt{\gamma(s_i+\lambda)}\big)(1-\sqrt{\gamma(s_i+\lambda)})^{k-1} \\
\leq\ &1 + (1-\gamma(s_i+\lambda))^{k-1} \leq 2.
\end{aligned}$$

For the second one:

$$0 \leq (1 + k\sqrt{\gamma(s_i+\lambda)})(1-\sqrt{\gamma(s_i+\lambda)})^k \ \leq\ (1-\gamma(s_i+\lambda))^k \leq 1.$$

□

Chapter 4

Dual Averaging Algorithm for Composite Least-Squares Problems

Abstract

We consider the minimization of composite objective functions composed of the expectation of quadratic functions and an arbitrary convex function. We study the stochastic dual averaging algorithm with a constant step-size, showing that it leads to a convergence rate of $O(1/n)$ without strong convexity assumptions. This thus extends earlier results on least-squares regression with the Euclidean geometry to (a) all convex regularizers and constraints, and (b) all geometries represented by a Bregman divergence. This is achieved by a new proof technique that relates stochastic and deterministic recursions.

This chapter is extracted from the paper: Stochastic Composite Least-Squares Regression with convergence rate $O(1/n)$, in collaboration with F. Bach published in the *Proceedings of the International Conference on Learning Theory (COLT)*, 2017.

4.1 Introduction

Many learning problems may be cast as the optimization of an objective function defined as an expectation of random functions, and which can be accessed only through samples. In this chapter, we consider *composite* problems of the form

$$\min_{\theta \in \mathbb{R}^d} \mathbb{E}_z \ell(z, \theta) + g(\theta), \tag{4.1}$$

where for any z, $\ell(z, \cdot)$ is a convex quadratic function (plus some linear terms) and g is any extended-value convex function.

In a machine learning context, $\ell(z, \theta)$ is the loss occurred for the observation z and the predictor parameterized by θ, $f(\theta) = \mathbb{E}_z \ell(z, \theta)$ is its generalization error, while the function g represents some additional regularization or constraints on the predictor. Thus in this chapter we consider composite least-squares regression problems, noting that solving such problems effectively leads to efficient algorithms for all smooth

losses by using an online Newton algorithm [Bach and Moulines, 2013], with the same running-time complexity of $O(d)$ per iteration for linear predictions.

When $g = 0$, averaged stochastic gradient descent with a constant step-size achieves the optimal convergence rate of $O(1/n)$ after n observations, even in ill-conditioned settings without strong convexity [Jain et al., 2016], with precise non-asymptotic results that depend on the statistical noise variance σ^2 of the least-squares problem, as $\sigma^2 d/n$, and on the squared Euclidean distance between the initial predictor θ_0 and the optimal predictor θ_*, as $\|\theta_0 - \theta_*\|_2^2/n$.

In this chapter, we extend this $O(1/n)$ convergence result in two different ways:

— **Composite problems**: we provide a new algorithm that deals with composite problems where g is (essentially) any extended-value convex function, such as the indicator function of a convex set for constrained optimization, or a norm or squared norm for additional regularization. This situation is common in many applications in machine learning and signal processing [see, e.g., Rish and Grabarnik, 2014, and references therein]. Because we consider large steps-sizes (that allow robustness to ill-conditioning), the new algorithm is *not* simply a proximal extension; for example, in the constrained case, averaged projected stochastic gradient descent with a constant step-size is not convergent, even for quadratic functions.

— **Beyond Euclidean geometry**: Following mirror descent [Nemirovski and Yudin, 1979] and recent work of Bauschke et al. [2016], our new algorithm can take into account a geometry obtained with a Bregman divergence D_h associated with a convex function h, which can typically be the squared Euclidean norm (leading to regular stochastic gradient descent in the non-composite case), the entropy function, or the squared ℓ_p-norm. This will allow convergence rates proportional to $D_h(\theta_*, \theta_0)/n$, which may be significantly smaller than $\|\theta_0 - \theta_*\|^2/n$ in many situations.

In order to obtain these two extensions, we consider the stochastic dual averaging algorithm of Nesterov [2009] and Xiao [2010] which we present in Section 4.2, and study under the particular set-up of *constant step-size with averaging*, showing in Section 4.3 that it also achieves a convergence rate of $O(1/n)$ even without strong convexity. This is achieved by a new proof technique that relates stochastic and deterministic recursions.

Given that known lower-bounds for this class of problems are proportional to $1/\sqrt{n}$ for function values, we established our $O(1/n)$ results with a different criterion, namely the Mahalanobis distance associated with the Hessian of the least-squares problem. In our simulations in Section 4.5, the two criteria behave similarly. Finally, in Section 4.4, we shed additional insights of the relationships between mirror descent and dual averaging, in particular in terms of continuous-time interpretations.

4.2 Dual Averaging Algorithm

In this section, we introduce dual averaging as well as related frameworks, together with new results in the deterministic case.

4.2.1 Assumptions

We consider the Euclidean space $\mathbb{R}^d$ of dimension d endowed with the natural inner product $\langle \cdot, \cdot \rangle$ and an arbitrary norm $\| \cdot \|$ (which may not be the Euclidean norm). We denote by $\| \cdot \|_*$ its dual norm and for any symmetric positive-definite matrix A, by $\| \cdot \|_A = \sqrt{\langle \cdot, A \cdot \rangle}$ the Mahalanobis norm. For a vector $\theta \in \mathbb{R}^d$, we denote by $\theta(i)$ its i-th coordinate and by $\|\theta\|_p = (\sum_{i=1}^d |\theta(i)|^p)^{1/p}$ its ℓ_p-norm. We also denote the convex conjugate of a function f by $f^*(\eta) = \sup_{\theta \in \mathbb{R}^d} \langle \eta, \theta \rangle - f(\theta)$. We remind that a function f is L-smooth with respect to a norm $\| \cdot \|$ if for all $(\alpha, \beta) \in \mathbb{R}^d \times \mathbb{R}^d$, $\|\nabla f(\alpha) - \nabla f(\beta)\|_* \leq L \|\alpha - \beta\|$ and is μ-strongly convex if for all $(\alpha, \beta) \in \mathbb{R}^d \times \mathbb{R}^d$ and $g \in \partial f(\beta)$, $f(\alpha) \geq f(\beta) + \langle g, \alpha - \beta \rangle + \frac{\mu}{2} \|\alpha - \beta\|^2$ [see, e.g., Shalev-Shwartz and Singer, 2006].

We consider problems of the form:

$$\min_{\theta \in \mathcal{X}} \psi(\theta) = f(\theta) + g(\theta), \tag{4.2}$$

where $\mathcal{X} \subset \mathbb{R}^d$ is a closed convex set with non empty interior. Throughout this chapter, we make the following general assumptions:

(A1) $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a proper lower semicontinuous convex function and is differentiable on $\mathring{\mathcal{X}}$ (the interior of $\mathcal{X}$).

(A2) $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a proper lower semicontinuous convex function.

(A3) $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ with $\overline{\text{dom } h} \cap \overline{\text{dom } g} = \mathcal{X}$, $\mathring{\text{dom}}\, h \cap \text{dom } g \neq \emptyset$. Moreover h is a Legendre function [Rockafellar, 1970, chap. 26]:

- h is a proper lower semicontinuous strictly convex function, differentiable on $\mathring{\text{dom}}\, h$.
- The gradient of h is diverging on the boundary of dom h (i.e., for any sequence (θ_n) converging to a boundary point of $\mathring{\text{dom}}\, h$, $\lim_{n \to +\infty} \|\nabla h(\theta_n)\| = \infty$). Note that ∇h is then a bijection from $\mathring{\text{dom}}\, h$ to $\mathring{\text{dom}}\, h^*$ whose inverse is the gradient of the conjugate ∇h^*.

(A4) The function $\psi = f + g$ attains its minimum over $\mathcal{X}$ at a certain $\theta_* \in \mathbb{R}^d$ (which may not be unique).

Note that we adopt the same framework as Bauschke et al. [2016] with the difference that a convex constraint $\mathcal{C}$ can be handled with more flexibility: either by considering a Legendre function h whose domain is $\mathcal{C}$ or by considering the hard constraint $g(\theta) = \mathbb{1}_C(\theta)$ (equal to 0 if $\theta \in \mathcal{C}$ and $+\infty$ otherwise).

4.2.2 Dual Averaging Algorithm

In this section we present the dual averaging algorithm (referred to from now on as "DA") for solving composite problems of the form of Eq. (4.2). It starts from $\theta_0 \in \mathring{\text{dom}}\, h$ and $\eta_0 = \nabla h(\theta_0)$ and iterates for $n \geq 1$ the recursion

$$\begin{aligned} \eta_n &= \eta_{n-1} - \gamma \nabla f(\theta_{n-1}) \\ \theta_n &= \nabla h_n^*(\eta_n), \end{aligned} \tag{4.3}$$

with $h_n = h + n\gamma g$ and $\gamma \in (0, \infty)$ (commonly referred to as the step-size in optimization or the learning rate in machine learning). We note that equivalently $\theta_n \in \arg\max_{\theta \in \mathbb{R}^d}\{\langle \eta_n, \theta\rangle - h_n(\theta)\}$. When $h = \frac{1}{2}\|\cdot\|_2^2$ and $g = 0$, we recover gradient descent.

Two iterates (η_n, θ_n) are updated in DA. The dual iterate η_n is simply proportional to the sum of the gradients evaluated in the primal iterates (θ_n). The update of the primal iterate θ_n is more complex and raises two different issues: its existence and its tractability. We discuss the first point in Appendix 4.A and assume, as of now, that the method is generally well defined in practice. The tractability of θ_n is essential and the algorithm is only used in practice if the functions h and g are simple in the sense that the gradient ∇h_n^* may be computed effectively. This is the case if there exists a closed form expression. Usual examples are given in Appendix 4.I.

Euclidean case and proximal operators. In the Euclidean case, Eq. (4.3) may be written in term of the proximal operator defined by Moreau [1962] as $\mathrm{Prox}_g(\eta) = \arg\min_{\theta \in \mathcal{X}}\{\frac{1}{2}\|\theta - \eta\|_2^2 + g(\theta)\}$:

$$\theta_n = \operatorname*{arg\,min}_{\theta \in \mathcal{X}}\Big\{\langle -\eta_n, \theta\rangle + n\gamma g(\theta) + \frac{1}{2}\|\theta\|_2^2\Big\} = \operatorname*{arg\,min}_{\theta \in \mathcal{X}}\Big\{\frac{1}{2}\|\theta - \eta_n\|_2^2 + n\gamma g(\theta)\Big\} = \mathrm{Prox}_{\gamma n g}(\eta_n).$$

DA is in this sense related to proximal gradient methods, also called forward-backward splitting methods [see, e.g., Beck and Teboulle, 2009, Wright et al., 2009, Combettes and Pesquet, 2011]. These methods are tailored to composite optimization problems: at each iteration f is linearized around the current iterate θ_n and they consider the following update

$$\theta_{n+1} = \operatorname*{arg\,min}_{\theta \in \mathcal{X}}\Big\{\langle \gamma \nabla f(\theta_n), \theta\rangle + \gamma g(\theta) + \frac{1}{2}\|\theta - \theta_n\|_2^2\Big\} = \mathrm{Prox}_{\gamma g}(\theta_n - \gamma \nabla f(\theta_n)).$$

Note the difference with DA which considers a dual iterate and a proximal operator for the function $n\gamma g$ instead of γg (see additional insights in Section 4.4).

From non-smooth to smooth optimization. DA was initially introduced by Nesterov [2009] to optimize a non-smooth function f with possibly convex constraints ($g = 0$ or $g = \mathbb{1}_{\mathcal{C}}$). It was extended to the general stochastic composite case by Xiao [2010] who defined the iteration as

$$\theta_n = \operatorname*{arg\,min}_{\theta \in \mathcal{X}}\Big\{\frac{1}{n}\sum_{i=0}^{n-1}\langle z_i, \theta\rangle + g(\theta) + \frac{\beta_n}{n}h(\theta)\Big\},$$

where z_i is an unbiased estimate [1] of a subgradient in $\partial f(\theta_i)$ and $(\beta_n)_{n \geq 1}$ a nonnegative and nondecreasing sequence of real numbers. This formulation is equivalent to Eq. (4.3) for constant sequences $\beta_n = 1/\gamma$. Xiao [2010] proved convergence rates of order $O(1/\sqrt{n})$ for convex problems with decreasing step-size $C/\sqrt{n}$ and $O(1/(\mu n))$

1. Their results remain true in the more general setting of online learning.

for problems with μ-strongly convex regularization with constant step-size $1/\mu$. DA was also studied with decreasing step-sizes in the distributed case by Duchi et al. [2012], Dekel et al. [2012], Colin et al. [2016] and combined with the alternating direction method of multipliers (ADMM) by Suzuki [2013]. It was further shown to be very efficient in manifold identification by Lee and Wright [2012] and Duchi and Ruan [2016].

Relationship with mirror descent. The DA method should be associated with its cousin mirror descent algorithm (referred to from now on as "MD"), introduced by Nemirovski and Yudin [1979] for the constrained case and written under its modern proximal form by Beck and Teboulle [2003]

$$\theta_n = \arg\min_{\theta \in \mathcal{X}} \left\{ \gamma \langle \nabla f(\theta_{n-1}), \theta \rangle + D_h(\theta, \theta_{n-1}) \right\},$$

where we denote by $D_h(\alpha, \beta) = h(\alpha) - h(\beta) - \langle \nabla h(\beta), \alpha - \beta \rangle$ the Bregman divergence associated with h. Moreover it was later extended to the general composite case by Duchi et al. [2010]

$$\theta_n = \arg\min_{\theta \in \mathcal{X}} \left\{ \gamma \langle \nabla f(\theta_{n-1}), \theta \rangle + \gamma g(\theta) + D_h(\theta, \theta_{n-1}) \right\}. \tag{4.4}$$

DA was initially motivated by Nesterov [2009] to avoid new gradients to be taken into account with less weight than previous ones. However, as an extension of the Euclidean case, DA essentially differs from MD on the way the regularization component is dealt with. See more comparisons in Section 4.4.

Relationship with online learning. DA was traditionally studied under the online learning setting [Zinkevich, 2003] of regret minimization and is related to the "follow the leader" approach [see, e.g., Kalai and Vempala, 2005] as noted by McMahan [2011]. More generally, the DA method may be cast in the primal-dual algorithmic framework of Shalev-Shwartz and Singer [2006] and Shalev-Shwartz and Kakade [2009].

4.2.3 Deterministic Convergence Result for Dual Averaging

In this section we present the convergence properties of the DA method for optimizing deterministic composite problems of the form in Eq. (4.2), for any smooth function f (see proof in Appendix 4.B).

Proposition 8. *Assume **(A1-4)**. For any step-size γ such that $h - \gamma f$ is convex on $\mathring{\mathcal{X}}$ we have for all $\theta \in \mathcal{X}$*

$$\psi(\theta_n) - \psi(\theta) \leq \frac{D_h(\theta, \theta_0)}{\gamma(n+1)}.$$

Moreover assume $g = 0$, and there exists $\mu \in \mathbb{R}$ such that $f - \mu h$ is also convex on $\mathring{\mathcal{X}}$

then we have for all $\theta \in \mathcal{X}$

$$f(\theta_n) - f(\theta) \leq (1 - \gamma\mu)^n \frac{D_h(\theta, \theta_0)}{\gamma}.$$

We can make the following remarks:

— We adapt the proofs of Chen and Teboule [1993], Bauschke et al. [2016] to the composite case and the DA method by including the regularization component g in the Bregman divergence. If g was differentiable we would simply use $D_{h_n} = D_{h+n\gamma g}$ and prove the following recursion:

$$\begin{aligned} D_{h_n}(\theta_*, \theta_n) - D_{h_{n-1}}(\theta_*, \theta_{n-1}) = &-D_{h_{n-1}}(\theta_n, \theta_{n-1}) + \gamma\langle \nabla f(\theta_{n-1}), \theta_{n-1} - \theta_n \rangle \\ &- \gamma(g(\theta_n) - g(\theta)) - \gamma\langle \nabla f(\theta_{n-1}), \theta_{n-1} - \theta_* \rangle. \end{aligned}$$

Since g is not differentiable, we extend instead the notion of Bregman divergence to the non-smooth case in Appendix 4.B.2 and show the proof works in the same way.

— **Related work**: A result on MD with analogue assumptions was first presented by Bauschke et al. [2016]. DA was analyzed for smooth functions in the non-composite case where $g = 0$, by Dekel et al. [2012] in the stochastic setting and by Lu et al. [2016] in the deterministic setting. The technique to extend the Bregman divergence to analyze the regularization component has its roots in the time-varying potential method in online learning [Cesa-Bianchi and Lugosi, 2006, Chapter 11.6] and the "follow the regularized leader" approach [Abernethy et al., 2008].

— This convergence rate is suboptimal for the class of addressed problems. Indeed accelerated gradient methods achieve the convergence rate of $O(L/n^2)$ in the composite setting [Nesterov, 2013], such a rate being optimal for optimizing smooth functions among first-order techniques that can access only sequences of gradients [Nesterov, 2004].

— Classical results on the convergence of optimization algorithms in non-Euclidean geometries assume on one hand that the function h is strongly convex and on the other hand the function f is Lipschitz or smooth. Following Bauschke et al. [2016], we consider a different assumption which combines the smoothness of f and the strong convexity of h on the single condition $h - \gamma f$ convex. For the Euclidean geometry where $h(\theta) = \frac{1}{2}\|\theta\|_2^2$, this condition is obviously equivalent to the smoothness of the function f with regards to the ℓ_2-norm. Moreover, under arbitrary norm $\|\cdot\|$, this is also equivalent to assuming h μ-strongly convex and f L-smooth (with respect to this norm). However it is much more general and may hold even when f is non-smooth, which precisely justifies the introduction of this condition [see examples described by Bauschke et al., 2016].

— The bound adapts to the geometry of the function h through the Bregman divergence between the starting point θ_0 and the solution θ_* and the step-size γ which is controlled by h. Therefore the choice of h influences the constant in

the bound. Examples are provided in Appendix 4.I.

4.3 Stochastic Convergence Results for Quadratic Functions

In this section, we consider a symmetric positive semi-definite matrix $\Sigma \in \mathbb{R}^{d\times d}$ and a convex quadratic function f defined as

$$(\mathbf{A5}) \qquad f(\theta) = \tfrac{1}{2}\langle \theta, \Sigma\theta\rangle - \langle q, \theta\rangle, \quad \text{with } q \in \mathbb{R}^d \text{ in the column space of } \Sigma,$$

so that f has a global minimizer $\theta_\Sigma \in \mathbb{R}^d$. Without loss of generality [2], Σ is assumed invertible, though its eigenvalues could be arbitrarily small. The global solution is known to be $\theta_\Sigma = \Sigma^{-1}q$, but the inverse of the Hessian is often too expensive to compute when d is large. The function may be simply expressed as $f(\theta_n) = \frac{1}{2}\langle \theta_n - \theta_\Sigma, \Sigma(\theta_n - \theta_\Sigma)\rangle + f(\theta_\Sigma)$ and the excess of the cost function $\psi = f + g$ as

$$\psi(\theta_n) - \psi(\theta_*) = \langle \theta_* - \theta_\Sigma, \Sigma(\theta_n - \theta_*)\rangle + g(\theta_n) - g(\theta_*) \text{ (linear part)} \\ + \frac{1}{2}\langle \theta_n - \theta_*, \Sigma(\theta_n - \theta_*)\rangle \text{ (quadratic part)} .$$

The first-order condition of the optimization problem in Eq. (4.2) is $0 \in \nabla f(\theta_*) + \partial g(\theta_*) + \partial \mathbb{1}_{\mathcal{X}}(\theta_*)$. By convexity of g, we have $g(\theta_n) - g(\theta_*) \geq \langle z, \theta_n - \theta_*\rangle$ for any $z \in \partial g(\theta_*)$. Moreover $\mathbb{1}_{\mathcal{X}}(\theta_n) - \mathbb{1}_{\mathcal{X}}(\theta_*) = 0 \geq \langle z, \theta_n - \theta_*\rangle$ for any $z \in \partial \mathbb{1}_{\mathcal{X}}(\theta_*)$ since $\theta_n, \theta_* \in \mathcal{X}$ by definition. Therefore this implies that the linear part $g(\theta_n) - g(\theta_*) + \langle \nabla f(\theta_*), \theta_n - \theta_*)\rangle$ is non-negative and we have the bound

$$\frac{1}{2}\|\theta_n - \theta_*\|_\Sigma^2 \leq \psi(\theta_n) - \psi(\theta_*). \tag{4.5}$$

We derive, in this section, convergence results in terms of the distance $\|\theta_n - \theta_*\|_\Sigma$ which takes into account the ill-conditioning of the matrix Σ and is a lower bound in the excess of function values. Furthermore it directly implies classical results for strongly convex problems.

In many practical situations, the gradient of f is not available for the recursion in Eq. (4.3), and we have only access to an unbiased estimate $\nabla f_{n+1}(\theta_n)$ of the gradient of f at θ_n. We consider in this case the stochastic dual averaging method (referred to from now on as "SDA") defined the same way as DA as

$$\begin{aligned} \eta_n &= \eta_{n-1} - \gamma \nabla f_n(\theta_{n-1}) \\ \theta_n &= \nabla h_n^*(\eta_n), \end{aligned} \tag{4.6}$$

for $\theta_0 \in \overset{\circ}{\text{dom}}\, h$ and $\eta_0 = \nabla h(\theta_0)$. Here we consider the stochastic approximation

2. By decomposing θ in $\theta = \theta_\| + \theta_\perp$ with $\theta_\perp \in \text{Null}(\Sigma)$ and $\langle \theta_\perp, \theta_\|\rangle = 0$ and considering $\psi(\theta) = f(\theta_\|) + \tilde{g}(\theta_\|)$ where $\tilde{g}(\theta_\|) = \inf_{\theta_\perp \in \text{Null}(\Sigma)} g(\theta_\perp + \theta_\|)$.

framework [Kushner and Yin, 2003]. That is, we let $(\mathcal{F}_n)_{n\geq 0}$ be an increasing family of σ-fields such that for each $\theta \in \mathbb{R}^d$ and for all $n \geq 1$ the random variable $\nabla f_n(\theta)$ is square-integrable and $\mathcal{F}_n$-measurable with $\mathbb{E}[\nabla f_n(\theta)|\mathcal{F}_{n-1}] = \nabla f(\theta)$. This includes (but also extends) the usual machine learning situation where ∇f_n is the gradient of the loss associated with the n-th independent observation. We will consider in the following two different gradient oracles.

4.3.1 Additive Noise

We study here the convergence of the SDA recursion defined in Eq. (4.6) under an additive noise model:

(A6) For all $n \geq 1$, $\nabla f_n(\theta) = \nabla f(\theta) - \xi_n$, where the noise $(\xi_n)_{n\geq 1}$ is a square-integrable martingale difference sequence (i.e., $\mathbb{E}[\xi_n|\mathcal{F}_{n-1}] = 0$) with bounded covariance $\mathbb{E}[\xi_n \otimes \xi_n] \preccurlyeq C$.

With this oracle and for the quadratic function f, SDA takes the form

$$\begin{aligned} \eta_n &= \eta_{n-1} - \gamma(\Sigma\theta_{n-1} - q) + \gamma\xi_n \\ \theta_n &= \nabla h_n^*(\eta_n). \end{aligned} \tag{4.7}$$

We obtain the following convergence result on the average $\bar{\theta}_n = \frac{1}{n}\sum_{k=0}^{n-1}\theta_k$ which is an extension of results from Bach and Moulines [2013] to non-Euclidean geometries and to composite settings (see proof in Appendix 4.C).

Proposition 9. *Assume **(A2-6)**. Consider the recursion in Eq. (4.7) for any constant step-size γ such that $h - \gamma f$ is convex. Then*

$$\frac{1}{2}\mathbb{E}\|\bar{\theta}_n - \theta_*\|_\Sigma^2 \leq 2\min\left\{\frac{D_h(\theta_*,\theta_0)}{\gamma n}; \frac{\|\nabla h(\theta_0) - \nabla h(\theta_*)\|_{\Sigma^{-1}}^2}{(\gamma n)^2}\right\} + \frac{4}{n}\operatorname{tr}\Sigma^{-1}C.$$

We can make the following observations:

— The proof in the Euclidean case [Bach and Moulines, 2013] highly uses the equality $\theta_n - \theta_\Sigma = (I - \gamma\Sigma)(\theta_{n-1} - \theta_\Sigma)$ which is no longer available in the non-Euclidean or proximal cases. Instead we adapt the classic proof of convergence of averaged SGD of Polyak and Juditsky [1992] which rests upon the expansion $\sum_{k=0}^n \nabla f_{k+1}(\theta_k) = \sum_{k=0}^n(\eta_k - \eta_{k+1})/\gamma = (\eta_0 - \eta_{n+1})/\gamma$. The crux of the proof is then to consider the difference between the iterations with and without noise, $\eta_n^{\rm sto} - \eta_n^{\rm det}$, which happens to satisfy a similar recursion as Eq. (4.7) but started from the solution θ_*. The quadratic nature of f is used twice: (a) to bound $\|\eta_n^{\rm sto} - \eta_n^{\rm det}\|_{\Sigma^{-1}} \sim \sqrt{n}$, and (b) to expand $\nabla f(\bar{\theta}_n) = \overline{\nabla f(\theta_n)} \sim \frac{\eta_n^{\rm sto} - \eta_0}{\gamma n} + 1/\sqrt{n}$.

— As for Proposition 8, the constraint on the step-size γ depends on the function h. Moreover the step-size γ is constant, contrary to previous works on SDA [Xiao, 2010] which prove results for decreasing step-size $\gamma_n = C/\sqrt{n}$ for the convex case (and with a convergence rate of only $O(1/\sqrt{n})$).

— The first term is the "bias" term. It only depends on the "distance" from the initial point θ_0 to the solution θ_* as the minimum of two terms. The first one

recovers the deterministic bound of Proposition 8. The second one, specific to quadratic objectives, leads to an accelerated rate of $O(1/n^2)$ for some good starting points such that $\|\nabla h(\theta_0) - \nabla h(\theta_*)\|^2_{\Sigma^{-1}} < \infty$, thus extending the result from Chapter 2.

— The second term is the "variance" term which depends on the noise in the gradients. When the noise is structured (such as for least-squares regression), i.e, there exists $\sigma > 0$ such that $C \preccurlyeq \sigma^2 \Sigma$, the variance term becomes $\frac{\sigma^2 d}{n}$ which is optimal over all estimators in $\mathbb{R}^d$ without regularization [Tsybakov, 2003]. However the regularization g does not bring statistical improvement as possible, for instance, with ℓ_1-regularization. We believe this is due to our proof technique. Indeed, in the case of linear constraints, Duchi and Ruan [2016] recently showed that the primal iterates (θ_n) follow a central limit theorem (CLT), namely $\sqrt{n}\bar{\theta}_n$ is asymptotically normal with a covariance precisely restricted to the active constraints. This supports that SDA may leverage the regularization (the active constraints in their case) to get better statistical performance. We leave such non-asymptotic results to future work.

Assumption **(A6)** on the gradient noise is quite general, since the noise (ξ_n) is allowed to be a martingale difference sequence (correct conditional expectation given the past, but not necessarily independence from the past). However it is not verified by the oracle corresponding to regular SDA for least-squares regression, where the noise combines both an additive and a multiplicative part, and its covariance is then no longer bounded in general (it will be for g the indicator function of a bounded set).

4.3.2 Least-Squares Regression

We consider now the least-squares regression framework, i.e, risk minimization with the square loss. Following Bach and Moulines [2013], we assume that:

(A7) The observations $(x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$, $n \geq 1$, are i.i.d. distributed with finite variances $\mathbb{E}\|x_n\|_2^2 < \infty$ and $\mathbb{E}y_n^2 < \infty$.

(A8) We consider the *least-squares regression* problem which is the minimization of the quadratic function $f(\theta) = \frac{1}{2}\mathbb{E}(\langle x_n, \theta\rangle - y_n)^2$.

(A9) We denote by $\Sigma = \mathbb{E}[x_n \otimes x_n]$ the population covariance matrix, which is the Hessian of f at all points. Without loss of generality, we reduce $\mathbb{R}^d$ to the minimal subspace where all x_n, $n \geq 1$, lie almost surely. Therefore Σ is invertible and all the eigenvalues of Σ are strictly positive, even if they may be arbitrarily small.

(A10) We denote the residual by $\xi_n = (y_n - \langle \theta_*, x_n\rangle)x_n$. We have $\mathbb{E}[\xi_n] = 0$ but $\mathbb{E}[\xi_n | x_n] \neq 0$ in general (unless the model is well-specified). There exists $\sigma > 0$ such that $\mathbb{E}[\xi_n \otimes \xi_n] \preccurlyeq \sigma^2 \Sigma$.

(A11) There exists $\kappa > 0$ such that for all $z \in \mathbb{R}^d$, $\mathbb{E}\langle z, x_n\rangle^4 \leq \kappa \langle z, \Sigma z\rangle^2$.

(A12) The function g is lower bounded by some constant which is assumed by sake of simplicity to be 0.

(A13) There exists $L > 0$ such that $Lh - \frac{1}{2}\|\cdot\|_\Sigma^2$ is convex.

Assumptions **(A7-9)** are standard for least-squares regression, while Assumption **(A10)** defines a bounded statistical noise. Assumption **(A11)** is commonly used in the analysis of least-mean-square algorithms [Macchi, 1995] and says the projection of the covariates x_n on any direction $z \in \mathbb{R}^d$ have a bounded *kurtosis*. It is true for Gaussian vectors with $\kappa = 3$. Assumption **(A13)** links up the geometry of the function h and the objective function f; for example for ℓ_p-geometries, L is proportional to $\mathbb{E}\|x\|_q^2$ where $1/p + 1/q = 1$ (see Corollary 7 in Appendix 4.I).

For the least-squares regression problem, the SDA algorithm defined in Eq. (4.6) takes the form:

$$\begin{aligned} \eta_n &= \eta_{n-1} - \gamma\big(\langle x_n, \theta_{n-1}\rangle - y_n\big)x_n \\ \theta_n &= \nabla h_n^*(\eta_n). \end{aligned} \tag{4.8}$$

This corresponds to a stochastic oracle of the form $\nabla f_n(\theta) = (\Sigma + \zeta_n)(\theta - \theta_\Sigma) - \xi_n$ for $\theta \in \mathbb{R}^d$, with $\zeta_n = x_n \otimes x_n - \Sigma$. This oracle combines an additive noise ξ_n satisfying the previous Assumption **(A6)** and a multiplicative noise ζ_n which is harder to analyze.

We obtain a similar result compared to Proposition 9 at the cost of additional corrective terms.

Proposition 10. *Assume **(A2-4)** and **(A7-13)**. Consider the recursion in Eq. (4.8) for any constant step-size γ such that $\gamma \leq \frac{1}{4\kappa L d}$. Then*

$$\frac{1}{2}\mathbb{E}\|\bar{\theta}_n - \theta_*\|_\Sigma^2 \leq 2\frac{D_h(\theta_*, \theta_0)}{\gamma n} + \frac{32d}{n}\big(\sigma^2 + \kappa\|\theta_* - \theta_\Sigma\|_\Sigma^2\big) + \frac{16\kappa d}{n^2}\left(\frac{5D_h(\theta_*, \theta_0)}{\gamma} + g(\theta_0)\right).$$

We can make the following remarks:

— The proof technique is similar to the one of Proposition 9. Nevertheless its complexity comes from the extra multiplicative noise ζ_n in the gradient estimate (see Appendix 4.D).

— The result is only proven for $\gamma \leq 1/(4\kappa L d)$ which seems to be a proof artifact. Indeed we empirically observed (see Section 4.5) that the iterates still converge to the solution for all $\gamma \leqslant 1/(2\mathbb{E}\|x_n\|_2^2)$.

— The global bound leads to a rate of $O(1/n)$ without strong convexity, which is optimal for stochastic approximation, even with strong convexity [Nemirovsky and Yudin, 1983]. We recover the terms of Proposition 9 pertubed by: (a) one corrective term of order $O(d/n)$ which depends on the distance between the solution θ_* and the global minimizer θ_Σ of the quadratic function f, which corresponds to the covariance of the multiplicative noise at the optimum, and (b) two residual terms of order $O(d/n^2)$. It would be interesting to study whether these two terms can be removed.

— As in Proposition 9, the bias is also $O\big(\frac{1}{(\gamma n)^2}\|\nabla h(\theta_0) - \nabla h(\theta_*)\|_{\Sigma^{-1}}^2\big)$ for specific starting points (see proof in Appendix 4.D for details).

— It is worth noting that in the constrained case ($g = \mathbb{1}_\mathcal{C}$ for a bounded convex set $\mathcal{C}$), the covariance of the noisy oracle is simply bounded by $(\kappa \operatorname{tr}\Sigma r^2 + \sigma^2)\Sigma$ where

we denote by $r = \max_{\theta \in \mathcal{C}} \|\theta - \theta_\Sigma\|_2$ (see Appendix 4.D.1 for details). Therefore Proposition 9 already implies $\frac{1}{2}\mathbb{E}\|\bar{\theta}_n - \theta_*\|_\Sigma^2 \leq 2\frac{D_h(\theta_*,\theta_0)}{\gamma n} + \frac{8d}{n}(\sigma^2 + \kappa r^2 \operatorname{tr}\Sigma)$. Moreover the result holds then for any step-size $\gamma \leqslant 1/L$, which is bigger than allowed for $g = 0$ [Bach and Moulines, 2013].

4.3.3 Convergence Results on the Objective Function

In this section we present the convergence properties of the SDA method on the objective function $\psi = f + g$ rather than on the norm $\|\cdot\|_\Sigma$.

We first start with a disclaimer: it is not possible to obtain general non-asymptotic results on the convergence of the SDA iterates in term of function values without additional assumptions on the regularization g. We indeed show in Appendix 4.E that, even in the simple case of a linear function $f(\theta) = \langle a, \theta \rangle$, for $a \in \mathbb{R}^d$, we can always find, for any finite time horizon N, a quadratic non-strongly convex regularization function g_N such that for any unstructured noise of variance σ^2, the function value $\psi_N(\theta) = f(\theta) + g_N(\theta)$ evaluated in the SDA iterates at time N is lowerbounded by

$$\psi_N(\bar{\theta}_N) - \psi_N(\theta_*) \geq \frac{\sigma^2}{12}.$$

This lower bound is specific to the SDA algorithm and we underline that the regularization g_N depends on the horizon N. However this result still prevents the possibility of a universal non-asymptotic convergence result on the function value for the SDA iterates for general quadratic and linear functions. We note that this does not apply to the setting of Proposition 9 and Proposition 10 since $\Sigma = 0$ for a linear function and the vector q defining the linear term $\langle q, \theta \rangle$ cannot be in the column space of Σ, thus violating Assumption **(A5)**. We conjecture that in the setting of Assumption **(A5)**, the lower bound is $O(1/\sqrt{n})$ as well.

We now provide some specific examples for which we can prove convergence in function values.

Quadratic objectives with smooth regularization. When there exists a constant $L_g \geq 0$ such that $L_g f - g$ is convex on $\mathring{\mathcal{X}}$ then results from Propositions 9 and 10 directly imply convergence of the composite objective to the optimum through

$$\psi(\bar{\theta}_n) - \psi(\theta_*) \leq \frac{(L_g + 1)}{2}\|\bar{\theta}_n - \theta_*\|_\Sigma^2 = O(1/n),$$

with precise constants from Propositions 9 and 10. Indeed we have in that case $(L_g + 1)f - \psi$ convex and this would be directly implied by Proposition 11 in Appendix 4.B.

An easy but still interesting application is the non-regularized case ($g = 0$) when the optimum θ_* is the global optimum θ_Σ of f, because then $\psi(\theta) - \psi(\theta_*) = \frac{1}{2}\|\theta - \theta_*\|_\Sigma^2$. Thus this extends previous results on function values [Bach and Moulines, 2013] to non-Euclidean geometries.

Constrained problems. When g is the indicator function of a convex set $\mathcal{C}$ then by definition the primal iterate $\theta_n \in \mathcal{C}$ and by convexity $\bar{\theta}_n \in \mathcal{C}$. Therefore $\psi(\bar{\theta}_n) = f(\bar{\theta}_n) + \mathbb{1}_{\mathcal{C}}(\bar{\theta}_n) = f(\bar{\theta}_n)$ and we obtain with the Cauchy-Schwarz inequality:

$$\begin{aligned} f(\bar{\theta}_n) - f(\theta_*) &= \langle \nabla f(\theta_*), \bar{\theta}_n - \theta_* \rangle + \frac{1}{2}\|\bar{\theta}_n - \theta_*\|_\Sigma^2 \\ &\leq \|\theta_* - \theta_\Sigma\|_2 \|\bar{\theta}_n - \theta_*\|_\Sigma + \frac{1}{2}\|\bar{\theta}_n - \theta_*\|_\Sigma^2 = O\Big(\frac{\|\theta_* - \theta_\Sigma\|_2}{\sqrt{n}}\Big), \end{aligned}$$

with precise constants from Propositions 9 and 10. Hence we obtain a global rate of order $O(1/\sqrt{n})$ for the convergence of the function value in the constrained case.

These rates may be accelerated to $O(1/n)$ for certain specific convex constraints or when the global optimum $\theta_\Sigma \in \mathcal{C}$; Duchi and Ruan [2016] recently obtained asymptotic convergence results for the iterates in the cases of linear and ℓ_2-ball constraints for linear objective functions. Their results can be directly extended to asymptotic convergence of function values and very probably to all strongly convex sets [see, e.g., Vial, 1983]. However, even for the simple ℓ_2-ball constrained problem, we were not able to derive non-asymptotic convergence rates for function values.

However the global rate of order $O(1/\sqrt{n})$ is statically non-improvable in general. In Appendix 4.F, we relate the stochastic convex optimization problem [Agarwal et al., 2012] to the statistical problem of convex aggregation of estimators [Tsybakov, 2003, Lecué, 2006]. These authors showed lower bounds on the performance of such estimators which provide us lower bounds on the performance of any stochastic algorithm to solve constrained problems. In Proposition 14 and Proposition 16 of Appendix 4.F, we derive more precisely lower bound results for linear and quadratic functions for certain ranges of n and d confirming the optimality of the convergence rate $O(1/\sqrt{n})$. This being said, in our experiments in Section 4.5, we observed that the convergence of function values follows closely the convergence in the Mahalanobis distance.

4.4 Parallel Between Dual Averaging and Mirror Descent

In this section we compare the behaviors of DA and MD algorithms, by highlighting their similarities and differences, in particular in terms of continuous-time interpretation.

4.4.1 Lazy versus Greedy Projection Methods

DA and MD are often described in the online-learning literature as "lazy" and "greedy" projection methods [Zinkevich, 2003]. Indeed, the difference between these two methods is more apparent in the Euclidean projection case (when $g = \mathbb{1}_{\mathcal{C}}$ and $h = \frac{1}{2}\|\cdot\|_2^2$). MD is then projected gradient descent and may be written under its

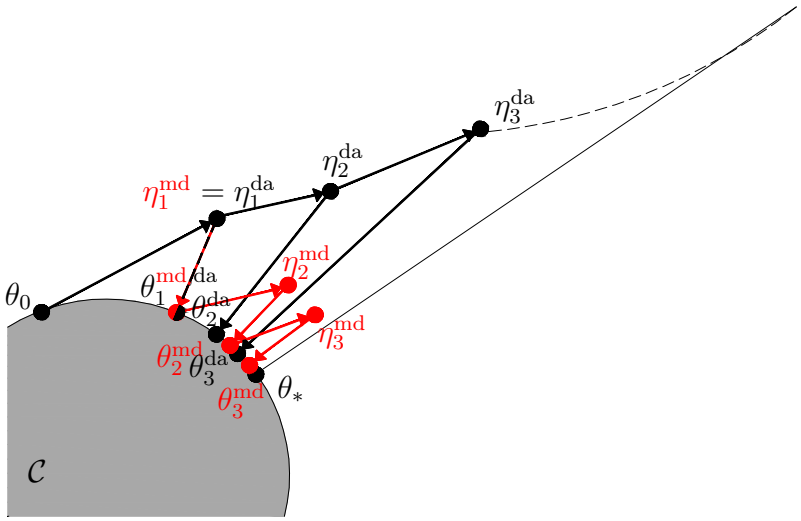


Figure 4-1 – Lazy versus greedy projection. In red: greedy projection. In black: lazy projection.

primal-dual form as:

$$\eta_n^{\mathrm{md}} = \theta_{n-1}^{\mathrm{md}} - g_n^{\mathrm{md}} \quad \text{with} \quad g_n^{\mathrm{md}} \in \partial f(\theta_{n-1}^{\mathrm{md}}) \quad \text{and} \quad \theta_n^{\mathrm{md}} = \operatorname*{arg\,min}_{\theta \in \mathcal{C}} \|\eta_n^{\mathrm{md}} - \theta\|_2.$$

Whereas DA takes the form

$$\eta_n^{\mathrm{da}} = \eta_{n-1}^{\mathrm{da}} - g_n^{\mathrm{da}} \quad \text{with} \quad g_n^{\mathrm{da}} \in \partial f(\theta_{n-1}^{\mathrm{da}}) \quad \text{and} \quad \theta_n^{\mathrm{da}} = \operatorname*{arg\,min}_{\theta \in \mathcal{C}} \|\eta_n^{\mathrm{da}} - \theta\|_2.$$

Therefore, imagining the subgradients g_n are provided by an adversary without the need to compute the primal sequence (θ_n), no projections are needed to update the dual sequence (η_n^{da}), and this one moves far away in the asymptotic direction of the gradient at the optimum $\nabla f(\theta_*)$. Furthermore the primal iterate θ_n^{da} is simply obtained, when required, by projecting back the dual iterate in the constraint set. Conversely, the MD dual iterate η_n^{md} update calls for $\theta_{n-1}^{\mathrm{md}}$, and therefore a projection step is unavoidable. Thereby MD iterates $(\eta_n^{\mathrm{md}}, \theta_n^{\mathrm{md}})$ are going, at each iteration, back-and-forth between the boundary and the outside of the convex set $\mathcal{C}$. This is illustrated in Figure 4-1

4.4.2 Strongly Convex Cases

MD converges linearly for smooth and strongly convex functions f, in the absence of a regularization component [Lu et al., 2016] or for Euclidean geometries [Nesterov, 2013]. However we were not able to derive faster convergence rates for DA when the function f or the regularization g are strongly convex. Moreover the only results we found in the literature are about (a) an alteration of the dual gradient method [Devolder et al., 2013, Section 4] which is itself a modification of DA with an additional projection step proposed by Nesterov [2013] for smooth optimization, (b) the strongly

convex regularization g which enables Xiao [2010] to obtain a $O(1/\mu n)$ convergence rate in the stochastic case.

At the simplest level, for $h = \frac{1}{2}\|\cdot\|_2^2$ and $f = 0$, MD is equivalent to the proximal point algorithm [Martinet, 1970] $\theta_n^{\mathrm{md}} = \arg\min_{\theta\in\mathbb{R}^d} \left\{g(\theta) + \frac{1}{\gamma}\|\theta - \theta_{n-1}^{\mathrm{md}}\|_2^2\right\}$, whereas DA, which is not anymore iterative, is such that $\theta_n^{\mathrm{da}} = \arg\min_{\theta\in\mathbb{R}^d} \left\{g(\theta) + \frac{1}{\gamma n}\|\theta\|_2^2\right\}$. For the squared ℓ_2-regularization $g(\theta) = \frac{\nu}{2}\|\theta - \theta_*\|_2^2$, we compute exactly (see Appendix 4.G)

$$g(\theta_n^{\mathrm{md}}) - g(\theta_*) = \left(\frac{1}{\gamma\nu}\right)^n [g(\theta_0^{\mathrm{md}}) - g(\theta_*)] \quad \text{and} \quad g(\theta_n^{\mathrm{da}}) - g(\theta_*) = \frac{g(\theta_0^{\mathrm{da}}) - g(\theta_*)}{(1+\nu\gamma n)^2}.$$

Therefore the convergence of DA can be dramatically slower than MD. However when noise is present, its special structure may be leveraged to get interesting results.

4.4.3 Continuous Time Interpretation of DA et MD

Following Nemirovsky and Yudin [1983], Bolte and Teboulle [2003], Krichene et al. [2015], Wibisono et al. [2016] we propose a continuous interpretation of these methods for g twice differentiable. Precise computations are derived in Appendix 4.H.

The MD iteration in Eq. (4.4) may be viewed as a forward-backward Euler discretization of the MD ODE [Bolte and Teboulle, 2003]:

$$\dot{\theta} = -\nabla^2 h(\theta)^{-1}[\nabla f(\theta) + \nabla g(\theta)]. \tag{4.9}$$

On the other hand, the ODE associated to DA takes the form

$$\dot{\theta} = -\nabla^2 (h(\theta) + tg(\theta))^{-1}(\nabla f(\theta) + \nabla g(\theta)). \tag{4.10}$$

It is worth noting that these ODEs are very similar, with an additional term $tg(\theta)$ in the inverse mapping $\nabla^2(h(\theta) + tg(\theta))^{-1}$ which may slow down the DA dynamics.

In analogy with the discrete case, the Bregman divergences D_h and D_{h+tg} are respectively Lyapunov functions for the MD and the DA ODEs [see, e.g., Krichene et al., 2015] and we notice in Appendix 4.H the continuous time argument really mimics the proof of Proposition 8 without the technicalities associated with discrete time. Moreover we recover the variational interpretation of Krichene et al. [2015], Wibisono et al. [2016], Wilson et al. [2016]: the Lyapunov function generates the dynamic in the sense that a function L is first chosen and secondly a dynamics, for which L is a Lyapunov function, is then designed. In this way MD and DA are the two different dynamics associated to the two different Lyapunov functions D_h and D_{h+tg}. We also provide in Appendix 4.H a slight extension to the noisy-gradient case.

4.5 Experiments

In this section, we illustrate our theoretical results on synthetic examples. We provide additional experiments on a machine learning benchmark in Appendix 4.K.

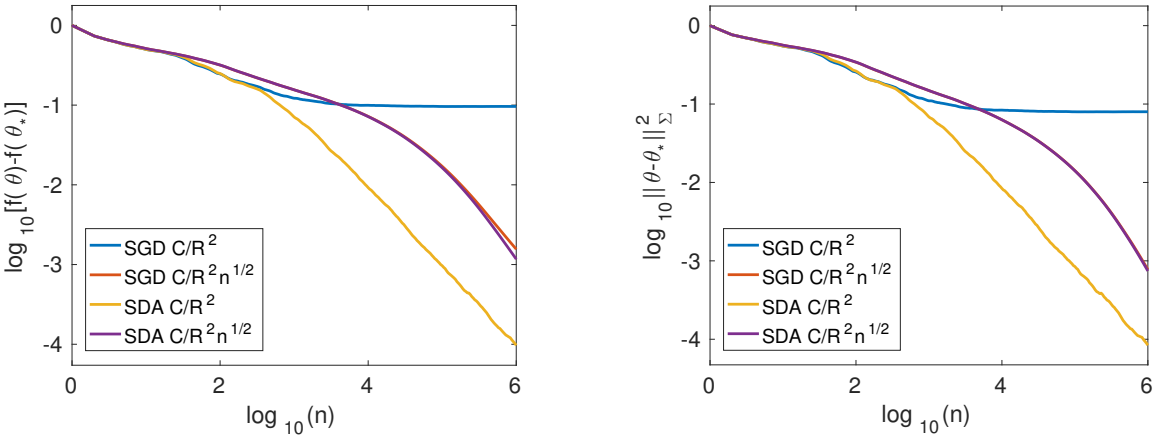


Figure 4-1 – Simplex-constrained least-squares regression with synthetic data. Left: Performance on the objective function. Right: Performance on the Mahalanobis norm $\|\cdot\|_\Sigma^2$.

Simplex-constrained least-squares regression with synthetic data. We consider normally distributed inputs $x_n \in \mathbb{R}^d$ with a covariance matrix Σ that has random eigenvectors and eigenvalues $1/k$, for $k = 1, \dots, d$ and a random global optimum $\theta_\Sigma \in [0, +\infty)^d$. The outputs y_n are generated from a linear function with homoscedatic noise with unit signal to noise-ratio ($\sigma^2 = 1$). We denote by $R^2 = \operatorname{tr}\Sigma$ the average radius of the data and we show results averaged over 10 replications.

We consider the problem of least-squares regression constrained on the simplex Δ_d of radius $r = \|\theta_\Sigma\|_1/2$, i.e., $\min_{\theta \in r\Delta_d} \mathbb{E}(\langle x_n, \theta\rangle - y_n)^2$, for $d = 100$. We compare the performance of SDA and SGD algorithms with different settings of the step-size γ_n, constant or proportional to $1/\sqrt{n}$. In the left plot of Figure 4-1 we show the performance on the objective function and on the right plot, we show the performance on the squared Mahalanobis norm $\|\cdot\|_\Sigma^2$. All costs are shown in log-scale, normalized so that the first iteration leads to $f(\theta_0) - f(\theta_*) = 1$. We can make the following observations (we only show results on Euclidean geometry since results under the negative entropy geometry were very similar):

— With constant step-size, SDA converges to the solution at rate $O(1/n)$ whereas the SGD algorithm does not converge to the optimal solution.

— With decaying step-size $\gamma_n = 1/(2R^2\sqrt{n})$, SDA and SGD converge first at rate $O(1/\sqrt{n})$, then at rate $O(1/n)$, taking finally advantage of the strong convexity of the problem.

— We note (a) there is no empirical difference between the performance on the objective function and the squared distance $\|\cdot\|_\Sigma^2$, (b) with decreasing step-size, SGD and SDA behave very similarly.

4.6 Conclusion

In this chapter, we proposed and analyzed the first algorithm to achieve a convergence rate of $O(1/n)$ for stochastic composite objectives, without the need for strong convexity. This was achieved by considering a constant step-size and averaging of the primal iterates in the dual averaging method.

Our results only apply to expectations of quadratic functions (but to any additional potentially non-smooth terms). In fact, constant step-size stochastic dual averaging is not convergent for general smooth objectives; however, as done in the non-composite case by Bach and Moulines [2013], one could iteratively solved quadratic approximations of the smooth problems with the algorithm we proposed in this chapter to achieve the same rate of $O(1/n)$, still with robustness to ill-conditioning and efficient iterations. Finally, it would be worth considering accelerated extensions to achieve a forgetting of initial conditions in $O(1/n^2)$ as was done for averaged gradient descent in Chapter 3. After having studied the stochastic optimization of quadratic functions, we now shift our focus towards two applications of the quadratic loss in machine learning.

Appendix

4.A Unambiguity of the Primal Iterate

We describe here conditions under which the primal iterate θ_n in Eq. (4.3) is correctly defined. Since h is strictly convex, h_n^* is continuously differentiable on $\overset{\circ}{\text{dom}}\, h_n^*$ [see Hiriart-Urruty and Lemaréchal, 2001, Theorem 4.1.1]. Therefore the primal iterate θ_n is well defined if the dual iterate $\eta_n \in \overset{\circ}{\text{dom}}\, h_n^*$. It is, for example, the case under two natural assumptions as shown by the next lemma which is an adaption of Lemma 2 by Bauschke et al. [2016].

Lemma 21. *We make the following assumptions:*
(B1) *h or g is supercoercive.*
(B2) $\arg\min_{\theta\in\mathcal{X}} \psi(\theta)$ *is compact and h bounded below.*
*Under **(B1)** or **(B2)** the primal iterates (θ_n) defined in Eq. (4.3) are well defined.*

Proof. Since h is strictly convex, h_n^* is continuously differentiable on $\overset{\circ}{\text{dom}}\, h_n^*$ [see Hiriart-Urruty and Lemaréchal, 2001, Theorem 4.1.1]. Therefore the primal iterate θ_n is well defined if the dual iterate $\eta_n \in \overset{\circ}{\text{dom}}\, h_n^*$.

— If h or g is supercoercive then h_n is supercoercive [see Bauschke and Combettes, 2011, Proposition 11.13] and it follows from Hiriart-Urruty and Lemaréchal [2001, Chapter E, Proposition 1.3.8] that $\text{dom}\, h_n^* = \mathbb{R}^d$.

— If $\arg\min_{\theta\in\mathcal{X}}\{\psi(\theta)\}$ is compact then $\psi + \mathbb{1}_{\mathcal{X}}$ is coercive. Moreover

$$\begin{aligned} h_n^*(\eta_n) &= \sup_{\theta\in\mathcal{X}} \big\{\langle \eta_n, \theta\rangle - h_n(\theta)\big\} \text{ since } \mathcal{X} \subset \overline{\text{dom}}\, h \\ &= -\inf_{\theta\in\mathcal{X}} \Big\{h(\theta) + \gamma \sum_{i=1}^{n} (g(\theta) + f(\theta_{i-1}) + \langle \nabla f(\theta_{i-1}), \theta - \theta_{i-1}\rangle)\Big\} \\ &\quad +\gamma \sum_{i=1}^{n} (f(\theta_{i-1}) - \langle \nabla f(\theta_{i-1}), \theta_{i-1}\rangle) \\ &\leq -\inf_{\theta\in\mathcal{X}} \big\{h(\theta) + n\gamma(g(\theta) + f(\theta))\big\} + \gamma \sum_{i=1}^{n} (f(\theta_{i-1}) - \langle \nabla f(\theta_{i-1}), \theta_{i-1}\rangle), \end{aligned}$$

by convexity of f. Therefore $\eta_n \in \overset{\circ}{\text{dom}}\, h_n^*$ since $\psi + \mathbb{1}_{\mathcal{X}}$ is coercive and h bounded below [see Bauschke and Combettes, 2011, Proposition 11.15].

□

4.B Proof of Convergence of Deterministic DA

We first describe a new notion of smoothness defined by Bauschke et al. [2016]. Then we present our extension of the Bregman divergence to the non-smooth function g to finally prove Proposition 8.

4.B.1 A *Lipschitz-Like/Convexity Condition*

Classical results on the convergence of optimization algorithms in non-Euclidean geometry assume on one hand that the function h is strongly convex and on the other hand the function f is Lipschitz or smooth. Following Bauschke et al. [2016], we consider a different assumption which combines the smoothness of f and the strong convexity of h on a single condition called *Lipschitz-like/Convexity Condition* by Bauschke et al. [2016] and denoted by **(LC)**:

(LC) There exists a constant $L \in (0, +\infty)$ such that $Lh - f$ is convex on $\mathring{\mathcal{X}}$.

For Euclidean geometry, this condition is obviously equivalent to the smoothness of the function f with regards to the ℓ_2-norm. Moreover, under an arbitrary norm $\|\cdot\|$, assuming h μ-strongly convex and f L-smooth clearly implies, by simple convex computation, **(LC)** with constant L/μ. However **(LC)** is much more general and may hold even when f is non-smooth what precisely justifies the introduction of this condition. Many examples are described by Bauschke et al. [2016]. Furthermore this notion has the elegance of pairing well with Bregman divergences and leading to more refined proofs as shown in the following proposition which summarizes equivalent properties of **(LC)**.

Proposition 11 (Bauschke et al. [2016])**.** *Assume **(A1-4)**. For $L > 0$ the following conditions are equivalent:*

- *$Lh - f$ is convex on $\mathring{\mathcal{X}}$, i.e., **(LC)** holds,*
- *$D_f(\alpha, \beta) \leq L D_h(\alpha, \beta)$ for all $(\alpha, \beta) \in \mathcal{X} \times \mathring{\mathcal{X}}$.*

Furthermore, when f and h are assumed twice differentiable, then the above is equivalent to

$$\nabla^2 f(\theta) \preccurlyeq L \nabla^2 h(\theta) \quad \text{for all } \theta \in \mathring{\mathcal{X}}.$$

4.B.2 Generalized Bregman Divergence

The Bregman divergence was defined by Bregman [1967] for a differentiable convex function h as

$$D_h(\alpha, \beta) = h(\alpha) - h(\beta) - \langle \nabla h(\beta), \alpha - \beta \rangle, \text{ for } (\alpha, \beta) \in \operatorname{dom} h \times \mathring{\operatorname{dom}}\, h. \tag{4.11}$$

It behaves as a squared distance depending on the function h and extends the computational properties of the squared ℓ_2-norm to non-Euclidean spaces. Indeed most proofs in Euclidean space rest upon the expansion $\|\theta_n - \theta_* - \gamma \nabla f(\theta_n)\|_2^2 = \|\theta_n - \theta_*\|_2^2 + \gamma^2 \|\nabla f(\theta_n)\|_2^2 - 2\gamma \langle \nabla f(\theta_n), \theta_n - \theta_* \rangle$ which is not available in non-Euclidean geometry. Therefore the Bregman divergence comes to rescue and is used to compute

a deviation between the current iterate of the algorithm and the solution of the problem and, seemingly, used as an non-Euclidean Lyapunov function [Chen and Teboule, 1993]. It has been widely used in optimization [see, e.g., Bauschke and Borwein, 1997, for a review].

We follow this path and include the regularization component g of the objective function $\psi = f + g$ in the Bregman divergence for the sake of the analysis. If g was differentiable we would simply use $D_{h+n\gamma g}$. Since g is not differentiable, D_{h_n} is not well defined. However for $(\alpha, \eta) \in \operatorname{dom} h \times \mathring{\operatorname{dom}}\, h_n^*$, we denote by extension for $\theta = \nabla h_n^*(\eta)$:

$$\tilde{D}_n(\alpha, \eta) = h_n(\alpha) - h_n(\theta) - \langle \eta, \alpha - \theta \rangle. \tag{4.12}$$

This extension is different from the one defined by Kiwiel [1997]. It is worth noting that if there exists μ such that $\alpha = \nabla h_n^*(\mu)$, we recover the classical formula $\tilde{D}_n(\alpha, \eta) = D_{h_n^*}(\eta, \mu)$ which is well defined since h_n^* is differentiable. Yet $\tilde{D}_n$ is defined more generally since such a μ does not always exist. The next lemma relates $\tilde{D}_n$ to D_h and is obvious if g is differentiable since $D_{h_n} = D_h + \gamma n D_g$.

Lemma 22. *Let $n \geq 0$, $\alpha \in \operatorname{dom} h$ and $\eta \in \mathring{\operatorname{dom}}\, h_n^*$, then with $\theta = \nabla h_n^*(\eta)$,*

$$\tilde{D}_n(\alpha, \eta) \geq D_h(\alpha, \theta). \tag{4.13}$$

Proof. $\theta = \nabla h_n^*(\eta)$, thus $\eta \in \partial h_n(\theta)$ and by elementary calculus rule $\partial h_n(\theta) = \nabla h(\theta) + n\gamma \partial g(\theta)$. Consequently $\eta - \nabla h(\theta) \in n\gamma \partial g(\theta)$ and by convexity of g

$$\tilde{D}_n(\alpha, \eta) - D_h(\alpha, \theta) = n\gamma \left[g(\alpha) - g(\theta) - \left\langle \frac{\eta - \nabla h(\theta)}{\gamma n}, \alpha - \theta \right\rangle \right] \geq 0.$$

□

4.B.3 Proof of Proposition 8

We assume their exists a constant $L > 0$ such that $Lh - f$ is convex on $\mathring{\mathcal{X}}$ and we assume the step-size $\gamma \leq 1/L$. We first show that the Bregman divergence decreases along the iterates [see, e.g., Chen and Teboule, 1993, Beck and Teboulle, 2003, Bach, 2015]. For all $\theta \in \mathcal{X}$,

$$\begin{aligned}
\tilde{D}_n(\theta, \eta_n) - \tilde{D}_{n-1}(\theta, \eta_{n-1}) &= h_{n-1}(\theta_{n-1}) - h_n(\theta_n) + h_n(\theta) - h_{n-1}(\theta) \\
&\quad - \langle \eta_n, \theta - \theta_n \rangle + \langle \eta_{n-1}, \theta - \theta_{n-1} \rangle \\
&= h_{n-1}(\theta_{n-1}) - h_{n-1}(\theta_n) - \gamma(g(\theta_n) - g(\theta)) \\
&\quad + \langle \eta_{n-1}, \theta_n - \theta_{n-1} \rangle + \langle \eta_n - \eta_{n-1}, \theta_n - \theta \rangle \\
&= -\tilde{D}_{n-1}(\theta_n, \eta_{n-1}) - \gamma(g(\theta_n) - g(\theta)) \\
&\quad - \gamma \langle \nabla f(\theta_{n-1}), \theta_n - \theta \rangle.
\end{aligned}$$

Therefore for all $\theta \in \mathcal{X}$,

$$\begin{aligned}
\tilde{D}_n(\theta, \eta_n) - \tilde{D}_{n-1}(\theta, \eta_{n-1}) = &-\tilde{D}_{n-1}(\theta_n, \eta_{n-1}) + \gamma \langle \nabla f(\theta_{n-1}), \theta_{n-1} - \theta_n \rangle \\
&- \gamma(g(\theta_n) - g(\theta)) - \gamma \langle \nabla f(\theta_{n-1}), \theta_{n-1} - \theta \rangle.
\end{aligned} \tag{4.14}$$

It follows from Proposition 11 and Lemma 22

$$f(\theta_n) - f(\theta_{n-1}) + \langle \nabla f(\theta_{n-1}), \theta_n - \theta_{n-1} \rangle \leq L D_h(\theta_n, \theta_{n-1}) \leq L D_{n-1}(\theta_n, \theta_{n-1}),$$

and from the convexity of f,

$$-\langle \nabla f(\theta_{n-1}), \theta_{n-1} - \theta \rangle \leq f(\theta) - f(\theta_{n-1}).$$

And Eq. (4.14) is bounded by

$$\tilde{D}_n(\theta, \eta_n) - \tilde{D}_{n-1}(\theta, \eta_{n-1}) \leq \gamma(\psi(\theta) - \psi(\theta_n)) + (\gamma L - 1) D_h(\theta_n, \theta_{n-1}).$$

Thus for $\gamma \leq 1/L$,

$$\tilde{D}_n(\theta, \eta_n) - \tilde{D}_{n-1}(\theta, \eta_{n-1}) \leq \gamma(\psi(\theta) - \psi(\theta_n)).$$

Taking $\theta = \theta_{n-1}$ we note that the sequence $\{\psi(\theta_n)\}_{n \geq 0}$ is decreasing and we obtain for $\gamma \leq 1/L$,

$$\psi(\theta_n) - \psi(\theta) \leq \frac{1}{n+1} \sum_{k=0}^{n} [\psi(\theta_i) - \psi(\theta)] \leq \frac{D_h(\theta, \theta_0) - \tilde{D}_n(\theta, \eta_n)}{\gamma(n+1)}. \tag{4.15}$$

We assume now that the non-smooth part $g = 0$ and there exists $\mu \geq 0$ such that $f - \mu h$ is convex. So Proposition 11 implies

$$-\langle \nabla f(\theta_{n-1}), \theta_{n-1} - \theta \rangle \leq f(\theta) - f(\theta_{n-1}) - \mu D_h(\theta, \theta_{n-1}),$$

which gives with Eq. (4.14) the better bound

$$D_h(\theta, \theta_n) - D_h(\theta, \theta_{n-1}) \leq \gamma(f(\theta) - f(\theta_n)) - \gamma\mu D_h(\theta, \theta_{n-1}) + (\gamma L - 1) D_h(\theta_n, \theta_{n-1}).$$

And for $\gamma \leq 1/L$, this can be simplified as

$$D_h(\theta, \theta_n) \leq (1 - \gamma\mu) D_h(\theta, \theta_{n-1}) + \gamma(f(\theta) - f(\theta_n)).$$

The sequence $\{f(\theta_n)\}_{n \geq 0}$ is still decreasing and we obtain by expanding the recursion

$$\begin{aligned}
D_h(\theta, \theta_n) &\leq (1 - \gamma\mu)^n D_h(\theta, \theta_0) + \sum_{k=1}^{n} (1 - \gamma\mu)^{n-k} \gamma(f(\theta) - f(\theta_k)) \\
&\leq (1 - \gamma\mu)^n D_h(\theta, \theta_0) + \sum_{k=1}^{n} (1 - \gamma\mu)^{n-k} \gamma(f(\theta) - f(\theta_k)) \\
&\leq (1 - \gamma\mu)^n D_h(\theta, \theta_0) + \sum_{k=1}^{n} (1 - \gamma\mu)^{n-k} \gamma(f(\theta) - f(\theta_n)) \\
&\leq (1 - \gamma\mu)^n D_h(\theta, \theta_0) + \gamma \frac{1 - (1 - \gamma\mu)^n}{\gamma\mu} (f(\theta) - f(\theta_n)).
\end{aligned}$$

Thus for all $\theta \in \mathcal{X}$,

$$\frac{1-(1-\gamma\mu)^n}{\mu}(f(\theta_n)-f(\theta))+D_h(\theta,\theta_n)\leq(1-\gamma\mu)^n D_h(\theta,\theta_0),$$

and

$$f(\theta_n)-f(\theta)\leq\frac{\gamma\mu(1-\gamma\mu)^n}{1-(1-\gamma\mu)^n}\frac{D_h(\theta,\theta_0)}{\gamma}\leq(1-\gamma\mu)^n\frac{D_h(\theta,\theta_0)}{\gamma},$$

since $(1-\gamma\mu)^2\leq 1-\gamma\mu$ implies $\gamma\mu/(1-(1-\gamma\mu)^n)\leq 1$.

4.C Proof of Proposition 9

In this section, we will prove Proposition 9. The proof relies on considering the difference between the iteration with noise we denote by (η_n,θ_n) and without noise we denote by (ω_n,ϕ_n), which happens to verify a similar recursion as the SDA recursion.

— We first show in Lemma 23 that the distance $\mathbb{E}\|\eta_n-\omega_n\|^2_{\Sigma^{-1}}$ is of order n.

— Then in Lemma 24 we show that $\mathbb{E}\|\bar{\theta}_n-\bar{\phi}_n\|^2_\Sigma$ is of order $O(1/n)$, by: (a) noticing that $\mathbb{E}\|\bar{\theta}_n-\bar{\phi}_n\|^2_\Sigma$ is of order $\frac{\mathbb{E}\|\eta_n-\omega_n\|^2_{\Sigma^{-1}}}{n^2}+\frac{\text{variance}}{n}$, (b) combining this with the result of Lemma 23.

4.C.1 Two Technical Lemmas

We first present and prove two technical lemmas.

Bound on the Difference of Two Dual Iterates

In the following lemma we show that the difference between two dual iterates that follow the same recursion is of order n. This will be used with the iteration with noise (η_n,θ_n) and without noise (ω_n,ϕ_n).

Lemma 23. *Let us consider two sequences of iterates (μ_k,α_k) and (ν_k,β_k) which satisfy the recursion $\mu_n-\nu_n=\mu_{n-1}-\nu_{n-1}-\gamma\Sigma(\alpha_{n-1}-\beta_{n-1})+\gamma\xi_n$, $\alpha_n=\nabla h_n^*(\mu_n)$ and $\beta_n=\nabla h_n^*(\nu_n)$ and assume that γ is such that $2h-\gamma f$ is convex then for all $n\geq 0$*

$$\mathbb{E}\|\mu_n-\nu_n\|^2_{\Sigma^{-1}}\leq\|\mu_0-\nu_0\|^2_{\Sigma^{-1}}+n\gamma^2\operatorname{tr}\Sigma^{-1}C.$$

Proof. We first expand the square.

$$\begin{aligned}\|\mu_{n+1}-\nu_{n+1}\|^2_{\Sigma^{-1}} &= \|\mu_n-\nu_n\|^2_{\Sigma^{-1}}+\gamma^2\|[\Sigma(\alpha_n-\beta_n)-\xi_{n+1}]\|^2_{\Sigma^{-1}}\\ &\quad -2\gamma\langle\Sigma(\alpha_n-\beta_n)-\xi_{n+1},\Sigma^{-1}(\mu_n-\nu_n)\rangle\\ &= \|\mu_n-\nu_n\|^2_{\Sigma^{-1}}+\gamma^2\|\alpha_n-\beta_n\|^2_\Sigma+\gamma^2\|\xi_{n+1}\|^2_{\Sigma^{-1}}\\ &\quad -2\gamma^2\langle\alpha_n-\beta_n,\xi_{n+1}\rangle-2\gamma\langle\alpha_n-\beta_n-\Sigma^{-1}\xi_{n+1},\mu_n-\nu_n\rangle.\end{aligned}$$

And taking the expectation

$$\begin{aligned}
\mathbb{E}[\|\mu_{n+1}-\nu_{n+1}\|^2_{\Sigma^{-1}}|\mathcal{F}_n] &= \|\mu_n-\nu_n\|^2_{\Sigma^{-1}}+\gamma^2\mathbb{E}[\|\xi_{n+1}\|^2_{\Sigma^{-1}}|\mathcal{F}_n] \\
&\quad +\gamma^2\|\alpha_n-\beta_n\|^2_\Sigma-2\gamma^2\mathbb{E}[\langle\alpha_n-\beta_n,\xi_{n+1}\rangle|\mathcal{F}_n] \\
&\quad -2\gamma\mathbb{E}[\langle\alpha_n-\beta_n)-\Sigma^{-1}\xi_{n+1},\mu_n-\nu_n\rangle|\mathcal{F}_n] \\
&= \|\mu_n-\nu_n\|^2_{\Sigma^{-1}}+\gamma^2\operatorname{tr}\Sigma^{-1}\mathbb{E}[\xi_{n+1}\otimes\xi_{n+1}|\mathcal{F}_n] \\
&\quad +\gamma^2\|\alpha_n-\beta_n\|^2_\Sigma-2\gamma^2\langle\alpha_n-\beta_n,\mathbb{E}[\xi_{n+1}|\mathcal{F}_n]\rangle \\
&\quad -2\gamma\langle\alpha_n-\beta_n)-\Sigma^{-1}\mathbb{E}[\xi_{n+1}|\mathcal{F}_n],\mu_n-\nu_n\rangle \\
&= \|\mu_n-\nu_n\|^2_{\Sigma^{-1}}+\gamma^2\operatorname{tr}\Sigma^{-1}C \\
&\quad +\gamma^2\|\alpha_n-\beta_n\|^2_\Sigma-2\gamma\langle\alpha_n-\beta_n,\mu_n-\nu_n\rangle.
\end{aligned}$$

Moreover, using the definition of α_n and β_n, with $\square=\gamma\|\alpha_n-\beta_n\|^2_\Sigma-2\langle\alpha_n-\beta_n,\mu_n-\nu_n\rangle$,

$$\begin{aligned}
\square &= \langle\gamma\Sigma(\alpha_n-\beta_n)-2(\mu_n-\nu_n),\alpha_n-\beta_n\rangle \\
&= \langle\gamma\nabla f(\alpha_n)-\nabla f(\beta_n)-2(\nabla h(\alpha_n)-\nabla h(\beta_n)),\alpha_n-\beta_n\rangle \\
&\quad -2\langle(\mu_n-\nabla h(\alpha_n))-(\nu_n-\nabla h(\beta_n)),\alpha_n-\beta_n\rangle \\
&= \langle\nabla(\gamma f-2h)(\alpha_n)-\nabla(\gamma f-2h)(\beta_n),\alpha_n-\beta_n\rangle \\
&\quad -2\langle(\mu_n-\nabla h(\alpha_n))-(\nu_n-\nabla h(\beta_n)),\alpha_n-\beta_n\rangle.
\end{aligned}$$

Using the h-smoothness of f and assuming that γ is such $2h-\gamma f$ is convex,

$$\langle\nabla(\gamma f-2h)(\alpha_n)-\nabla(\gamma f-2h)(\beta_n),\alpha_n-\beta_n\rangle\leq 0,$$

and as explained in the proof of Lemma 22, $\mu_n-\nabla h(\alpha_n)\in\partial n\gamma g(\alpha_n)$ and $\nu_n-\nabla h(\beta_n)\in\partial n\gamma g(\beta_n)$ and consequently

$$\langle(\mu_n-\nabla h(\alpha_n))-(\nu_n-\nabla h(\beta_n)),\alpha_n-\beta_n\rangle\geq 0,$$

by convexity of g. This explains that

$$\gamma\|\alpha_n-\beta_n\|^2_\Sigma-2\langle\alpha_n-\beta_n,\mu_n-\nu_n\rangle\leq 0.$$

Then, taking the global expectation, we have shown that

$$\mathbb{E}\|\mu_{n+1}-\nu_{n+1}\|^2_{\Sigma^{-1}}\leq\mathbb{E}\|\mu_n-\nu_n\|^2_{\Sigma^{-1}}+\gamma^2\operatorname{tr}\Sigma^{-1}C,$$

which concludes the proof. □

Bound on the Difference of the Average of Two Primal Iterates

In the following lemma we adapt the classic proof of averaged SGD by Polyak and Juditsky [1992] to show that the difference between two averaged primal iterates, which follow the same recursion, is of order $O(1/n)$.

Lemma 24. *Let us consider two sequences of iterates (μ_k,α_k) and (ν_k,β_k) which*

satisfy the recursion $\mu_n - \nu_n = \mu_{n-1} - \nu_{n-1} - \gamma\Sigma(\alpha_{n-1} - \beta_{n-1}) + \gamma\xi_n$, $\alpha_n = \nabla h_n^*(\mu_n)$ *and* $\beta_n = \nabla h_n^*(\nu_n)$ *and assume that* γ *is such that* $2h - \gamma f$ *is convex then for all* $n \geq 0$

$$\mathbb{E}\|\bar{\alpha}_n - \bar{\beta}_n\|_\Sigma^2 \leq 4\frac{\|\mu_0 - \nu_0\|_{\Sigma^{-1}}^2}{(\gamma n)^2} + \frac{4}{n}\operatorname{tr}\Sigma^{-1}C.$$

Proof. Let us consider two sequences of iterates (μ_k, α_k) and (ν_k, β_k) which satisfy the recursion $\mu_n - \nu_n = \mu_{n-1} - \nu_{n-1} - \gamma\Sigma(\alpha_{n-1} - \beta_{n-1}) + \gamma\xi_n$, $\alpha_n = \nabla h_n^*(\mu_n)$ and $\beta_n = \nabla h_n^*(\nu_n)$. This can be written as

$$\Sigma(\alpha_n - \beta_n) = \frac{\mu_n - \nu_n - \mu_{n+1} + \nu_{n+1}}{\gamma} + \xi_{n+1}.$$

Thus we obtain

$$\Sigma^{1/2}\sum_{i=0}^{n-1}(\alpha_i - \beta_i) \quad = \quad \frac{\Sigma^{-1/2}(\mu_0 - \nu_0 - \mu_n + \nu_n)}{\gamma} + \sum_{i=0}^{n-1}\Sigma^{-1/2}\xi_{i+1}.$$

Finally, using that by convexity $(a+b)^2 \leq 2(a^2+b^2)$, this leads to

$$n^2\mathbb{E}\|\bar{\alpha}_n - \bar{\beta}_n\|_\Sigma^2 \leq 2\mathbb{E}\left\|\frac{\Sigma^{-1/2}(\mu_0 - \nu_0)}{\gamma} + \sum_{i=0}^{n-1}\Sigma^{-1/2}\xi_{i+1}\right\|_2^2 + 2\mathbb{E}\left\|\frac{\Sigma^{-1/2}(\mu_n - \nu_n)}{\gamma}\right\|_2^2.$$

Using martingale second moment expansions, we obtain

$$\mathbb{E}\|\bar{\alpha}_n - \bar{\beta}_n\|_\Sigma^2 \leq 2\mathbb{E}\frac{\|\mu_0 - \nu_0\|_{\Sigma^{-1}}^2}{(\gamma n)^2} + 2\frac{\mathbb{E}\|\mu_n - \nu_n\|_{\Sigma^{-1}}^2}{(\gamma n)^2} + \frac{2}{n^2}\sum_{i=0}^{n-1}\operatorname{tr}\Sigma^{-1}\mathbb{E}(\xi_{i+1}\otimes\xi_{i+1}).$$

We compute $\sum_{i=0}^{n-1}\operatorname{tr}\Sigma^{-1}\mathbb{E}(\xi_{i+1}\otimes\xi_{i+1}) = \sum_{i=0}^{n-1}\operatorname{tr}\Sigma^{-1}C = n\operatorname{tr}\Sigma^{-1}C$ and, using Lemma 23, we bound $\mathbb{E}\|\mu_n - \nu_n\|_{\Sigma^{-1}}^2$ as

$$\frac{\mathbb{E}\|\mu_n - \nu_n\|_{\Sigma^{-1}}^2}{(\gamma n)^2} \leq \frac{\|\mu_0 - \nu_0\|_{\Sigma^{-1}}^2}{(\gamma n)^2} + \frac{1}{n}\operatorname{tr}\Sigma^{-1}C.$$

This implies the final bound

$$\mathbb{E}\|\bar{\alpha}_n - \bar{\beta}_n\|_\Sigma^2 \leq 4\frac{\|\mu_0 - \nu_0\|_{\Sigma^{-1}}^2}{(\gamma n)^2} + \frac{4}{n}\operatorname{tr}\Sigma^{-1}C.$$

□

4.C.2 Application of Lemma 24 to Prove Proposition 9

First of all we define the sequence

$$\eta_n^* = \nabla h(\theta_*) - n\gamma\nabla f(\theta_*). \tag{4.16}$$

By definition of θ_*, $-\nabla f(\theta_*) \in \partial g(\theta_*)$ then $\eta_n^* \in \partial(h + n\gamma g)\theta_*$ and $\theta_* = \nabla h_n^*(\eta_n^*)$. Therefore the sequence η_n^* is obtained by iterating DA started from the solution of the problem θ_*.

We note then than Lemma 24 applied to $(\mu_n = \eta_n, \alpha_n = \theta_n)$ and $(\nu_n = \eta_n^*, \beta_n = \theta_*)$ gives the first bound of Proposition 9.

On the other hand, when considering the noiseless iterates (ω_n, ϕ_n) defined by $\omega_n = \omega_{n-1} - \gamma\Sigma(\phi_{n-1} - \theta_\Sigma)$ and $\phi_n = \nabla h_n^*(\omega_n)$, started from the same point $\phi_0 = \theta_0$, we obtain, following Proposition 8, for γ such that $h - \gamma f$ is convex, the bound

$$\frac{1}{2}\|\bar{\phi}_n - \theta_*\|_\Sigma^2 \leq \psi(\bar{\phi}_n) - \psi(\theta_*) \leq \frac{D_h(\theta_*, \theta_0)}{\gamma n}.$$

Therefore, considering the difference between the semi-stochastic and the noiseless iterate $(\eta_n - \omega_n)$ which verifies the same equation $\eta_n - \omega_n = \eta_{n-1} - \omega_{n-1} - \gamma\Sigma(\theta_{n-1} - \phi_{n-1}) + \gamma\xi_n$ with $\theta_0 - \phi_0 = 0$ as initial value, we may apply Lemma 24 to show

$$\mathbb{E}\|\bar{\theta}_n - \bar{\phi}_n\|_\Sigma^2 \leq \frac{4}{n} \operatorname{tr} \Sigma^{-1} C.$$

And by the Cauchy-Schwarz inequality

$$\begin{aligned} \mathbb{E}\|\bar{\theta}_n - \theta_*\|_\Sigma^2 &\leq 2\mathbb{E}\|\bar{\theta}_n - \bar{\phi}_n\|_\Sigma^2 + 2\mathbb{E}\|\bar{\phi}_n - \theta_*\|_\Sigma^2 \\ &\leq \frac{8}{n} \operatorname{tr} \Sigma^{-1} C + 4\frac{D_h(\theta_*, \theta_0)}{\gamma n}, \end{aligned}$$

which proves the second bound of Proposition 9.

It is worth noting that the condition on the step-size of Lemma 23 is less restrictive than in Proposition 9. Indeed for all γ such that $2h - \gamma f$ is convex, the difference between the dual iterates of the stochastic and deterministic recursions stay close but the deterministic iterates only converge to the solution for γ such that $h - \gamma f$ is convex.

4.D Proof of Proposition 10

In this section, we prove Proposition 10. The proof technique is similar to Proposition 9 but with the additional difficulty of the multiplicative noise.

We first note that Assumption **(A11)** is equivalent by the Cauchy-Schwarz inequality to

$$\mathbb{E}\langle x_n, Mx_n\rangle\langle x_n, Nx_n\rangle \leq \kappa \operatorname{tr}(M\Sigma) \operatorname{tr}(N\Sigma), \tag{4.17}$$

for all positive semi-definite symmetric matrices M and N, see, e.,g., proof in Chapter 3. We will often use, in the following demonstrations, Eq. (4.17) and its direct corollary

$$\langle x_n, Mx_n\rangle x_n \otimes x_n \preccurlyeq \kappa \operatorname{tr}(M\Sigma)\Sigma, \tag{4.18}$$

without always referring to it.

4.D.1 A Simple Proof for the Bounded Constrained Case

We first prove Proposition 10 for the constrained case. It is then a simple corollary of Proposition 9.

Let us denote by $\mathcal{C}$ a bounded convex set and consider the constrained problem ($g = \mathbf{1}_{\mathcal{C}}$). We remind that the general stochastic oracle for SDA in least-squares regression is

$$\nabla f_n(\theta) = (\Sigma + \zeta_n)(\theta - \theta_\Sigma) - \xi_n, \text{ for } \theta \in \mathbb{R}^d,$$

with $\zeta_n = x_n \otimes x_n - \Sigma$. We denote by $r = \max_{\theta \in \mathcal{C}} \|\theta - \theta_\Sigma\|_2$ and we show that the noise covariance is directly bounded, despite the multiplicative noise:

$$\mathbb{E}\Big[\big(\nabla f_n(\theta) - \nabla f(\theta)\big) \otimes \big(\nabla f_n(\theta) - \nabla f(\theta)\big)\Big] \preccurlyeq 2\mathbb{E}\big[\zeta_n(\theta - \theta_\Sigma) \otimes (\theta - \theta_\Sigma)\zeta_n\big] + 2\mathbb{E}\xi_n \otimes \xi_n,$$

and using Assumption **(A11)**

$$\mathbb{E}\big[\zeta_n(\theta - \theta_\Sigma) \otimes (\theta - \theta_\Sigma)\zeta_n\big] \preccurlyeq r^2 \mathbb{E}\zeta_n\zeta_n \preccurlyeq r^2\kappa(\operatorname{tr}\Sigma)\Sigma.$$

Therefore

$$\mathbb{E}\Big[\big(\nabla f_n(\theta) - \nabla f(\theta)\big) \otimes \big(\nabla f_n(\theta) - \nabla f(\theta)\big)\Big] \preccurlyeq 2\big(\sigma^2 + r^2\kappa(\operatorname{tr}\Sigma)\big)\Sigma.$$

Hence Proposition 9 already implies for all step-size such that $h - \gamma f$ is convex

$$\frac{1}{2}\mathbb{E}\|\bar{\theta}_n - \theta_*\|_\Sigma^2 \leq 2\frac{D_h(\theta_*, \theta_0)}{\gamma n} + \frac{8d}{n}(\sigma^2 + \kappa r^2 \operatorname{tr}\Sigma).$$

4.D.2 A General Result

We prove in this section a more general result than Proposition 10 under the additional assumption

(A14) There exists $b \in [0, 1]$ and $\mu_b > 0$ such that $h - \frac{\mu_b}{2}\|\cdot\|_{\Sigma^b}^2$ is convex.

Proposition 12. *Assume **(A2-4)** and **(A7-14)**. Consider the recursion in Eq. (4.8). For any constant step-size γ such that $\gamma \leq \min\{\frac{\mu_b}{4\kappa \operatorname{tr}\Sigma^{1-b}}, \frac{1}{\kappa L d}\}$. Then*

$$\begin{aligned}\frac{1}{2}\mathbb{E}\|\bar{\theta}_n - \theta_*\|_\Sigma^2 \leq{}& 2\frac{D_h(\theta_*, \theta_0)}{\gamma n} + \frac{24}{n}\operatorname{tr}\Sigma^{-1}C + \frac{16\kappa d\gamma}{n\mu_b}\operatorname{tr} C\Sigma^{-b} \\ &+ \frac{8\kappa d}{n}\left(\frac{4\kappa\gamma \operatorname{tr}\Sigma^{1-b}}{\mu_b} + 3\right)\|\theta_* - \theta_\Sigma\|_\Sigma^2 + 80\frac{\kappa d}{\gamma n^2}D_h(\theta_*, \theta_0) + \frac{16\kappa d}{n^2}g(\theta_0).\end{aligned}$$

We note Assumption **(14)** is always satisfied for $b = 1$, for which it is Assumption **(13)**. Therefore Proposition 12 directly implies Proposition 10 as a corollary. We prove now two auxiliary lemmas which will be used in the proof of the Proposition 12.

4.D.3 Two Auxiliary Results for Least-Squares Objectives

For $b \in [0,1]$, we denote by T_b the operator $T_b = \mathbb{E}[\langle x, \Sigma^{-b}x\rangle x \otimes x]$. We first prove that, for least-square objectives, the sum of the function evaluated along the primal iterates remains bounded.

Lemma 25. *Let us consider the recursion $\eta_n = \eta_{n-1} - \gamma x_n \otimes x_n(\theta_{n-1} - \theta_*) + \gamma\xi_n$ and assume g is positive and there exist μ_b such that $h - \frac{\mu_b}{2}\|\cdot\|^2_{\Sigma^b}$ is convex and κ such that $T_b \preccurlyeq \kappa \operatorname{tr}(\Sigma^{1-b})\Sigma$, then for $\gamma \leq \mu_b/(4\kappa \operatorname{tr}\Sigma^{1-b})$ and $\theta \in \mathcal{X}$ we have*

$$\mathbb{E}\sum_{i=0}^{n}[\psi(\theta_i) - \psi(\theta)] + \Big(1 - 4\gamma\kappa \operatorname{tr}(\Sigma^{1-b})/\mu_b\Big)\sum_{i=0}^{n}\frac{1}{2}\mathbb{E}\|\theta_i - \theta\|^2_\Sigma$$
$$\leq \frac{D_h(\theta,\theta_0) - \mathbb{E}D_h(\theta,\theta_{n+1})}{\gamma} + (n+1)\gamma/\mu_b \operatorname{tr}\Sigma^{-b}C + 4(n+1)\kappa \operatorname{tr}(\Sigma^{1-b})/\mu_b f(\theta) + g(\theta_0).$$

We note that we can also obtain a bound depending on $2\psi(\theta)$ rather than $4f(\theta)$ with a similar proof.

Proof. Let denote by $f_n(\theta) = x_n \otimes x_n(\theta - \theta_\Sigma) + \xi_n$. Then following the proof of Proposition 8 (see Eq. (4.14)) we have the expansion

$$\tilde{D}_n(\theta,\eta_n) - \tilde{D}_{n-1}(\theta,\eta_{n-1}) \leq -\gamma(g(\theta_n) - g(\theta)) - \gamma\langle \nabla f_n(\theta_{n-1}), \theta_{n-1} - \theta\rangle$$
$$- D_h(\theta_n,\theta_{n-1}) + \gamma\langle \nabla f_n(\theta_{n-1}), \theta_{n-1} - \theta_n\rangle. \quad (4.19)$$

Since $h - \frac{\mu_b}{2}\|\cdot\|^2_{\Sigma^b}$ is convex, using Proposition 11, we get that $D_h(\theta_n,\theta_{n-1}) \geq \frac{\mu_b}{2}\|\theta_n - \theta_{n-1}\|^2_{\Sigma^b}$. Let denote by $A = -D_h(\theta_n,\theta_{n-1}), +\gamma\langle \nabla f_n(\theta_{n-1}), \theta_{n-1} - \theta_n\rangle$,

$$\begin{aligned}
A &\leq -\frac{\mu_b}{2}\|\theta_n - \theta_{n-1}\|^2_{\Sigma^b} + \gamma\langle x_n \otimes x_n(\theta_{n-1} - \theta_\Sigma) + \xi_n, \theta_{n-1} - \theta_n\rangle \\
&\leq -\frac{\mu_b}{2}\|\theta_n - \theta_{n-1}\|^2_{\Sigma^b} \\
&\quad + \Big\langle \frac{\gamma\Sigma^{-b/2}}{\sqrt{\mu_b}}[x_n \otimes x_n(\theta_{n-1} - \theta_\Sigma) + \xi_n], \Sigma^{b/2}\sqrt{\mu_b}\theta_{n-1} - \theta_n\Big\rangle \\
&\leq \frac{\gamma^2\mu_b}{2}\|x_n \otimes x_n(\theta_{n-1} - \theta_\Sigma) + \xi_n\|^2_{\Sigma^{-b}} \\
&\quad - \tfrac{1}{2}\|\gamma\Sigma^{b/2}\sqrt{\mu_b}(\theta_n - \theta_{n-1}) - \tfrac{\gamma\Sigma^{-b/2}}{\sqrt{\mu_b}}[x_n \otimes x_n(\theta_{n-1} - \theta_\Sigma) + \xi_n]\|^2_2 \\
&\leq \frac{\gamma^2}{2\mu_b}\|x_n \otimes x_n(\theta_{n-1} - \theta_\Sigma) + \xi_n\|^2_{\Sigma^{-1}} \\
&\leq \frac{\gamma^2}{\mu_b}\|\theta_{n-1} - \theta_\Sigma\|^2_{T_b} + \frac{\gamma^2}{\mu_b}\|\xi_n\|^2_{\Sigma^{-b}}.
\end{aligned}$$

Thus, taking the conditional expectation and assuming that κ is such that $T_b \preccurlyeq \kappa \operatorname{tr}(\Sigma^{1-b})\Sigma$ we obtain

$$-\mathbb{E}[D_h(\theta_n,\theta_{n-1})\mathcal{F}_{n-1}] + \gamma\mathbb{E}\langle \nabla f_n(\theta_{n-1}), \theta_{n-1} - \theta_n\rangle\mathcal{F}_{n-1}]$$

$$\leq \frac{\gamma^2 \kappa \operatorname{tr}(\Sigma^{1-b})}{\mu_b} \|\theta_{n-1} - \theta_\Sigma\|_\Sigma^2 + \frac{\gamma^2}{\mu_b} \operatorname{tr} \Sigma^{-b} C.$$

Taking again the conditional expectation in Eq. (4.19), we have for $\theta \in \mathcal{X}$

$$\begin{aligned}
\mathbb{E}[\tilde{D}_n(\theta, \eta_n)|\mathcal{F}_{n-1}] - \tilde{D}_{n-1}(\theta, \eta_{n-1}) \leq & \frac{\gamma^2 \kappa}{\mu_b} \operatorname{tr}(\Sigma^{1-b}) \|\theta_{n-1} - \theta_\Sigma\|_\Sigma^2 + \frac{\gamma^2}{\mu_b} \operatorname{tr} \Sigma^{-b} C \\
& - \gamma \mathbb{E}[\langle x_n \otimes x_n(\theta_{n-1} - \theta_\Sigma) + \xi_n, \theta_{n-1} - \theta \rangle | \mathcal{F}_{n-1}] \\
& - \gamma(\mathbb{E}[g(\theta_n)|\mathcal{F}_{n-1}] - g(\theta)) \\
\leq & \frac{\gamma^2 \kappa}{\mu_b} \operatorname{tr}(\Sigma^{1-b}) \|\theta_{n-1} - \theta_\Sigma\|_\Sigma^2 - \gamma \langle \theta_{n-1} - \theta_\Sigma, \Sigma(\theta_{n-1} - \theta) \rangle \\
& + \frac{\gamma^2}{\mu_b} \operatorname{tr} \Sigma^{-b} C - \gamma(\mathbb{E}[g(\theta_n)|\mathcal{F}_{n-1}] - g(\theta)).
\end{aligned}$$

And we note that

$$-\gamma \langle \theta_{n-1} - \theta_\Sigma, \Sigma(\theta_{n-1} - \theta_*) \rangle = -\gamma[f(\theta_{n-1}) - f(\theta_*)] - \frac{\gamma}{2} \|\theta_{n-1} - \theta\|_\Sigma^2.$$

Therefore

$$\begin{aligned}
\mathbb{E}[\tilde{D}_n(\theta_*, \eta_n)|\mathcal{F}_{n-1}] - \tilde{D}_{n-1}(\theta_*, \eta_{n-1}) \leq & -\gamma[f(\theta_{n-1}) - f(\theta_*) + \mathbb{E}[g(\theta_n)|\mathcal{F}_{n-1}] - g(\theta_*)] \\
& - \frac{\gamma}{2} \Big(1 - 4 \frac{\gamma \kappa}{\mu_b} \operatorname{tr}(\Sigma^{1-b}) \Big) \|\theta_{n-1} - \theta_\Sigma\|_\Sigma^2 \\
& + 2 \frac{\gamma^2 \kappa}{\mu_b} \operatorname{tr}(\Sigma^{1-b}) \|\theta - \theta_\Sigma\|_\Sigma^2 + \frac{\gamma^2}{\mu_b} \operatorname{tr} \Sigma^{-b} C.
\end{aligned}$$

Taking the total expectation we obtain

$$\begin{aligned}
& \mathbb{E} f(\theta_{n-1}) - f(\theta_*) + \mathbb{E} g(\theta_n) - g(\theta_*)] + \frac{1}{2} \Big(1 - 4 \frac{\gamma \kappa}{\mu_b} \operatorname{tr}(\Sigma^{1-b}) \Big) \|\theta_{n-1} - \theta_\Sigma\|_\Sigma^2 \\
& \qquad \leq \frac{\mathbb{E} \tilde{D}_{n-1}(\theta_*, \eta_{n-1}) - \mathbb{E} \tilde{D}_n(\theta_*, \eta_n)}{\gamma} + 2 \frac{\gamma \kappa}{\mu_b} \operatorname{tr}(\Sigma^{1-b}) \|\theta - \theta_\Sigma\|_\Sigma^2 + \frac{\gamma}{\mu_b} \operatorname{tr} \Sigma^{-b} C,
\end{aligned}$$

which, summing from $i = 0$ to $i = n$, leads to

$$\begin{aligned}
& \sum_{i=0}^n [\mathbb{E} f(\theta_i) - f(\theta_*) + \mathbb{E} g(\theta_i) - g(\theta_*)]] + \Big(1 - 4 \frac{\gamma \kappa}{\mu_b} \operatorname{tr}(\Sigma^{1-b}) \Big) \sum_{i=0}^n \frac{1}{2} \|\theta_i - \theta_\Sigma\|_\Sigma^2 \leq \\
& \qquad \frac{D_h(\theta_*, \theta_0) - \mathbb{E} \tilde{D}_{n+1}(\theta_*, \eta_{n+1})}{\gamma} + 4 \frac{\gamma \kappa}{\mu_b} \operatorname{tr}(\Sigma^{1-b})(n+1) \|\theta - \theta_\Sigma\|_\Sigma^2 \\
& \qquad\qquad + (n+1) \frac{\gamma}{\mu_b} \operatorname{tr} \Sigma^{-b} C - \mathbb{E} g(\theta_{n+1}) + g(\theta_0).
\end{aligned}$$

The result follows if g is non negative. □

We now present an extension of Lemma 23 to least-squares objectives.

Lemma 26. *Let us consider two sequences of iterates (μ_k, α_k) and (ν_k, β_k) which satisfy the recursion $\mu_n - \nu_n = \mu_{n-1} - \nu_{n-1} - \gamma x_n \otimes x_n(\alpha_{n-1} - \beta_{n-1}) + \gamma\xi_n$, $\alpha_n = \nabla h_n^*(\mu_n)$ and $\beta_n = \nabla h_n^*(\nu_n)$ and denote by $C = \mathbb{E}[x_n \otimes x_n]$ for $n \geq 0$. Assume that γ is such that $h - \gamma T$ is convex. Then*

$$\mathbb{E}\|\mu_n - \nu_n\|_{\Sigma^{-1}}^2 \leq \mathbb{E}\|\mu_0 - \nu_0\|_{\Sigma^{-1}}^2 + 2\gamma^2 n \operatorname{tr}\Sigma^{-1}C.$$

We note that the condition $h - \gamma T$ is rather restrictive since bounds on T are often of the form d times a matrix. For instance Eq. (4.17) directly implies $T \preccurlyeq \kappa d\Sigma$. Even for independent normal data x_n with diagonal covariance matrix Σ we are able to derive the equality $T = (d+2)\Sigma$.

Proof. We expand

$$\begin{aligned}\|\mu_n - \nu_n\|_{\Sigma^{-1}}^2 &= \|\mu_{n-1} - \nu_{n-1}\|_{\Sigma^{-1}}^2 + \gamma^2\|x_n \otimes x_n(\alpha_{n-1} - \beta_{n-1}) + \xi_n\|_{\Sigma^{-1}}^2 \\ &\quad -2\gamma\langle x_n \otimes x_n(\alpha_{n-1} - \beta_{n-1}) + \xi_n, \Sigma^{-1}(\mu_{n-1} - \nu_{n-1})\rangle.\end{aligned}$$

Taking conditional expectations, we get

$$\begin{aligned}\mathbb{E}[\|\mu_n - \nu_n\|_{\Sigma^{-1}}^2|\mathcal{F}_{n-1}] &= \|\mu_{n-1} - \nu_{n-1}\|_{\Sigma^{-1}}^2 + \gamma^2\mathbb{E}[\|x_n \otimes x_n(\alpha_{n-1} - \beta_{n-1}) + \xi_n\|_{\Sigma^{-1}}^2|\mathcal{F}_{n-1}] \\ &\quad -2\gamma\mathbb{E}[\langle x_n \otimes x_n(\alpha_{n-1} - \beta_{n-1}) + \xi_n, \Sigma^{-1}(\mu_{n-1} - \nu_{n-1})\rangle|\mathcal{F}_{n-1}] \\ &= \|\mu_{n-1} - \nu_{n-1}\|_{\Sigma^{-1}}^2 + \gamma^2\mathbb{E}[\|x_n \otimes x_n(\alpha_{n-1} - \beta_{n-1}) + \xi_n\|_{\Sigma^{-1}}^2|\mathcal{F}_{n-1}] \\ &\quad -2\gamma\langle \alpha_{n-1} - \beta_{n-1}, \mu_{n-1} - \nu_{n-1}\rangle.\end{aligned}$$

Using $(a+b)^2 \leq 2a^2 + 2b^2$ and denoting by $B = \mathbb{E}[\|x_n \otimes x_n(\alpha_{n-1} - \beta_{n-1}) + \xi_n\|_{\Sigma^{-1}}^2|\mathcal{F}_{n-1}]$, this leads to

$$\begin{aligned}B &\leq 2\mathbb{E}[\|x_n \otimes x_n(\alpha_{n-1} - \beta_{n-1})\|_{\Sigma^{-1}}^2|\mathcal{F}_{n-1}] + 2\mathbb{E}[\|\xi_n\|_{\Sigma^{-1}}^2|\mathcal{F}_{n-1}] \\ &\leq 2\|\alpha_{n-1} - \beta_{n-1}\|_{\mathbb{E}[x_n \otimes x_n \Sigma^{-1} x_n \otimes x_n|\mathcal{F}_{n-1}]}^2 + 2\operatorname{tr}\Sigma^{-1}\mathbb{E}[\varepsilon_n \otimes \varepsilon_n|\mathcal{F}_{n-1}] \\ &\leq 2\|\alpha_{n-1} - \beta_{n-1}\|_T^2 + 2\operatorname{tr}\Sigma^{-1}C,\end{aligned}$$

with $T = \mathbb{E}[x \otimes x\Sigma^{-1}x \otimes x]$. Thus we obtain

$$\begin{aligned}\mathbb{E}[\|\mu_n - \nu_n\|_{\Sigma^{-1}}^2|\mathcal{F}_{n-1}] &\leq \|\mu_{n-1} - \nu_{n-1}\|_{\Sigma^{-1}}^2 + 2\gamma^2\operatorname{tr}\Sigma^{-1}C \\ &\quad -2\gamma\langle \mu_{n-1} - \nu_{n-1} - \gamma T(\alpha_{n-1} - \beta_{n-1}), \alpha_{n-1} - \beta_{n-1}\rangle \\ &\leq \|\mu_{n-1} - \nu_{n-1}\|_{\Sigma^{-1}}^2 + 2\gamma^2\operatorname{tr}\Sigma^{-1}C,\end{aligned}$$

assuming that γ is such $h - \gamma\frac{1}{2}\|\cdot\|_T^2$ is convex (as in the proof of Lemma 23). Taking global expectations, we have shown that

$$\mathbb{E}\|\mu_n - \nu_n\|_{\Sigma^{-1}}^2 \leq \mathbb{E}\|\mu_{n-1} - \nu_{n-1}\|_{\Sigma^{-1}}^2 + 2\gamma^2\operatorname{tr}\Sigma^{-1}C.$$

□

4.D.4 Bound on the Difference between Two Averages of Primal Variables

We present now the following lemma with is an analogue of Lemma 24 for the least-squares problem. It shows that the difference between the average of two sequences of primal iterates which follow the same recursion is $O(1/n)$.

Lemma 27. *Let us consider two sequences of iterates (μ_k, α_k) and (ν_k, β_k) which satisfy the recursion $\mu_n - \nu_n = \mu_{n-1} - \nu_{n-1} - \gamma x_n \otimes x_n(\alpha_{n-1} - \beta_{n-1}) + \gamma\xi_n$, $\alpha_n = \nabla h_n^*(\mu_n)$ and $\beta_n = \nabla h_n^*(\nu_n)$. For $n \geq 0$ denote by $C = \mathbb{E}[x_n \otimes x_n]$. Assume that γ is such that $h - \gamma T$ is convex and there exists κ such that $T \preccurlyeq \kappa d\Sigma$. Then*

$$\mathbb{E}\|\bar{\alpha}_n - \bar{\beta}_n\|_\Sigma^2 \leq 4\frac{\kappa d - 1}{n^2}\sum_{i=0}^{n-1}\mathbb{E}\|\alpha_i - \beta_i\|_\Sigma^2 + 4\frac{\|\eta_0 - \mu_0\|_{\Sigma^{-1}}^2}{(\gamma n)^2} + \frac{8}{n}\operatorname{tr}\Sigma^{-1}C.$$

Proof. Using the expansion $\mu_n - \nu_n = \mu_{n-1} - \nu_{n-1} - \gamma x_n \otimes x_n(\alpha_{n-1} - \beta_{n-1}) + \gamma\xi_n$, we derive

$$\begin{aligned}\Sigma(\alpha_n - \beta_n) &= (\Sigma - x_{n+1}\otimes x_{n+1})(\alpha_n - \beta_n) + x_{n+1}\otimes x_{n+1}(\alpha_n - \beta_n) \\ &= (\Sigma - x_{n+1}\otimes x_{n+1})(\alpha_n - \beta_n) + \frac{\mu_n - \nu_n - \mu_{n+1} + \nu_{n+1}}{\gamma} + \xi_{n+1}.\end{aligned}$$

We obtain by summing n times

$$\Sigma^{1/2}\sum_{i=0}^{n-1}(\alpha_i - \beta_i) = \sum_{i=0}^{n-1}\Sigma^{-1/2}X_{i+1} + \frac{\Sigma^{-1/2}(\mu_0 - \nu_0 - \mu_n + \nu_n)}{\gamma} + \sum_{i=0}^{n-1}\Sigma^{-1/2}\xi_{i+1},$$

where we denote by $X_i = (\Sigma - x_i \otimes x_i)(\alpha_{i-1} - \beta_{i-1})$ which is a square-integrable martingale difference sequence. We use $(a+b)^2 \leq 2(a^2 + b^2)$ to obtain

$$\|(\bar{\alpha}_n - \bar{\beta}_n)\|_\Sigma^2 \leq \frac{2}{n^2}\Big\|\sum_{i=0}^{n-1}\Sigma^{-1/2}X_{i+1} + \frac{\Sigma^{-1/2}(\mu_0 - \nu_0)}{\gamma} + \sum_{i=0}^{n-1}\Sigma^{-1/2}\xi_{i+1}\Big\|_2^2 + 2\frac{\|\mu_n - \nu_n\|_{\Sigma^{-1}}^2}{(\gamma n)^2}.$$

Therefore using martingale square moment inequalities which here amount to considering the variance of the sum as the sum of the variance, we have

$$\begin{aligned}\mathbb{E}\|(\bar{\alpha}_n - \bar{\beta}_n)\|_\Sigma^2 \leq \frac{4}{n^2}\sum_{i=0}^{n-1}\mathbb{E}\|X_{i+1}\|_{\Sigma^{-1}}^2 + 2\frac{\|\mu_0 - \nu_0\|_{\Sigma^{-1}}^2}{(\gamma n)^2} \\ + 2\frac{\mathbb{E}\|\mu_n - \nu_n\|_{\Sigma^{-1}}^2}{(\gamma n)^2} + \frac{4}{n^2}\sum_{i=0}^{n-1}\operatorname{tr}\Sigma^{-1}\mathbb{E}(\xi_{i+1}\otimes\xi_{i+1}). \quad (4.20)\end{aligned}$$

— The variance term may be bounded as

$$\sum_{i=0}^{n-1} \operatorname{tr} \Sigma^{-1} \mathbb{E}(\xi_{i+1} \otimes \xi_{i+1}) \leq \sum_{i=0}^{n-1} \operatorname{tr} \Sigma^{-1} C \leq n \operatorname{tr} \Sigma^{-1} C.$$

— Following Lemma 26 we bound the dual iterates $\mathbb{E}\|\mu_n - \nu_n\|_{\Sigma^{-1}}^2$ as

$$\frac{\mathbb{E}\|\mu_n - \nu_n\|_{\Sigma^{-1}}^2}{(\gamma n)^2} \leq \frac{\|\mu_0 - \nu_0\|_{\Sigma^{-1}}^2}{(\gamma n)^2} + \frac{2}{n} \operatorname{tr} \Sigma^{-1} C.$$

— The martingale difference sequence (X_i) satisfies

$$\begin{aligned}
\mathbb{E}\|\Sigma^{-1/2} X_{i+1}\|_2^2 &\leq \mathbb{E}\langle (\Sigma - x_{i+1} \otimes x_{i+1})(\alpha_i - \beta_i), \Sigma^{-1}(\Sigma - x_{i+1} \otimes x_{i+1})(\alpha_i - \beta_i) \rangle \\
&\leq \langle \alpha_i - \beta_i, \mathbb{E}[(\Sigma - x_{i+1} \otimes x_{i+1})^\top \Sigma^{-1} (\Sigma - x_{i+1} \otimes x_{i+1})](\alpha_i - \beta_i) \rangle \\
&\leq \langle \alpha_i - \beta_i, [\mathbb{E}(x_{i+1} \otimes x_{i+1})^\top \Sigma^{-1} x_{i+1} \otimes x_{i+1} - \Sigma](\alpha_i - \beta_i) \rangle \\
&\leq \langle \alpha_i - \beta_i, [T - \Sigma](\alpha_i - \beta_i) \rangle \\
&\leq (\kappa d - 1) \|\alpha_i - \beta_i\|_\Sigma^2.
\end{aligned}$$

Consequently we obtain in Eq. (4.20)

$$\mathbb{E}\|\Sigma^{1/2}(\bar{\alpha}_n - \bar{\beta}_n)\|_2^2 \leq 4 \frac{\kappa d - 1}{n^2} \sum_{i=0}^{n-1} \mathbb{E}\|\alpha_i - \beta_i\|_\Sigma^2 + 4 \frac{\|\mu_0 - \nu_0\|_{\Sigma^{-1}}^2}{(\gamma n)^2} + \frac{8}{n} \operatorname{tr} \Sigma^{-1} C.$$

□

4.D.5 Application of Lemma 27 to the Proof of Proposition 12

We are now able to prove Proposition 12 using Lemma 27.

Firstly we can directly apply Lemma 27 to $(\mu_n = \eta_n, \alpha_n = \theta_n)$ and $(\nu_n = \eta_n^*, \beta_n = \theta_*)$ where (η_n^*, θ_*) are defined in Eq. (4.16). This implies

$$\mathbb{E}\|\Sigma^{1/2}(\bar{\theta}_n - \theta_*)\|_2^2 \leq 4 \frac{\kappa d - 1}{n^2} \sum_{i=0}^{n-1} \mathbb{E}\|\theta_i - \theta_*\|_\Sigma^2 + 4 \frac{\|\nabla h(\theta_0) - \nabla h(\theta_*)\|_{\Sigma^{-1}}^2}{(\gamma n)^2} + \frac{8}{n} \operatorname{tr} \Sigma^{-1} C.$$

Following Lemma 25, the primal variables (θ_i) satisfy

$$\sum_{i=0}^{n-1} \mathbb{E}\|\theta_i - \theta_*\|_\Sigma^2 \leq 2 \frac{D_h(\theta_*, \theta_0)}{\gamma} + 2 \frac{n\gamma}{\mu_b} \operatorname{tr} \Sigma^{-b} C + \frac{8 n \gamma \kappa \operatorname{tr} \Sigma^{1-b}}{\mu_b} f(\theta_*) + 2 g(\theta_0).$$

This leads to the final bound

$$\begin{aligned}
\mathbb{E}\|\Sigma^{1/2}(\bar{\theta}_n - \theta_*)\|_2^2 &\leq 8 \frac{\kappa d - 1}{\gamma n^2} D_h(\theta_*, \theta_0) + 4 \frac{\|\nabla h(\theta_0) - \nabla h(\theta_*)\|_{\Sigma^{-1}}^2}{(\gamma n)^2} \\
&\quad + 8 \frac{1}{n} \operatorname{tr} \Sigma^{-1} C + 8 \frac{\kappa d - 1}{n} \frac{\gamma}{\mu_b} [\operatorname{tr} \Sigma^{-b} C + 4 \kappa \operatorname{tr} \Sigma^{1-b} f(\theta_*)] + 8 \frac{\kappa d - 1}{n^2} g(\theta_0). \quad (4.21)
\end{aligned}$$

This bound depends on $\|\cdot\|_{\Sigma^{-1}}$ which may be infinite. For this reason we compare again the noisy iterate θ_n to the noiseless iterate we still denote by (ϕ_n). We remind these iterates verify the recursion

$$\omega_n = \omega_{n-1} - \gamma\Sigma(\phi_{n-1} - \theta_\Sigma).$$

Therefore the difference $(\eta_n - \omega_n)$ satisfies the same form of recursion as (η_n):

$$\eta_n - \omega_n = \nabla\eta_{n-1} - \omega_{n-1} - \gamma x_n \otimes x_n(\theta_{n-1} - \phi_{n-1}) + \gamma\varepsilon_n,$$

with a different noise $\varepsilon_n = \xi_n - [x_n \otimes x_n - \Sigma](\phi_{n-1} - \theta_\Sigma)$ and 0 for initial value. Although the noise ε_n is different from ξ_n, its covariance is still bounded by

$$\begin{aligned}
\frac{1}{3}\mathbb{E}[\varepsilon_n \otimes \varepsilon_n] &\preccurlyeq \mathbb{E}[\xi_n \otimes \xi_n] + \mathbb{E}[[x_n \otimes x_n - \Sigma](\phi_{n-1} - \theta_*) \otimes (\phi_{n-1} - \theta_*)[x_n \otimes x_n - \Sigma]] \\
&\quad + \mathbb{E}[[x_n \otimes x_n - \Sigma](\theta_* - \theta_\Sigma) \otimes (\theta_* - \theta_\Sigma)[x_n \otimes x_n - \Sigma]] \\
&\preccurlyeq \mathbb{E}[\xi_n \otimes \xi_n] - \mathbb{E}[\Sigma(\phi_{n-1} - \theta_*)^{\otimes 2}\Sigma] - \mathbb{E}[\Sigma(\theta_* - \theta_\Sigma)^{\otimes 2}\Sigma] \\
&\quad + \mathbb{E}[x_n \otimes x_n(\phi_{n-1} - \theta_*)^{\otimes 2}x_n \otimes x_n] + \mathbb{E}[x_n \otimes x_n(\theta_* - \theta_\Sigma)^{\otimes 2}x_n \otimes x_n] \\
&\preccurlyeq \mathbb{E}[\xi_n \otimes \xi_n] + (\kappa - 1)\big(\|\phi_{n-1} - \theta_*\|_\Sigma^2 + \|\theta_* - \theta_\Sigma\|_\Sigma^2\big)\Sigma,
\end{aligned}$$

where we have use that for $z \in \mathbb{R}^d$, $\mathbb{E}\langle z, x_n\rangle^4 \leq \kappa\langle z, \Sigma z\rangle$. We may apply Proposition 8 and obtain

$$\mathbb{E}[\varepsilon_n \otimes \varepsilon_n] \preccurlyeq 3C + \frac{6(\kappa - 1)}{\gamma n} D_h(\theta_*, \theta_0)\Sigma + 6(\kappa - 1)f(\theta_*).$$

Thereby Lemma 27 can be applied with $\theta_0 = \alpha_0$ and we get

$$\mathbb{E}\|\bar{\theta}_n - \bar{\phi}_n\|_\Sigma^2 \leq 4\frac{\kappa d - 1}{n^2}\sum_{i=0}^{n-1}\mathbb{E}\|\theta_i - \phi_i\|_\Sigma^2 + \frac{8}{n}\operatorname{tr}\Sigma^{-1}\mathbb{E}[\varepsilon_n \otimes \varepsilon_n].$$

As before we apply Lemma 25 to have

$$\begin{aligned}
\sum_{i=0}^{n-1}\mathbb{E}\|\theta_i - \phi_i\|_\Sigma^2 &\leq \left[2\sum_{i=0}^{n-1}\mathbb{E}\|\theta_i - \theta_*\|_\Sigma^2 + 2\sum_{i=0}^{n-1}\|\phi_i - \theta_*\|_\Sigma^2\right] \\
&\leq \left[\frac{8D_h(\theta_*, \theta_0)}{\gamma} + \frac{n\gamma}{\mu_b}4\operatorname{tr}\Sigma^{-b}C + \frac{16n\gamma\kappa\operatorname{tr}\Sigma^{1-b}}{\mu_b}f(\theta_*) + 4g(\theta_0)\right].
\end{aligned}$$

Therefore

$$\begin{aligned}
\mathbb{E}\|\bar{\theta}_n - \bar{\phi}_n\|_\Sigma^2 \leq 4\frac{\kappa d - 1}{n^2}&\left[8\frac{D_h(\theta_*, \theta_0)}{\gamma} + \frac{n\gamma}{\mu_b}4\operatorname{tr}\Sigma^{-b}C + \frac{16n\gamma\kappa\operatorname{tr}\Sigma^{1-b}}{\mu_b}f(\theta_*) + 4g(\theta_0)\right] \\
&+ \frac{8}{n}\left[3\operatorname{tr}\Sigma^{-1}C + \frac{6(\kappa - 1)}{\gamma n}D_h(\theta_*, \theta_0)d + 6(\kappa - 1)f(\theta_*)d\right].
\end{aligned}$$

And rearranging terms we obtain

$$\mathbb{E}\|\bar{\theta}_n - \bar{\phi}_n\|_\Sigma^2 \leq 80\frac{\kappa d}{\gamma n^2} D_h(\theta_*, \theta_0) + \frac{64\kappa^2 d\gamma}{n\mu_b} \operatorname{tr} \Sigma^{1-b} f(\theta_*) + \frac{16\kappa d\gamma}{n\mu_b} \operatorname{tr} C\Sigma^{-b} + \frac{16\kappa d}{n^2} g(\theta_0) + \frac{24}{n} \operatorname{tr} \Sigma^{-1} C + \frac{48\kappa d}{n} f(\theta_*).$$

And by the Cauchy-Schwarz inequality ($\mathbb{E}\|\bar{\theta}_n - \theta_*\|_\Sigma^2 \leq 2\mathbb{E}\|\bar{\theta}_n - \bar{\phi}_n\|_\Sigma^2 + 2\mathbb{E}\|\bar{\phi}_n - \theta_*\|_\Sigma^2$)

$$\frac{1}{2}\mathbb{E}\|\bar{\theta}_n - \theta_*\|_\Sigma^2 \leq 2\frac{D_h(\theta_*, \theta_0)}{\gamma n} + \frac{24}{n} \operatorname{tr} \Sigma^{-1} C + \frac{16\kappa d\gamma}{n\mu_b} \operatorname{tr} C\Sigma^{-b} + \frac{16\kappa d}{n}\Big(\frac{4\kappa\gamma \operatorname{tr} \Sigma^{1-b}}{\mu_b} + 3\Big) f(\theta_*) + 80\frac{\kappa d}{\gamma n^2} D_h(\theta_*, \theta_0) + \frac{16\kappa d}{n^2} g(\theta_0),$$

which proves the second bound of Proposition 12.

4.D.6 A Corollary of Proposition 12 for h with an Euclidean Behavior

When h rather behaves as an Euclidean norm, we may replace Assumptions **(A12-13)** by the following:

(A12') There exists $\mu_h > 0$ such that $h - \frac{\mu_h}{2}\|\cdot\|_2^2$ is convex.

(A13 ') There exists R^2 such that $\mathbb{E}[\|x_n\|_2^2 x_n \otimes x_n] \preccurlyeq R^2\Sigma$.

And Proposition 12 implies the following corollary.

Corollary 6. *Assume For any constant step-size γ such that $\gamma \leq \min\{\frac{\mu_h}{4\kappa R^2}, \frac{R^2}{4\kappa d}\}$. Then*

$$\frac{1}{2}\mathbb{E}\|\bar{\theta}_n - \theta_*\|_\Sigma^2 \leq 2\frac{D_h(\theta_*, \theta_0)}{\gamma n} + \frac{8}{n}\Big(3 + \frac{4\gamma\kappa R^2}{\mu_h}\Big)\Big(\sigma^2 d + \kappa d\|\theta_* - \theta_\Sigma\|_\Sigma^2\Big) + \frac{16\kappa d}{n^2}\Big(\frac{5D_h(\theta_*, \theta_0)}{\gamma} + g(\theta_0)\Big).$$

This corollary would pave the way for a general result for larger step-size γ without the condition $\gamma \leq \frac{R^2}{4\kappa d}$. Unfortunately the latter seems not improvable, as noted after Lemma 26.

4.E Lower Bound for Non-Strongly Convex Quadratic Regularization

We derive, in this section, a lower bound on the performance of SDA when f is the linear form $f(\theta) = \langle a, \theta\rangle$ with $a \in \mathbb{R}^d$ and g is a non-strongly convex quadratic function. We assume that the vector a is not available and we only have access to

estimates of the gradient

$$\nabla f_n(\theta) = a + \xi_n \text{ for } n \geq 1, \tag{4.22}$$

where (ξ_n) is an uncorrelated zero-mean noise sequence with bounded covariance.

Proposition 13. *For any $d \geq 2$, $L > 0$; $\gamma > 0$ and finite time horizon $N \geq 1$, there exists a quadratic function g L-smooth such that for any uncorrelated zero-mean noise sequence (ξ_n) with bounded covariance $\mathbb{E}[\xi_n \otimes \xi_n] = \sigma^2 L I_d$, SDA with constant step-size γ applied with the oracle Eq. (4.22) satisfies*

$$\psi(\bar{\theta}_N) - \psi(\theta_*) \geq \frac{\sigma^2}{12} \min\{(L\gamma)^2, 1\}.$$

Proof. For sake of clarity, we consider $d = 2$ and $a = 0$. Thus $f(\theta) = \mathbb{E}\langle \xi_n, \theta \rangle = 0$. Let $g(\theta) = \frac{1}{2}\langle \theta, A\theta \rangle$ be a quadratic form with $A = \begin{pmatrix} L & 0 \\ 0 & \mu \end{pmatrix}$ for $L \geq \mu > 0$ with μ possibly arbitrary small. The noise (ξ_n) is assumed to be uncorrelated zero-mean with bounded covariance $\mathbb{E}[\xi_n \otimes \xi_n] = \sigma^2 L I_2$. The stochastic dual algorithm with step-size γ takes the form:

$$\begin{aligned} \theta_n &= \nabla h_n^*(-n\gamma \bar{\xi}_n) \\ &= \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} \Big\{ \langle \bar{\xi}_n, \theta \rangle + \frac{1}{2}\langle \theta, A\theta \rangle + \frac{1}{2n\gamma} \|\theta\|_2^2 \Big\} \\ &= \gamma n (I + \gamma n A)^{-1} \bar{\xi}_n. \end{aligned}$$

And

$$\begin{aligned} \bar{\theta}_n &= \frac{\gamma}{n} \sum_{k=1}^{n-1} \sum_{j=1}^{k} k (I + \gamma k A)^{-1} \frac{1}{k} \xi_j \\ &= \frac{\gamma}{n} \sum_{j=1}^{n-1} \Big(\sum_{k=j}^{n-1} (I + \gamma k A)^{-1} \Big) \xi_j. \end{aligned}$$

Therefore using standard martingale square moment inequalities

$$\begin{aligned} \mathbb{E}\langle \bar{\theta}_n, A\bar{\theta}_n \rangle &= \frac{\gamma^2}{n^2} \sum_{j=1}^{n} \mathbb{E}\Big\langle \xi_j \Big(\sum_{k=j}^{n} (I + \gamma k A)^{-1} \Big), A \Big(\sum_{k=j}^{n} (I + \gamma k A)^{-1} \Big) \xi_j \Big\rangle \\ &= \frac{\gamma^2 \sigma^2 L}{n^2} \operatorname{tr} \sum_{j=1}^{n} \Big(\sum_{k=j}^{n} (I + \gamma k A)^{-1} \Big) A \Big(\sum_{k=j}^{n} (I + \gamma k A)^{-1} \Big) I_2 \\ &= \frac{\gamma^2 \sigma^2 L}{n^2} \sum_{j=1}^{n} \Big[L \Big(\sum_{k=j}^{n} \frac{1}{1 + \gamma L k} \Big)^2 + \mu \Big(\sum_{k=j}^{n} \frac{1}{1 + \gamma \mu k} \Big)^2 \Big]. \end{aligned}$$

And

$$\begin{aligned}\mathbb{E}\langle\bar{\theta}_n, A\bar{\theta}_n\rangle &\geq \frac{\gamma^2\sigma^2 L}{n^2}\Big[\frac{L}{(1+\gamma L n)^2}+\frac{\mu}{(1+\gamma\mu n)^2}\Big]\sum_{j=1}^{n}(n-j)^2 \\ &\geq \frac{n\sigma^2\gamma^2 L}{3}\Big[\frac{L}{(1+\gamma L n)^2}+\frac{\mu}{(1+\gamma\mu n)^2}\Big] \geq \frac{n\sigma^2\gamma^2}{3}\frac{\mu}{(1+\gamma\mu n)^2} \\ &\geq \frac{\sigma^2 L}{12}\min\Big(n\mu\gamma^2, \frac{1}{\mu n}\Big).\end{aligned}$$

Conclude by taking $\mu = L/N$.

The proof is the same for $d \geq 2$ by considering $A = \mathrm{diag}(L, \ldots, L, L\mu)$ with $d-1$ L. □

4.F Lower Bound for Stochastic Approximation Problems

In this section we relate the problem of aggregation of estimators to the stochastic convex optimization problem, i.e., minimizing a convex function, given only unbiased estimates of its gradients. We will consider the regression and the classification with hinge loss problems which will individually provide lower bounds for quadratic and linear functions. We follow here Tsybakov [2003], Lecué [2006], Agarwal et al. [2012].

4.F.1 Oracle Complexity of Stochastic Convex Optimization

Beforehand we describe the stochastic oracle model formalism as done by Nemirovsky and Yudin [1983], Agarwal et al. [2012], Raginsky and Rakhlin [2011]. For a given class of problems we aim to determine lower bounds on the number of queries to a stochastic first-order oracle needed to optimize to a certain precision any function in this class. To this end we have the following definition.

Definition 3 (Agarwal et al. [2012])**.** *For a given constraint convex set $\mathcal{C}$, and a function class $\mathcal{S}$, a first-order stochastic oracle is a random mapping $\pi : \mathcal{C} \times \mathcal{S} \to \mathbb{R} \times \mathbb{R}^d$ of the form*

$$\phi(\theta, f) = (\tilde{f}(\theta), g(\theta)),$$

such that

$$\mathbb{E}\tilde{f}(\theta) = f(\theta); \qquad \mathbb{E}g(\theta) = \nabla f(\theta),$$

and there exists a constant $C < \infty$ such that for every $\theta \in \mathbb{R}^d$

$$\mathbb{E}[\|g(\theta) - \nabla f(\theta)\|^2] \leq C(1 + \|\theta\|^2).$$

The class of first-order stochastic oracle is denoted by Φ. A stochastic approximation algorithm M is a method which approximately minimizes a function f by querying, at each iteration i, the oracle at the point θ_i. The oracle answers with the

information $\phi(\theta_i, f)$ and the method uses all the information $\{\phi(\theta_0, f), \ldots, \phi(\theta_i, f)\}$ to build a new point θ_{i+1}. For $n \in \mathbb{N}$ we denote by $\mathcal{M}_n$ the class of all such methods that are allowed to make n queries. As done by Agarwal et al. [2012], we denote the error of the method M on the function f after n steps as

$$\varepsilon_n(M, f, \mathcal{C}, \phi) = f(\theta_n) - \min_{\theta \in \mathcal{C}} f(\theta).$$

Given a class of functions $\mathcal{S}$, an oracle ϕ and a convex constraint set $\mathcal{C}$, Agarwal et al. [2012] also defines the minimax error as

$$\varepsilon_n^*(\mathcal{S}, \mathcal{C}, \phi) = \inf_{M \in \mathcal{M}_n} \sup_{f \in \mathcal{S}} \mathbb{E}_\phi \varepsilon_n(M, f, \mathcal{C}, \phi).$$

We will lower bound this minimax error by relating convex stochastic approximation with convex aggregation of estimators [Juditsky and Nemirovski, 2000, Tsybakov, 2003].

4.F.2 Convex Aggregation of Estimators

Let $(\mathcal{X}, \mathcal{A})$ be a measurable space and $\mathcal{Y} \subset \mathbb{R}$. We consider random variables (X, Y) on $\mathcal{X} \times \mathcal{Y}$ with probability distribution denoted by π. We observe n i.i.d. pairs $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ which follow the law π and we want to predict the output Y for any feature $X \in \mathcal{X}$ by a prediction $f(X)$ for a measurable function f from $\mathcal{X}$ to $\mathbb{R}$. For this purpose we want to minimize the risk defined by

$$A(f) = \mathbb{E}[\ell(f(X), Y)],$$

for any measurable function f from $\mathcal{X}$ to $\mathbb{R}$ and $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ a loss function.

We consider we have access to d different arbitrary estimators $\mathcal{F} = \{f_1, \ldots, f_d\}$ with values in $\mathcal{Y}$. We denote their convex hull by $\mathcal{C} = conv(f_1, \ldots, f_d)$. The aim of convex aggregation is to build a new estimator which is a convex combination of the different f_i and behaves as the best among the estimators f_i. The aggregation problem is equivalent to a minimization problem over the simplex Δ_d since for $f \in \mathcal{C}$ there is $\theta \in \Delta_d$ such that $f = \sum_{i=1}^d \theta(i) f_i$. Therefore, defining $B(\theta) = A\Big(\sum_{i=1}^d \theta(i) f_i\Big)$, we have

$$\min_{f \in \mathcal{C}} A(f) = \min_{\theta \in \Delta_d} B(\theta).$$

We denote by $F : \mathcal{X} \to \mathbb{R}^d, x \mapsto (f_1(x), \ldots, f_d(x))$ the function whose the ith coordinate is the function f_i, and we have

$$B(\theta) = \mathbb{E}[\ell(\langle F(X), \theta \rangle, Y)].$$

Therefore the convex aggregation problem of minimizing $A(f)$ over the convex hull of $\mathcal{F}$ is formally equivalent to the stochastic approximation problem of minimizing, over the simplex Δ_d, the function $B(\theta) = \mathbb{E}[\ell(\langle F(X), \theta \rangle, Y)]$, given only unbiased estimates of its gradient $\nabla B_n(\theta) = \nabla \ell(\langle F(x_n), \theta \rangle, y_n)$. Hence lower bounds on convex

aggregation problems provide lower bounds on stochastic approximation problems studied in this chapter.

4.F.3 Aggregation in Regression and Application to Oracle Complexity of Stochastic Quadratic Optimization

We first consider the regression problem for which $\mathcal{Y} = \mathbb{R}$. We rely substantially on Tsybakov [2003]. The regression model is

$$Y_i = f_*(X_i) + \xi_i, \text{ for } i = 1, \dots, n,$$

where $X_1, \dots, X_n$ are i.i.d. random vectors of $\mathcal{X}$ of law P^X and ξ_i are i.i.d. Gaussian $\mathcal{N}(0, \sigma^2)$ random variables such that $(\xi_1, \dots, \xi_n)$ is independent of $(X_1, \dots, X_n)$ and $f_* : \mathcal{X} \to \mathbb{R}$ is the regression function. Regression problem aims to estimate the unknown regression function f_* based on the data D_n by minimizing the risk

$$A_{\text{reg}}(f, f_*) = \mathbb{E}(f(X) - f_*(X))^2.$$

The problem of the optimal rate of convex aggregation has been studied by Tsybakov [2003]. We reintroduce his notations and assumptions for sake of completeness.

Let denote by $\mathcal{F}_0 = \{f : \|f\|_\infty \leq L\}$ for $L > 0$ and assume that

(B1) There exists a cube $S \subset \mathcal{X}$ such that P^X admits a bounded density μ on S w.r.t. the Lebesgue measure and $\mu(x) \geq \mu_0 > 0$ for all $x \in S$.

(B2) There exists a constant c_0 such that $d \leq c_0 \exp(n)$.

We have the following result

Theorem 9 (Theorem 2, Tsybakov [2003])**.** *Under assumptions **(B1-2)** we have*

$$\sup_{f_1, \dots, f_d \in \mathcal{F}_0} \inf_{T_n} \sup_{f_* \in \mathcal{F}_0} \left[\mathbb{E}_{D_n} A_{reg}(T_n, f_*) - \min_{f \in \mathcal{C}} A_{reg}(f, f_*) \right] \geq c \zeta_n(d),$$

for some constant $c > 0$ and any integer n, where $\inf_{T_n}$ denotes the infimum over all estimators, $\mathbb{E}_{D_n}$ denotes the expectation with regard to the probability distribution of the data D_n and

$$\zeta_n(d) = \begin{cases} d/n & \text{if } d \leq \sqrt{n} \\ \sqrt{\frac{1}{n} \log\left(\frac{d}{\sqrt{n}} + 1\right)} & \text{if } d > \sqrt{n}. \end{cases}$$

We relate now the problem of convex aggregation of regression functions to the problem of stochastic quadratic functions optimization. Consider $\mathcal{F} = \{f_1, \dots, f_d\}$ the set of estimators given by Proposition 9 and denote by $F : \mathcal{X} \to \mathbb{R}^d, x \mapsto (f_1(x), \dots, f_d(x))$. For $f \in \mathcal{C}$, there is $\theta \in \Delta_d$ such that $f = \sum_{i=1}^d \theta(i) f_i$ and we obtain

$$A_{\text{reg}}(f, f_*) = \mathbb{E}[(\langle \theta, F(X) \rangle - f_*(X))^2] = B(\theta),$$

where $B(\theta) = \langle \theta, \mathbb{E}[F(X) \otimes F(X)], \theta \rangle - 2\langle \theta \mathbb{E}[f_*(X) F(X)] \rangle + \mathbb{E}[f_*(X)^2]$ is a quadratic

function. This set enables us to construct a difficult subclass of quadratic functions:

$$\mathcal{G}_{\text{quad}} = \Big\{ B(\theta) = \frac{1}{2}\mathbb{E}[(\langle \theta, F(X)\rangle - f_*(X))^2]; f_* \in \mathcal{F}_0 \Big\}.$$

We also define the first-order stochastic oracle ϕ_{quad} on $\mathcal{G}_{\text{quad}}$ as follows

$$\phi_{\text{quad}}(\theta, f) = \left(\frac{1}{2}(\langle \theta, F(x)\rangle - f_*(x))^2, (\langle \theta, F(x)\rangle - f_*(x))F(x) \right), \text{ for } x \sim P^X.$$

We can optimize B with a stochastic approximation algorithm $M \in \mathcal{M}_n$ to obtain $\theta_n \in \Delta_d$ and therefore build a estimator $T_n = \sum_{i=1}^d \theta_n(i) f_i$ which belongs to $\mathcal{C}$. Moreover we have

$$A_{\text{reg}}(T_n, f_*) = B(\theta_n) \text{ and } \min_{f \in \mathcal{C}} A_{\text{reg}}(f, f_*) = \min_{\theta \in \Delta_d} B(\theta).$$

Consequently, for the oracle ϕ_{quad} and the class $\mathcal{G}_{\text{quad}}$ Proposition 9 implies that

$$\varepsilon_n^*(\mathcal{G}_{\text{quad}}, \Delta_d, \phi_{\text{quad}}) \geq c\zeta_n(d). \tag{4.23}$$

And we have proven the following minimax oracle complexity.

Proposition 14. *Let Δ_d be the simplex. Then there exists universal constants $c_0 > 0$ and $c > 0$ such that the minimax oracle complexity over the class $\mathcal{S}_{quad}$ of quadratic functions satisfies the following lower bounds:*

— *For $d \leq \sqrt{n}$*

$$\sup_{\phi \in \Phi} \varepsilon_n^*(\mathcal{S}_{quad}, \Delta_d, \phi) \geq c\frac{d}{n}.$$

— *For $\sqrt{n} \leq d \leq c_0 \exp(n)$*

$$\sup_{\phi \in \Phi} \varepsilon_n^*(\mathcal{S}_{quad}, \Delta_d, \phi) \geq c\sqrt{\frac{1}{n}\log\Big(\frac{d}{\sqrt{n}} + 1\Big)}.$$

We note that without assumption on d the lower-bound for the class of quadratic functions is of order $O(1/n)$ but in high-dimensional settings it becomes of order $(1/\sqrt{n})$. Nevertheless we will see in the next section this lower-bound is always of order $(1/\sqrt{n})$ for the class of linear functions.

4.F.4 Aggregation in Classification and Application to Oracle Complexity of Stochastic Linear Optimization

We consider now the classification problem with the hinge loss for which $\mathcal{Y} = \{-1, 1\}$. We follow very closely the framework of Lecué [2006, 2007] and use their notations. We still consider random variables (X, Y) on $\mathcal{X} \times \mathcal{Y}$ with probability distribution denoted by π. We observe n i.i.d. pairs $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$

which follow the law π and we want to predict the label Y for any feature $X \in \mathcal{X}$ by minimizing the hinge risk defined by

$$A_{\text{cla}}(f) = \mathbb{E} \max(1 - Yf(X), 0),$$

for any measurable function f from $\mathcal{X}$ to $\mathbb{R}$. We consider we have access to d different estimators $\mathcal{F} = \{f_1, \ldots, f_d\}$ with values in $[-1, 1]$. We denote their convex hull by $\mathcal{C} = conv(f_1, \ldots, f_d)$. Lecué [2006, Theorem 1] and Lecué [2007, Theorem 2] provide a lower bound on this aggregation problem for classification we adapt to our specific case.

Proposition 15 (Adaptation of Theorem 2 of Lecué [2007] for $\kappa = \infty$)**.** *Let d, n be two integers such that $2 \log_2 d \leq n$. We assume that the input space $\mathcal{X}$ is infinite. There exists an absolute constant $c > 0$, and a set of prediction rules $\mathcal{F} = \{f_1, \ldots, f_n\}$ such that for any real-valued procedure T_n, there exists a probability measure π, for which*

$$\mathbb{E}_{D_n}[A_{cla}(T_n)] - \min_{f \in \mathcal{C}}(A_{cla}(f)) \geq c\sqrt{\frac{\log d}{n}}.$$

Proof. Theorem 2 of Lecué [2007] is stated under an additional Margin assumption MAH(κ) (see definition and notation below Eq. (9) in Lecué [2007]) on the probability distribution π, i.e., there exists a constant c_0 such that

$$\mathbb{E}[|f(X) - f^*(X)|] \leq c_0(A(f) - A^*)^{1/\kappa},$$

for any function f on $\mathcal{X}$ with values in $[-1, 1]$. Therefore taking $\kappa \to \infty$, we can always consider $c_0 = 2$. And the constant $c(\kappa)$ in Theorem 2 of Lecué [2007] is

$$c(\kappa) = c_0^{\kappa}(4e)^{-1}2^{-2\kappa(\kappa-1)/(2\kappa-1)}(\log 2)^{-\kappa/(2\kappa-1)},$$

which goes when $\kappa \to \infty$ to $c_\infty = \sqrt{2}/(4e\sqrt{\log 2})$. Hence taking $\kappa \to \infty$ in Theorem 2 of Lecué [2007] implies Proposition 15. We could also have plugged arguments of the proof of Theorem 14.5 of Devroye et al. [1996] to directly prove this result. □

We relate now the problem of convex aggregation of classifiers to the problem of optimizing a linear function on the simplex. Consider the set of prediction rules $\mathcal{F} = \{f_1, \ldots, f_n\}$ given by Proposition 15 and denote by $F : \mathcal{X} \to \mathbb{R}^d, x \mapsto (f_1(x), \ldots, f_d(x))$. For $f \in \mathcal{C}$, there is $\theta \in \Delta_d$ such that $f = \sum_{i=1}^d \theta(i) f_i$ and we obtain

$$A_{\text{cla}}(f) = \mathbb{E} \max(1 - Y\langle F(X), \theta\rangle, 0).$$

On the other hand, when the f_i are valued in $[-1, 1]$, the classification problem becomes equivalent to maximize the expectation $\mathbb{E}Yf(X)$ since the hinge loss is linear on $[-1, 1]$:

$$Y \in \{-1, 1\}, f(X) \in [-1, 1] \Longrightarrow Yf(X) \in [-1, 1] \Longrightarrow \mathbb{E} \max(1 - Yf(X), 0) = 1 - \mathbb{E}Yf(X).$$

Combining both, we obtain that

$$A_{\text{cla}}(f) = 1 - \langle \mathbb{E}[YF(X)], \theta \rangle = 1 + C(\theta),$$

where $C(\theta) = -\langle \mathbb{E}[YF(X)], \theta \rangle$ is a linear function. This set enables us to construct a difficult subclass of linear functions

$$\mathcal{G}_{\text{lin}} = \{C(\theta) = -\langle \mathbb{E}[YF(X)], \theta \rangle; (X, Y) \sim \pi\}.$$

We also define the first-order stochastic oracle ϕ_{lin} on $\mathcal{G}_{\text{lin}}$ as follows

$$\phi_{\text{lin}}(\theta, f) = \Big(\langle yF(x), \theta \rangle, yF(x) \Big), \text{ for } (x, y) \sim \pi.$$

As before we may optimize C with a stochastic approximation algorithm $M \in \mathcal{M}_n$ to obtain $\theta_n \in \Delta_d$ and therefore build a estimator $T_n = \sum_{i=1}^d \theta_n(i) f_i$ which belongs to $\mathcal{C}$. Moreover we have

$$A_{\text{cla}}(T_n) = C(\theta_n) \text{ and } \min_{f \in \mathcal{C}} A_{\text{cla}}(f) = \min_{\theta \in \Delta_d} C(\theta).$$

Consequently, for the oracle ϕ_{lin} and the class $\mathcal{G}_{\text{lin}}$ Proposition 9 implies that

$$\varepsilon_n^*(\mathcal{G}_{\text{lin}}, \Delta_d, \phi_{\text{lin}}) \geq c \sqrt{\frac{\log d}{n}}. \tag{4.24}$$

And we have proven the following minimax oracle complexity.

Proposition 16. *Let Δ_d be the simplex. Then there exists universal constant $c > 0$ such that the minimax oracle complexity over the class $\mathcal{S}_{lin}$ of linear functions satisfies the following lower bound for $2 \log_2 d \leq n$*

$$\sup_{\phi \in \Phi} \varepsilon_n^*(\mathcal{S}_{lin}, \Delta_d, \phi) \geq c \sqrt{\frac{\log(d)}{n}}.$$

4.G Lower Bounds on the Rates of Convergence of DA and MD Algorithms

Let us consider in this section that $f = 0$, $g(\theta) = \frac{1}{2\nu} \|\theta - \theta_*\|_2^2$ and $h = \frac{1}{2} \|\theta\|_2^2$. In this case, for $n \geq 1$, MD iterates (θ_n^{md}) verify

$$\theta_n^{\text{md}} = \arg\min_{\theta \in \mathbb{R}^d} \Big\{ \frac{1}{2\nu} \|\theta - \theta_*\|_2^2 + \frac{1}{2\gamma} \|\theta - \theta_{n-1}^{\text{md}}\|_2^2 \Big\}.$$

Therefore $\theta_n^{\text{md}} = \theta_* + \frac{1}{\gamma\nu}(\theta_{n-1}^{\text{md}} - \theta_*)$, $\theta_n^{\text{md}} - \theta_* = \frac{1}{(\gamma\nu)^n}(\theta_0^{\text{md}} - \theta_*)$ and

$$g(\theta_n^{\text{md}}) - g(\theta_*) = \frac{g(\theta_0^{\text{md}}) - g(\theta_*)}{(\gamma\nu)^{2n}}.$$

Whereas DA iterates (θ_n^{da}) satisfy

$$\theta_n^{\mathrm{da}} = \operatorname*{arg\,min}_{\theta\in\mathbb{R}^d} \Big\{ \frac{1}{2\nu}\|\theta-\theta_*\|_2^2 + \frac{1}{2\gamma n}\|\theta\|_2^2 \Big\}.$$

We compute $\theta_n^{\mathrm{da}} = \frac{\gamma\nu n}{\gamma\nu n+1}\theta_*$ and

$$g(\theta_n^{\mathrm{da}}) - g(\theta_*) = \frac{g(\theta_0^{\mathrm{da}}) - g(\theta_*)}{(1+\nu\gamma n)^2}.$$

4.H Continuous Time Interpretation of DA et MD

Following Nemirovsky and Yudin [1983], Bolte and Teboulle [2003], Krichene et al. [2015] we propose a continuous interpretation of these methods for g twice differentiable. We note this could be extended for g non-smooth with differential inclusions.

Derivation of the ordinary differential equation (ODE). The first-order optimality condition of the MD iteration in Eq. (4.4) $\gamma\nabla f(\theta_n) + \gamma\nabla g(\theta_{n+1}) + \nabla h(\theta_{n+1}) - \nabla h(\theta_n)$ can be rearranged as

$$\frac{\nabla h(\theta_{n+1}) - \nabla h(\theta_n)}{\gamma} = -\nabla f(\theta_n) - \nabla g(\theta_{n+1}).$$

Noting $\partial_t \nabla h(\theta) = \nabla^2 h(\theta)\dot\theta$, this is exactly a forward-backward Euler discretization of the MD ODE [Bolte and Teboulle, 2003]:

$$\dot\theta = -\nabla^2 h(\theta)^{-1}[\nabla f(\theta) + \nabla g(\theta)]. \tag{4.25}$$

On the other hand, considering the DA iteration in Eq. (4.3) we obtain

$$\frac{\eta_n - \eta_{n-1}}{\gamma} = -\nabla f(\theta_{n-1}) \qquad \text{and} \qquad \eta_n = n\gamma\nabla g(\theta_n) + \nabla h(\theta_n). \tag{4.26}$$

Combining both parts in Eq. (4.26) leads to the single equation

$$n\gamma\frac{\nabla g(\theta_n) - \nabla g(\theta_{n-1})}{\gamma} + \nabla g(\theta_{n-1}) + \frac{\nabla h(\theta_n) - \nabla h(\theta_{n-1})}{\gamma} = -\nabla f(\theta_{n-1}),$$

which is the explicit Euler discretization of the ODE $\partial_t(t\nabla g(\theta) + \nabla h(\theta)) = -\nabla f(\theta)$. Therefore the ODE associated to DA takes the form

$$\dot\theta = -\nabla^2(h(\theta) + tg(\theta))^{-1}(\nabla f(\theta) + \nabla g(\theta)). \tag{4.27}$$

It is worth noting that this ODE is very similar to the MD ODE in Eq. (4.25), with an additional term $tg(\theta)$ in the inverse mapping $\nabla^2(h(\theta) + tg(\theta))^{-1}$ which may thus slow down the DA dynamic.

Lyapunov analyzes. Lyapunov functions are used to prove convergence of the solutions of ODEs. In analogy with the discrete case, the Bregman divergence is a Lyapunov function for these ODEs [see, e.g., Bolte and Teboulle, 2003, Krichene et al., 2015] since

$$\begin{aligned}\partial_t D_h(\theta_*, \theta(t)) &= \partial_t[h(\theta_*) - h(\theta(t)) - \langle \nabla h(\theta(t)), \theta_* - \theta(t)\rangle] \\ &= -\langle \nabla h(\theta(t)), \dot\theta(t)\rangle + \langle \nabla^2 h(\theta(t))\dot\theta(t), \theta(t) - \theta_*\rangle + \langle \nabla h(\theta(t)), \dot\theta(t)\rangle \\ &= \langle \nabla^2 h(\theta(t))\dot\theta(t), \theta(t) - \theta_*\rangle.\end{aligned}$$

For the MD ODE in Eq. (4.25) we obtain

$$\begin{aligned}\partial_t D_h(\theta_*, \theta(t)) &= -\langle \nabla f(\theta(t)) + \nabla g(\theta(t)), \theta(t) - \theta_*\rangle \\ &\leq \psi(\theta_*) - \psi(\theta(t)) \qquad \text{(by convexity of } \psi).\end{aligned}$$

Integrating, this yields with Jensen inequality

$$\psi(\bar\theta(t)) - \psi(\theta_*) \leq \frac{1}{t}\int_0^t \big(\psi(\theta(s)) - \psi(\theta_*)\big) ds \leq \frac{D_h(\theta_*, \theta(0)) - D_h(\theta_*, \theta(t))}{t},$$

for $\bar\theta(t) = \frac{1}{t}\int_0^t \theta(s) ds$. This is the same convergence result as in the discrete time. For the DA ODE in Eq. (4.27) we obtain

$$\begin{aligned}\partial_t D_{h+tg}(\theta_*, \theta(t)) &= \partial_t[(h+tg)(\theta_*) - (h+tg)(\theta(t)) - \langle \nabla(h+tg)(\theta(t)), \theta_* - \theta(t)\rangle] \\ &= g(\theta_*) - \langle (\nabla h(\theta(t)) + t\nabla g(\theta(t))), \dot\theta(t)\rangle + g(\theta(t)) \\ &\quad + \langle \partial_t(\nabla h + t\nabla g)(\theta(t)), \theta(t) - \theta_*\rangle + \langle (\nabla + t\nabla g) h(\theta(t)), \dot\theta(t)\rangle \\ &= g(\theta_*) - g(\theta_t) - \langle \nabla f(\theta(t)), \theta(t) - \theta_*\rangle.\end{aligned}$$

Therefore by convexity of f, $\partial_t D_{h+tg}(\theta_*, \theta(t)) \leq \psi(\theta_*) - \psi(\theta(t))$ and we obtain

$$\psi(\bar\theta(t)) - \psi(\theta_*) \leq \frac{D_h(\theta_*, \theta(0)) - D_{h+tg}(\theta_*, \theta(t))}{t}.$$

The continuous time argument really mimics the proof of Proposition 8 without the technicalities associated with the discrete time. We remind that we recover the variational interpretation of Krichene et al. [2015], Wibisono et al. [2016], Wilson et al. [2016]: the Lyapunov function generates the dynamic in the sense that a function L is first chosen and secondly a dynamics, for which L is a Lyapunov function, is then designed. In this way MD and DA are the two different dynamics associated to the two different Lyapunov functions D_h and D_{h+tg}.

Extension to the noisy-gradient case. We consider now we only have access to noisy estimates of the gradient as in Section 4.3 and propose a continuous-time interpretation of these stochastic methods. Stochastic MD and SDA may be viewed, in their primal-dual forms, as discretizations of the following stochastic differential

equations (SDE). For stochastic MD

$$\mathrm{d}\eta(t) = -[\nabla f(\theta(t)) + \nabla g(\theta(t))]\mathrm{d}t + \sigma \mathrm{d}W(t)\mathrm{d}t \qquad \text{and} \qquad \eta(t) = \nabla h(\theta(t)),$$

and for SDA

$$\mathrm{d}\eta(t) = -\nabla f(\theta(t))\mathrm{d}t + \sigma \mathrm{d}W(t)\mathrm{d}t \qquad \text{and} \qquad \eta(t) = \nabla (h + tg)(\theta(t)),$$

where W_t is a Wiener process and $\sigma > 0$. We note that the regularization g does not take part in the SDA SDE which explains this dynamic is efficient in presence of noise. In contrast, the stochastic MD SDE is corrupted by the presence of the gradient ∇g which may not behaves well for non-smooth g. This continuous-time interpretation of stochastic algorithms could lead to further insights but is outside the scope of this chapter.

4.I Examples of Different Geometries

We describe now different examples of concrete geometries and how SDA is then implemented for well known regularizations g.

Euclidean distance. The simplest geometry is obtained by taking the function $h(\theta) = \frac{1}{2}\|\theta\|_2^2$, which is a Legendre function on $\operatorname{dom} h = \mathbb{R}^d$. Its associated Bregman divergence is also the squared Euclidean distance $D_h(\alpha, \beta) = \frac{1}{2}\|\alpha - \beta\|_2^2$. Therefore **(LC)** is equivalent to the smoothness of the function f and we return to classic results on proximal gradient descent.

— Projection: Let $g = \mathbb{1}_{\mathcal{C}}$ be the indicator of a convex set $\mathcal{C}$. The SDA method yields to the projected method

$$\theta_n = \min_{\theta \in \mathcal{C}} \Big\| \theta + \gamma \sum_{k=0}^{n-1} \nabla f_{k+1}(\theta_k) \Big\|_2^2.$$

— ℓ_2-regularization: Let $g = \frac{1}{2}\|\cdot\|_Q^2$ where $Q \succcurlyeq 0$, we directly have $\nabla h_n^*(\eta) = (I + n\gamma Q)^{-1}\eta$ and the SDA method comes back to

$$\theta_n = \theta_{n-1} - (\gamma^{-1} I + nQ)^{-1}(Q\theta_{n-1} + \nabla f_n(\theta_{n-1})), \text{ for } n \geq 1,$$

which is a standard gradient descent on $f + g$ with a structured decreasing step-size $\gamma_n = (\gamma^{-1} I + nQ)^{-1}$.

— ℓ_1-regularization: Let $g = \lambda\|\cdot\|_1$, we can compute the primal iterate with, for $i = 1, \ldots, d$, $\nabla_i h_n^*(\eta) = \operatorname{sign}(\eta(i)) \max(|\eta(i)| - n\gamma\lambda, 0)$. Therefore the SDA method is equivalent, for $i = 1, \ldots, d$, to the iteration:

$$\theta_n(i) = -\operatorname{sign}\Big(\sum_{k=0}^{n-1} \nabla_i f_{k+1}(\theta_k)\Big) \max\Big(\Big|\sum_{k=0}^{n-1} \nabla_i f_{k+1}(\theta_k)\Big| - n\gamma\lambda, 0\Big).$$

Yet since convergence results hold on the average of the iterates $\bar{\theta}_n$, SDA provides less sparse solutions than other methods which rather consider final iterates as outputs.

Kullback-Leibler divergence. The negative entropy $h(\theta) = \sum_{i=1}^n \theta(i)\log(\theta(i))$ is a Legendre function on dom $h = (0,\infty)^n$ whose associated Bregman divergence is the Kullback-Leibler divergence

$$D_h(\alpha,\beta) = \sum_{i=1}^n \alpha(i)\log\Big(\frac{\alpha(i)}{\beta(i)}\Big) + \sum_{i=1}^n (\beta(i)-\alpha(i)),$$

and its conjugate gradient mapping is $\nabla_i h^*(\eta) = \exp(\eta_i) - 1$ for $i = 1,\dots,d$.

When constrained on the simplex Δ_d, h is 1-strongly convex with respect to the ℓ_1-norm [see, e.g., Beck and Teboulle, 2003, Proposition 5.1], and **(LC)** holds, for example, if f is smooth with regards to the ℓ_1-norm. This illustrates one of the non-Euclidean benefit since Lipschitz constants under the ℓ_∞-norm are smaller than under the ℓ_2-norm.

This geometry is particularly appropriated to constrained minimization on the simplex Δ_d. With $g(\theta) = \mathbb{1}_{\Delta_d}$, SDA update is the dual averaging analogue of the exponentiated gradient algorithm [Kivinen and Warmuth, 1997]:

$$\theta_n(i) = \frac{\exp(\eta_n(i))}{\sum_{j=1}^d \exp(\eta_n(j))} \text{ for } i = 1,\dots,d.$$

ℓ_p-norm. The choice $h = \frac{1}{2(p-1)}\|\cdot\|_p^2$ for $p \in (1,2]$ is believed to adapt to the geometry of learning problem and is often used with $p = 1 + 1/\log(d)$ in association with ℓ_1-regularization [see, e.g., Duchi et al., 2010]. Its Fenchel conjugate is the squared conjugate norm $h^* = \frac{1}{2(q-1)}\|\cdot\|_q^2$ for $1/p + 1/q = 1$ and its conjugate gradient mapping is $\nabla_i h^*(\eta) = \frac{\operatorname{sign}(\eta(i))|\eta(i)|^{q-1}}{(q-1)\|\eta\|_q^{q-2}}$ [see, e.g., Gentile and Littlestone, 1999]. For ℓ_1-regularization, this yields to:

$$\nabla_i h_n^*(\eta) = \nabla_i h^*\big(\operatorname{sign}(\eta(i))\max(|\eta(i)| - n\gamma\lambda, 0)\big) \text{ for } i = 1,\dots,d.$$

The function h is 1-strongly convex with respect to the ℓ_p-norm [see, e.g., Hanner, 1956]. Therefore **(LC)** holds if f is smooth with respect to the the ℓ_p-norm. However when the function f considered is quadratic as in Section 4.3, we can directly show that **(LC)** holds under tighter conditions on the Hessian matrix Σ (see proof in Appendix 4.J).

Proposition 17. *Assume that $f(\theta) = \frac{1}{2}\langle\theta,\Sigma\theta\rangle$ and $h(\theta) = \frac{1}{2(p-1)}\|\theta\|_p^2$. Then $h - \gamma f$ is convex for any constant step-size γ such that*

$$\gamma \le \min_\alpha \frac{\|\alpha\|_p^2}{\langle\alpha,\Sigma\alpha\rangle}.$$

When $\Sigma = \mathbb{E}(x \otimes x)$ is a covariance matrix as in Section 4.3.2, $\langle \alpha, \Sigma\alpha \rangle = \mathbb{E}\langle x, \alpha \rangle^2 \leq \mathbb{E}\|x\|_q^2 \|\alpha\|_p^2$ by Hölder inequality, and Proposition 17 admits the following corollary.

Corollary 7. *Assume that* $f(\theta) = \frac{1}{2}\mathbb{E}(\langle x, \theta \rangle - y)^2$, $h(\theta) = \frac{1}{2}\|\theta\|_p^2$ *and* q *such that* $1/p + 1/q = 1$. *Then* $h - \gamma f$ *is convex for any constant step-size* γ *such that*

$$\gamma \leq 1/\mathbb{E}\|x\|_q^2.$$

Therefore we may use the algorithm with bigger step-size than in the Euclidean case. Moreover when the algorithm is started from $\theta_0 = 0$, the Bregman divergence is $D_h(\theta_*, \theta_0) = \frac{1}{2(p-1)}\|\theta_*\|_p^2$ and the bias in Proposition 8 would be bounded by $\frac{\mathbb{E}\|x\|_q^2 \|\theta_*\|_p^2}{2(p-1)}$.

For high-dimension problems, taking $q = 1 + \log(d)$ (with $p \sim 1$ and $q \sim +\infty$) yields to bounds depending on the ℓ_1-norm of the optimal predictor and the ℓ_∞-norm of the features which is advisable for sparse problems.

4.J Proof of Proposition 17

We consider here $h(\theta) = \frac{1}{2(p-1)}\|\theta\|_p^2$. For $\theta \in \mathbb{R}^d$, h is twice differentiable. Its gradient is

$$\nabla_i h(\theta) = \frac{\operatorname{sign}(\theta(i))|\theta(i)|^{p-1}}{(p-1)\|\theta\|_p^{p-2}},$$

and its Hessian may be written for $\alpha = \frac{2-p}{(p-1)}\|\theta\|_p^{-2(p-1)}$, $u(i) = \|\theta\|_p^{2-p}\theta(i)^{p-2}$ and $v(i) = \theta(i)^{p-1}$ for $i = 1, \dots, d$, as

$$\nabla^2 h(\theta) = \operatorname{Diag}(u) + \alpha v v^\top,$$

The function $h - \gamma f$ is convex if and only if $\nabla^2 h(\theta) \preccurlyeq \gamma\Sigma$ for all $\theta \in \mathbb{R}^d$. This condition is equivalent to

$$\min_\theta \min_\alpha \frac{\langle \alpha, \nabla^2 h(\theta)\alpha \rangle}{\langle \alpha, \Sigma\alpha \rangle} \geq \gamma.$$

A sufficient condition is that $\operatorname{Diag} u \succcurlyeq \gamma\Sigma$. After a change of variables, u may be written as $u(i) = \eta(i)^{p-2}$ where $\eta(i) = |\theta(i)|/\|\theta\|_p$ satisfies $\sum_{i=1}^d \eta(i)^p = 1$ and $\eta(i) \geq 0$. Hence for all $\theta, \alpha \in \mathbb{R}^d$

$$\langle \alpha, \nabla^2 h(\theta)\alpha \rangle \geq \sum_{i=1}^d \alpha(i)^2 u(i) = \sum_{i=1}^d \alpha(i)^2 \eta(i)^{p-2},$$

which implies

$$\min_{\theta \in \mathbb{R}^d} \langle \alpha, \nabla^2 h(\theta)\alpha \rangle \geq \min_{\eta \in \mathbb{R}^d} \sum_{i=1}^d \alpha(i)^2 \eta(i)^{p-2} \text{ such that } \sum_{i=1}^d \eta(i)^p = 1 \text{ and } \eta(i) \geq 0.$$

This optimization problem is equivalent with $v(i) = \eta(i)^p$ to the one the simplex Δ_d

$$\min_{v \in \mathbb{R}^d} \sum_{i=1}^d \alpha(i)^2 v(i)^{1-2/p} \text{ such that } \sum_{i=1}^d v(i) = 1 \text{ and } \nu(i) \geq 0,$$

for which we define the Lagrangian $\mathcal{L}(v, \lambda, \mu) = \sum_{i=1}^d \alpha(i)^2 v(i)^{1-2/p} - \langle \lambda, v \rangle + \nu(1 - \sum_{i=1}^d v(i))$ for $\lambda \in \mathbb{R}_+^d$ and $\mu \in \mathbb{R}$. Its gradient is $\nabla_{v(i)} \mathcal{L}(v, \lambda, \mu) = (1-2/p)\alpha(i)^2/v(i)^{2/p} - \lambda(i) - \nu$. Writing the KKT condition for this problem [see, e.g., Boyd and Vandenberghe, 2004], we have that (v, λ, ν) is optimal if and only if $(1 - 2/p)\alpha(i)^2/v(i)^{2/p} - \lambda(i) - \nu = 0$, $\sum_{i=1}^d v(i) = 1$ and for all i; $\lambda(i) \geq 0$, $v(i) \geq 0$ and $\lambda(i)v(i) = 0$. These conditions are satisfied by $v(i) = \frac{\alpha(i)^p}{\sum_{i=1}^d a(j)^p}$, $\alpha(i) = 0$ and $\nu = (1 - 2p)(\sum_{i=1}^d a(j)^p)^{2/p}$. Hence the minimum value is

$$\sum_{i=1}^d \alpha(i)^2 v(i)^{1-2/p} = \sum_{i=1}^d \alpha(i)^2 \frac{\alpha(i)^{p-2}}{(\sum_{i=1}^d a(j)^p)^{1-2/p}} = \frac{\sum_{i=1}^d a(j)}{(\sum_{i=1}^d a(j)^p)^{1-2/p}} = \|\alpha\|_p^2.$$

Consequently

$$\langle \alpha, \nabla^2 h(\theta) \alpha \rangle \geq \|\alpha\|_p^2,$$

and $h - \gamma f$ is convex for $\gamma \leq \min_{\alpha \in \mathbb{R}^d} \frac{\|\alpha\|_p^2}{\langle \alpha, \Sigma \alpha \rangle}$.

4.K Standard Benchmarks

We have considered the *sido* dataset which is often used for comparing large-scale optimization algorithms. This is a *finite* binary classification dataset with finite number of observations with outputs in $\{-1, 1\}$. We have followed the following experimental protocol: (1) remove all outliers, i.e., sample points x_n whose norms is greater than 5 times the average norm. (2) divide the dataset in two equal parts, one for training, one for testing, (3) start the algorithms from $\theta_0 = 0$, (4) sample within the training dataset with replacement, for 100 times the number of observations in the training set; a dashed line marks the first effective pass in all plots, (5) compute averaged cost on training and testing data based on 10 replications. All cost are shown in log-scale, normalized to that the first iteration leads to $\psi(\theta_0) - \psi(\theta_*) = 1$.

We solved a ℓ_1-regularized least-squares regression for three different values of ℓ_1-regularization: (1) one with the λ_* which corresponds to the best generalization error after 500 effective passes through the train set, (2) one with $\lambda_*/8$ and (3) one with $256\lambda_*$.

We compare five algorithms: averaged SGD with constant step-size, average SGD with decreasing step-size $C/(R^2\sqrt{n})$, SDA with constant step-size, SDA with decreasing step-size $C/(R^2\sqrt{n})$ and SAGA with constant step-size [Defazio et al., 2014], which showed state-of-the-art performance in the set-up of finite data sets. We consider the theoretical value of step-size which ensures convergence. We note the behaviors are comparable to the situation where step-sizes with the best testing error after one effective pass through the data (testing powers of 4 times the theoretical step-size)

are used.

We can make the following observations:

— We show results for $\lambda = \lambda_*$ in Figure 4.K.1. SAGA, constant-step-size SDA and constant-step-size SGD exhibit the best behavior for both settings of step-size. However the training error of SGD does not converge to 0. On the other hand, SGD and SDA with step-size decaying as $C/R^2\sqrt{n}$ are slower. SAGA and constant-step-size SDA exhibit some overfitting after more than 10 passes on the regularized objective ψ.

— We show results for $\lambda = \lambda_*/8$ in Figure 4.K.2. The problem is then very little regularized and the behavior of constant-step-size SGD gets closer to constant-step-size SDA. There is here still overfitting for the regularized objective ψ.

— We show results for $\lambda = 256\lambda_*$ in Figure 4.K.1. The problem is then much more regularized. In this case the regularization has an important weight and the stochasticity of the quadratic objective plays a minor role. Therefore SAGA exhibits the best behavior, despite strong early oscillations, with a linear convergence but reaches a saturation point after few passes over the data. On the other hand, constant-step-size SDA exhibits a sublinear convergence which is faster at the beginning and catches up with SAGA at the end. Constant-step-size SGD is not converging to the solution.

To conclude, constant-step-size SDA behaves similarly to SAGA which is specially dedicated to the set-up of finite data sets. For larger datasets, where only a single pass is possible, SAGA could not be run. Moreover SAGA does not come with generalization guarantees while SDA does (if a single pass is made).

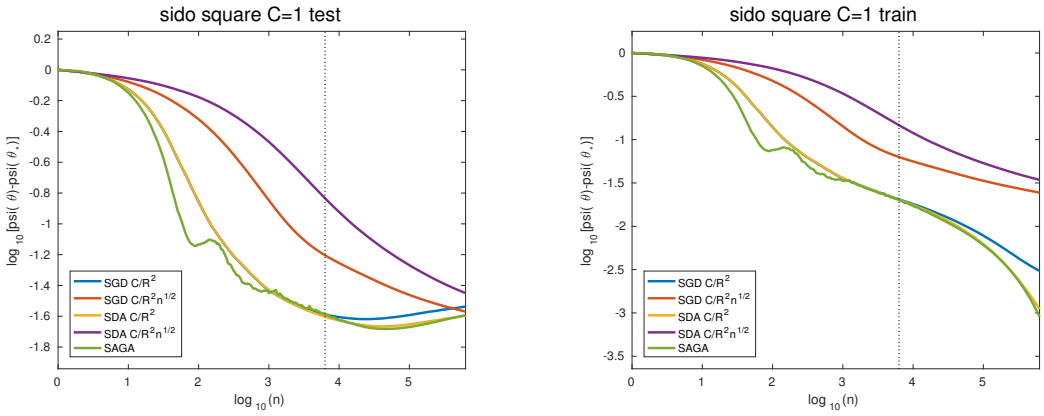


Figure 4.K.1 – Test and train performances for ℓ_1-regularized least-squares regression on the *sido* dataset with $\lambda = \lambda_{\text{opt}}$. Left: test performance. Right: train performance.

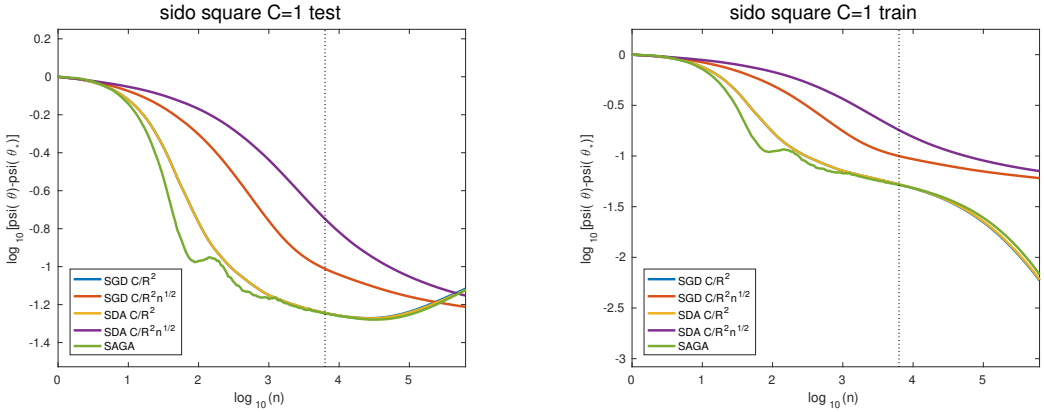


Figure 4.K.2 – Test and train performances for ℓ_1-regularized least-squares regression on the *sido* dataset with $\lambda = \frac{\lambda_{\text{opt}}}{8}$. Left: test performance. Right: train performance.

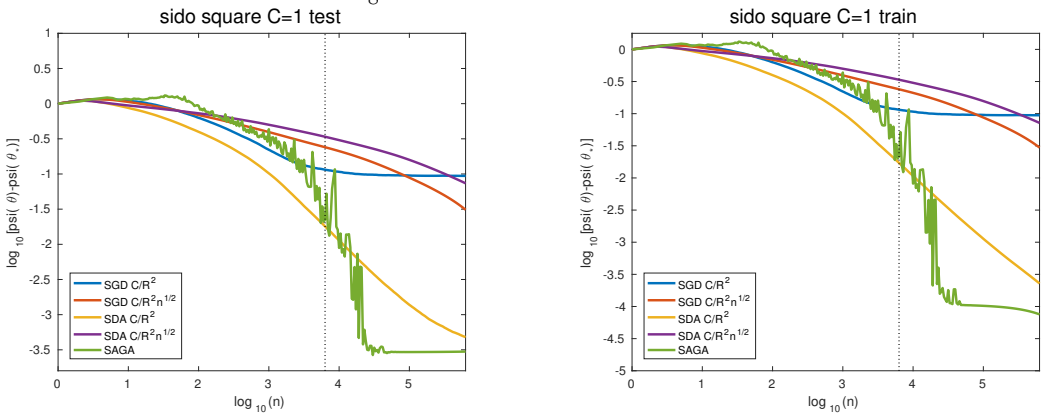


Figure 4.K.3 – Test and train performances for ℓ_1-regularized least-squares regression on the *sido* dataset with $\lambda = 256\lambda_{\text{opt}}$. Left: test performance. Right: train performance.

Part II

Applications of the Quadratic Loss in Machine Learning

Chapter 5

Application to Discriminative Clustering

Abstract

Clustering high-dimensional data often requires some form of dimensionality reduction, where clustered variables are separated from "noise-looking" variables. We cast this problem as finding a low-dimensional projection of the data which is well-clustered. This yields a one-dimensional projection in the simplest situation with two clusters, and extends naturally to a multi-label scenario for more than two clusters. In this chapter, (a) we first show that this joint clustering and dimension reduction formulation is equivalent to previously proposed discriminative clustering frameworks, thus leading to convex relaxations of the problem; (b) we propose a novel sparse extension, which is still cast as a convex relaxation and allows estimation in higher dimensions; (c) we propose a natural extension for the multi-label scenario; (d) we provide a new theoretical analysis of the performance of these formulations with a simple probabilistic model, leading to scalings over the form $d = O(\sqrt{n})$ for the affine invariant case and $d = O(n)$ for the sparse case, where n is the number of examples and d the ambient dimension; and finally, (e) we propose an efficient iterative algorithm with running-time complexity proportional to $O(nd^2)$, improving on earlier algorithms for discriminative clustering with the square loss, which had quadratic complexity in the number of examples.

This chapter is extracted from the paper Robust Discriminative Clustering with Sparse Regularizers, in collaboration with B. Palaniappan and F. Bach, accepted in the *Journal of Machine Learning Research*.

5.1 Introduction

Clustering is an important and commonly used pre-processing tool in many machine learning applications, with classical algorithms such as K-means [MacQueen, 1967], linkage algorithms [Gower and Ross, 1969] or spectral clustering [Ng et al., 2002]. In high dimensions, these unsupervised learning algorithms typically have

problems identifying the underlying optimal discrete nature of the data; for example, they are quickly perturbed by adding a few noisy dimensions. Clustering high-dimensional data thus requires some form of dimensionality reduction, where clustered variables are separated from non-informative "noise-looking" (e.g., Gaussian) variables.

Several frameworks aim at linearly separating noise from signal, that is finding projections of the data that extracts the signal and removes the noise. They differ in the ways signals and noise are defined. A line of work that dates back to projection pursuit [Friedman and Stuetzle, 1981] and independent component analysis [Hyvärinen et al., 2004] defines the noise as Gaussian while the signal is non-Gaussian [Blanchard et al., 2006, Le Roux and Bach, 2013, Diederichs et al., 2013]. In this work, we follow De la Torre and Kanade [2006], Ding and Li [2007], along the alternative route where one defines the signal as being clustered while the noise is any non-clustered variable. In the simplest situation with two clusters, we may project the data into a one-dimensional subspace. Given a data matrix $X \in \mathbb{R}^{n \times d}$ composed of n d-dimensional points, the goal is to find a direction $w \in \mathbb{R}^d$ such that $Xw \in \mathbb{R}^n$ is well-clustered, e.g., by K-means. This is equivalent to identifying both a direction to project, represented as $w \in \mathbb{R}^d$ and the labeling $y \in \{-1, 1\}^n$ that represents the partition into two clusters.

Most existing formulations are non-convex and typically perform a form of alternating optimization [De la Torre and Kanade, 2006, Ding and Li, 2007], where given $y \in \{-1, 1\}^n$, the projection w is found by linear discriminant analysis (or any binary classification method), and given the projection w, the clustering is obtained by thresholding Xw or running K-means on Xw. As shown in Section 5.2, this alternating minimization procedure happens to be equivalent to maximizing the (centered) correlation between $y \in \{-1, 1\}^n$ and the projection $Xw \in \mathbb{R}^d$, that is

$$\max_{w \in \mathbb{R}^d, y \in \{-1,1\}^n} \frac{(y^\top \Pi_n X w)^2}{\|\Pi_n y\|_2^2 \, \|\Pi_n X w\|_2^2},$$

where $\Pi_n = I_n - \frac{1}{n} 1_n 1_n^\top$ is the usual centering projection matrix (with $1_n \in \mathbb{R}^n$ being the vector of all ones, and I_n the $n \times n$ identity matrix). This correlation is equal to one when the projection is perfectly clustered (independently of the number of elements per cluster). Existing methods are alternating minimization algorithms with no theoretical guarantees.

In this chapter, we relate this formulation to discriminative clustering formulations [Xu et al., 2004, Bach and Harchaoui, 2007], which consider the problem

$$\min_{v \in \mathbb{R}^d, \; b \in \mathbb{R}, \; y \in \{-1,1\}^n} \frac{1}{n} \|y - Xv - b 1_n\|_2^2, \tag{5.1}$$

with the intuition of finding labels y which are easy to predict by an affine function of the data. In particular, we show that given the relationship between the number of positive labels and negative labels (i.e., the squared difference between the respective number of elements), these two problems are equivalent, and hence discriminative

clustering explicitly performs joint dimension reduction and clustering.

While the discriminative framework is based on convex relaxations and has led to interesting developments and applications [Zhang et al., 2009, Li et al., 2009, Joulin et al., 2010a,b, Wang et al., 2010, Niu et al., 2013, Huang et al., 2015], it has several shortcomings when used with the square loss: (a) the running-time complexity of the semi-definite formulations is at least quadratic in n, and typically much more, (b) no theoretical analysis has ever been performed, (c) no convex sparse extension has been proposed to handle data with many irrelevant dimensions, (d) balancing of the clusters remains an issue, as it typically adds an extra hyperparameter which may be hard to set. In this chapter, we focus on addressing these concerns.

When there are more than two clusters, one considers either the *multi-label* or the *multi-class* settings. The multi-class problem assumes that the data are clustered into distinct classes, i.e., a single class per observation, whereas the multi-label problem assumes the data share different labels, i.e., multiple labels per observation. We show in this work that discriminative clustering framework extends more naturally to multi-label scenarios and that this extension has the same convex relaxation.

A summary of the contributions of this chapter follows:

— In Section 5.2, we relate discriminative clustering with the square loss to a joint clustering and dimension reduction formulation. The proposed formulation takes care of the balancing hyperparameter implicitly.

— We propose in Section 5.3 a novel sparse extension to discriminative clustering and show that it can still be cast through a convex relaxation.

— When there are more than two clusters, we extend naturally the sparse formulation to a multi-label scenario in Section 5.4.

— We then proceed to provide a theoretical analysis of the proposed formulations with a simple probabilistic model in Section 5.5, which effectively leads to scalings over the form $d = O(\sqrt{n})$ for the affine invariant case and $d = O(n)$ for the 1-sparse case.

— Finally, we propose in Section 5.6 efficient iterative algorithms with running-time complexity for each step equal to $O(nd^2)$, the first to be linear in the number of observations n for discriminative clustering with the square loss.

Throughout this chapter we assume that $X \in \mathbb{R}^{n \times d}$ is *centered*, a common pre-processing step in unsupervised (and supervised) learning. This implies that $X^\top 1_n = 0$ and $\Pi_n X = X$.

5.2 Joint Dimension Reduction and Clustering

In this section, we focus on the single binary label case, where we first study the usual non-convex formulation, before deriving convex relaxations based on semi-definite programming. Some of the following results are already known in the literature however we state them here for completeness.

5.2.1 Non-Convex Formulation

Following De la Torre and Kanade [2006], Ding and Li [2007], Ye et al. [2008], we consider a cost function which depends on $y \in \{-1,1\}^n$ and $w \in \mathbb{R}^d$, which is such that alternating optimization is exactly (a) running K-means with two clusters on Xw to obtain y given w (when we say "running K-means", we mean solving the vector quantization problem exactly), and (b) performing linear discriminant analysis to obtain w given y.

Proposition 18 (Joint clustering and dimension reduction for two clusters)**.** *Given $X \in \mathbb{R}^{n\times d}$ such that $X^\top 1_n = 0$ and X has rank d, consider the optimization problem*

$$\max_{w\in\mathbb{R}^d, y\in\{-1,1\}^n} \frac{(y^\top Xw)^2}{\|\Pi_n y\|_2^2 \, \|Xw\|_2^2}. \tag{5.2}$$

Given y, the optimal w is obtained as $w = (X^\top X)^{-1}X^\top y$, while given w, the optimal y is obtained by running K-means on Xw.

This equivalence might be straightforward however it has not been precisely stated in the literature to the best of our knowledge.

Proof. Given y, we need to optimize the Rayleigh quotient $\frac{w^\top X^\top y y^\top Xw}{w^\top X^\top Xw}$ with a rank-one matrix in the numerator, which leads to $w = (X^\top X)^{-1}X^\top y$. Given w, we show in Appendix 5.A, that the averaged distortion measure of K-means once the means have been optimized is exactly equal to $(y^\top Xw)^2/\|\Pi_n y\|_2^2$. □

Algorithm. The proposition above leads to an alternating optimization algorithm. Note that K-means in one dimension may be run *exactly* in $O(n \log n)$ [Bellman, 1973]. After having optimized with respect to w in Eq. (5.2), we then need to maximize with respect to y the function $\frac{y^\top X(X^\top X)^{-1}X^\top y}{\|\Pi_n y\|_2^2}$, which happens to be exactly performing K-means on the whitened data (which is now in high dimension and not in 1 dimension). At first, it seems that dimension reduction is *simply* equivalent to whitening the data and performing K-means; while this is a formally correct statement, the resulting K-means problem is not easy to solve as the clustered dimension is hidden in noise; for example, algorithms such as K-means++ [Arthur and Vassilvitskii, 2007], which have a multiplicative theoretical guarantee on the final distortion measure, are not provably effective here because the minimal final distortion is not small (since the clusters are corrupted by some noisy dimensions), and the multiplicative guarantee is then meaningless.

5.2.2 Convex Relaxation and Discriminative Clustering

The discriminative clustering formulation in Eq. (5.1) may be optimized for any $y \in \{-1,1\}^n$ in closed form with respect to b as $b = \frac{1_n^\top (y - Xv)}{n} = \frac{1_n^\top y}{n}$ since X is

centered. Substituting b in Eq. (5.1) leads us to

$$\min_{v\in\mathbb{R}^d} \frac{1}{n}\|\Pi_n y - Xv\|_2^2 = \frac{1}{n}\|\Pi_n y\|_2^2 - \max_{w\in\mathbb{R}^d} \frac{(y^\top Xw)^2}{\|Xw\|_2^2}, \tag{5.3}$$

where v is obtained from any solution w as $v = w\frac{y^\top Xw}{\|Xw\|_2^2}$. Thus, given

$$\frac{(y^\top 1_n)^2}{n^2} = \frac{1}{n^2}\big(\#\{i, y_i = 1\} - \#\{i, y_i = -1\}\big)^2 = \alpha \in [0, 1], \tag{5.4}$$

which characterizes the asymmetry between clusters and with $\|\Pi_n y\|^2 = n(1-\alpha)$, we obtain from Eq. (5.3), an equivalent formulation to Eq. (5.2) (with the added constraint) as

$$\min_{y\in\{-1,1\}^n,\ v\in\mathbb{R}^d} \frac{1}{n}\|\Pi_n y - Xv\|_2^2 \ \text{ such that } \ \frac{(y^\top 1_n)^2}{n^2} = \alpha. \tag{5.5}$$

This is exactly equivalent to a discriminative clustering formulation with the square loss [Bach and Harchaoui, 2007] with an explicit cluster balance constraint. Consequently we have formally established that the discriminative clustering formulation in Eq. (5.5) is related to the joint clustering and dimension reduction formulation in Eq. (5.2). Following Bach and Harchaoui [2007], we may optimize Eq. (5.5) in closed form with respect to v as $v = (X^\top X)^{-1}X^\top y$. Substituting v in Eq. (5.5) leads us to

$$\min_{y\in\{-1,1\}^n} \frac{1}{n} y^\top \big(\Pi_n - X(X^\top X)^{-1}X^\top\big) y \ \text{ such that } \ \frac{(y^\top 1_n)^2}{n^2} = \alpha. \tag{5.6}$$

This combinatorial optimization problem is NP-hard in general [Karp, 1972, Garey et al., 1976]. Hence in practice, it is classical to consider the following convex relaxation of Eq. (5.6) [Luo et al., 2010]. For any admissible $y \in \{-1, +1\}^n$, the matrix $Y = yy^\top \in \mathbb{R}^{n\times n}$ is a rank-one symmetric positive semi-definite matrix with unit diagonal entries and conversely any such Y may be written in the form $Y = yy^\top$ such that y is admissible for Eq. (5.6). Moreover by rewriting Eq. (5.6) as

$$\min_{y\in\{-1,1\}^n} \frac{1}{n} \operatorname{tr} yy^\top \big(\Pi_n - X(X^\top X)^{-1}X^\top\big) \ \text{ such that } \ \frac{1_n^\top (yy^\top) 1_n}{n^2} = \alpha,$$

we see that the objective and constraints are linear in the matrix $Y = yy^\top$ and Eq. (5.6) is equivalent to

$$\min_{Y\succcurlyeq 0,\ \operatorname{rank}(Y)=1\ \operatorname{diag}(Y)=1} \frac{1}{n} \operatorname{tr} Y\big(\Pi_n - X(X^\top X)^{-1}X^\top\big) \text{ such that } \frac{1_n^\top Y 1_n}{n^2} = \alpha.$$

Then dropping the non-convex rank constraint leads us to the following classical convex relaxation:

$$\min_{Y \succcurlyeq 0,\ \operatorname{diag}(Y)=1} \frac{1}{n} \operatorname{tr} Y\big(\Pi_n - X(X^\top X)^{-1}X^\top\big) \text{ such that } \frac{1_n^\top Y 1_n}{n^2} = \alpha. \tag{5.7}$$

This is the standard (unregularized) formulation, which is cast as a semi-definite program. The complexity of interior-point methods is $O(n^7)$, but efficient algorithms in $O(n^2)$ for such problems have been developed due to the relationship with the max-cut problem [Journée et al., 2010, Wen et al., 2012]. We note that convex relaxation techniques are also used for semi-supervised methods [De Bie and Cristianini, 2003].

Given the solution Y, one may traditionally obtain a candidate $y \in \{-1, 1\}^n$ by running K-means on the largest eigenvector of Y or by sampling [Goemans and Williamson, 1995]. In this chapter, we show in Section 5.5 that it may be advantageous to consider the first two eigenvectors.

5.2.3 Unsuccessful Full Convex Relaxation

The formulation in Eq. (5.7) imposes an extra parameter α that characterises the cluster imbalance. It is tempting to find a direct relaxation of Eq. (5.2). It turns out to lead to a trivial relaxation, which we outline below.

When optimizing Eq. (5.2) with respect to w, we obtain the following optimization problem

$$\max_{y \in \{-1,1\}^n} \frac{y^\top X(X^\top X)^{-1}X^\top y}{y^\top \Pi_n y},$$

leading to a quasi-convex relaxation as

$$\max_{Y \succcurlyeq 0,\ \operatorname{diag}(Y)=1} \frac{\operatorname{tr} YX(X^\top X)^{-1}X^\top}{\operatorname{tr} \Pi_n Y},$$

whose solution is found by solving a sequence of convex problems [Boyd and Vandenberghe, 2004, Section 4.2.5]. As shown in Appendix 5.B, this may be exactly reformulated as a single convex problem:

$$\max_{M \succcurlyeq 0,\ \operatorname{diag}(M)=1+\frac{1^\top M 1}{n^2}} \operatorname{tr} MX(X^\top X)^{-1}X^\top.$$

Unfortunately, this relaxation always leads to trivial solutions, and we thus need to consider the relaxation in Eq. (5.7) for several values of $\alpha = 1_n^\top Y 1_n / n^2$ (and then the non-convex algorithm can be run from the rounded solution of the convex problem, using Eq. (5.2) as a final objective). Alternatively, we may solve the following *penalized* problem for several values of $\nu \geqslant 0$:

$$\min_{Y \succcurlyeq 0,\ \operatorname{diag}(Y)=1} \frac{1}{n} \operatorname{tr} Y\big(\Pi_n - X(X^\top X)^{-1}X^\top\big) + \frac{\nu}{n^2} 1_n^\top Y 1_n. \tag{5.8}$$

For $\nu = 0$, $Y = 1_n 1_n^\top$ is always a trivial solution. As outlined in our theoretical section and as observed in our experiments, it is sufficient to consider $\nu \in [0, 1]$.

By convex duality [Borwein and Lewis, 2000, Sec. 4.3], both constrained relaxation in Eq. (5.7) and penalized relaxation in Eq. (5.8) are formally equivalent for specific choice of constraint parameter α and penalization parameter ν. We will see in Section 5.6 that the formulation in Eq. (5.8) is more suitable for algorithmic design [Bach et al., 2012].

5.2.4 Equivalent Relaxations

Optimizing Eq. (5.5) with respect to v in closed form as in Section 5.2.2 is feasible with no regularizer or with a quadratic regularizer. However, if one needs to add more complex regularizers, we need a different relaxation. Therefore we now propose a new formulation of the discriminative clustering framework. We start from the penalized version of Eq. (5.5),

$$\min_{y \in \{-1,1\}^n,\ v \in \mathbb{R}^d} \frac{1}{n} \|\Pi_n y - Xv\|_2^2 + \nu \frac{(y^\top 1_n)^2}{n^2}, \tag{5.9}$$

which we expand as:

$$\min_{y \in \{-1,1\}^n,\ v \in \mathbb{R}^d} \frac{1}{n} \operatorname{tr} \Pi_n y y^\top - \frac{2}{n} \operatorname{tr} X v y^\top + \frac{1}{n} \operatorname{tr} X^\top X v v^\top + \nu \frac{(y^\top 1_n)^2}{n^2}, \tag{5.10}$$

and relax as, using $Y = yy^\top$, $P = yv^\top$ and $V = vv^\top$,

$$\begin{aligned} \min_{V,P,Y} \quad & \frac{1}{n} \operatorname{tr} \Pi_n Y - \frac{2}{n} \operatorname{tr} P^\top X + \frac{1}{n} \operatorname{tr} X^\top X V + \nu \frac{1_n^\top Y 1_n}{n^2} \\ \text{s.t.} \quad & \begin{pmatrix} Y & P \\ P^\top & V \end{pmatrix} \succcurlyeq 0,\ \operatorname{diag}(Y) = 1. \end{aligned} \tag{5.11}$$

When optimizing Eq. (5.11) with respect to V and P, we get exactly Eq. (5.8). Indeed, the optimum is attained for $V{=}(X^\top X)^{-1} X^\top Y X (X^\top X)^{-1}$ and $P{=}YX(X^\top X)^{-1}$ as shown in Appendix 5.C.1. Therefore, the convex relaxation in Eq. (5.11) is equivalent to Eq. (5.8).

However, we get an interesting behavior when optimizing Eq. (5.11) with respect to P and Y also in closed form. For $\nu = 1$, we obtain, as shown in Appendix 5.C.2, the following closed form expressions:

$$\begin{aligned} Y &= \operatorname{Diag}(\operatorname{diag}(XVX^\top))^{-1/2} XVX^\top \operatorname{Diag}(\operatorname{diag}(XVX^\top))^{-1/2} \\ P &= \operatorname{Diag}(\operatorname{diag}(XVX^\top))^{-1/2} XV, \end{aligned}$$

leading to the problem:

$$\min_{V \succcurlyeq 0} \quad 1 - \frac{2}{n} \sum_{i=1}^n \sqrt{(XVX^\top)_{ii}} + \frac{1}{n} \operatorname{tr}(VX^\top X). \tag{5.12}$$

The formulation above in Eq. (5.12) is interesting for several reasons: (a) it is formulated as an optimization problem in $V \in \mathbb{R}^{d\times d}$, which will lead to algorithms whose running time will depend on n linearly (see Section 5.6), (b) it allows for easy adding of regularizers (see Section 5.3), which may be formulated as convex functions of $V = vv^\top$. At first sight this seems to be valid only for $\nu = 1$. However we now propose a reformulation which can handle all possible $\nu \in [0, 1)$ through a simple data augmentation.

Reformulation for any ν. When $\nu \in [0, 1)$, we may reformulate the objective function in Eq. (5.9) as follows:

$$\begin{aligned}
\|\Pi_n y - Xv\|_2^2 + \nu\frac{(y^\top 1_n)^2}{n} &= \|\Pi_n y - Xv + \nu\frac{y^\top 1_n}{n}1_n\|_2^2 - \frac{\big(\nu y^\top 1_n\big)^2}{n} + \nu\frac{\big(y^\top 1_n\big)^2}{n} \\
&= \|y - Xv - (1-\nu)\frac{y^\top 1_n}{n}1_n\|_2^2 + \frac{\nu\big((1-\nu)y^\top 1_n\big)^2}{(1-\nu)n} \\
&= \min_{b\in\mathbb{R}} \|y - Xv - b1_n\|_2^2 + \frac{\nu n}{1-\nu}b^2, \qquad (5.13)
\end{aligned}$$

since $\|y - Xv - b1_n\|_2^2 + \frac{\nu n}{1-\nu}b^2$ can be optimized in closed form with respect to b as $b = (1-\nu)\frac{y^\top 1_n}{n}$. Note that the weighted imbalance ratio $(1-\nu)\frac{y^\top 1_n}{n}$ is made as an optimization variable in Eq. (5.13). Thus we have the following reformulation

$$\begin{aligned}
&\min_{v\in\mathbb{R}^d,\ y\in\{-1,1\}^n} \frac{1}{n}\|\Pi_n y - Xv\|_2^2 + \nu\frac{(y^\top 1_n)^2}{n^2} \\
= &\min_{v\in\mathbb{R}^d, b\in\mathbb{R},\ y\in\{-1,1\}^n} \frac{1}{n}\|y - Xv - b1_n\|_2^2 + \frac{\nu}{1-\nu}b^2, \qquad (5.14)
\end{aligned}$$

which is a non-centered penalized formulation on a higher-dimensional problem in the variable $\binom{v}{b} \in \mathbb{R}^{d+1}$. In the rest of the chapter, we will focus on the case $\nu = 1$ for ease of exposition. This enables the use of the formulation in Eq. (5.12), which is easier to optimize. It is worth noting that this is not an algorithmic restriction. Of course any problem with $\nu \in [0, 1)$ can be treated with equal ease by adding a constant term and a quadratic regularizer.

5.3 Regularization

There are several natural possibilities. We consider norms Ω such that $\Omega(w)^2 = \Gamma(ww^\top)$ for a certain convex function Γ; all norms have that form [Bach et al., 2012, Proposition 5.1]. When $\nu = 1$, Eq. (5.12) then becomes

$$\max_{V\succcurlyeq 0} \frac{2}{n}\sum_{i=1}^n \sqrt{(XVX^\top)_{ii}} - \frac{1}{n}\operatorname{tr}(VX^\top X) - \Gamma(V). \qquad (5.15)$$

The quadratic regularizers $\Gamma(V) = \operatorname{tr}\Lambda V$ have already been tackled by Bach and Harchaoui [2007]. They consider the regularized version of problem in Eq. (5.3)

$$\min_{v\in\mathbb{R}^d} \frac{1}{n}\|\Pi_n y - Xv\|_2^2 + v^\top \Lambda v, \tag{5.16}$$

optimize in closed form with respect to v as $v = (X^\top X + n\Lambda)^{-1}X^\top y$. Substituting v in Eq. (5.16) leads them to

$$\min_{Y\succcurlyeq 0,\ \operatorname{diag}(Y)=1} \frac{1}{n}\operatorname{tr} Y\big(\Pi_n - X(X^\top X + n\Lambda)^{-1}X\big).$$

In this chapter, we propose a novel sparse extension to discriminative clustering framework with the square loss. Specifically we formulate a non-trivial sparse regularizer which is a combination of weighted squared ℓ_1-norm and ℓ_2-norm. It leads to

$$\Gamma(V) = \operatorname{tr}[\operatorname{Diag}(a)V\operatorname{Diag}(a)] + \|\operatorname{Diag}(c)V\operatorname{Diag}(c)\|_1, \tag{5.17}$$

such that $\Gamma(vv^\top) = \sum_{i=1}^d a_i^2 v_i^2 + \big(\sum_{i=1}^d c_i|v_i|\big)^2$. This allows to treat all situations simultaneously, with $\nu = 1$ or with $\nu \in [0,1)$. To be more precise, when $\nu \in [0,1)$, we can consider in Eq. (5.14), a problem of size $d+1$ with a design matrix $[X, 1_n] \in \mathbb{R}^{n\times(d+1)}$, a direction of projection $\binom{v}{b} \in \mathbb{R}^{d+1}$ and different weights for the last variable with $a_{d+1} = \frac{\nu}{1-\nu}$ and $c_{d+1} = 0$.

Note that the sparse regularizers on V introduced in this chapter are significantly different when compared to the sparse regularizers on variable v in Eq. (5.3), for example, considered by Wang et al. [2013]. A straightforward sparse regularizer on v in Eq. (5.3), despite leading to a sparse projection, does not yield natural generalizations of the discriminative clustering framework in terms of theory or algorithms.

In our analysis and experiments for the balanced clusters (when $\nu = 1$), the sparse regularization $\Gamma = \lambda\|\cdot\|_1$, for $\lambda \in \mathbb{R}$ will often be considered. This is equivalent to setting $a = 0_d$ and $c = \sqrt{\lambda}1_d$ in Eq. (5.17). The problem in Eq. (5.15) then becomes

$$\max_{V\succcurlyeq 0} \frac{2}{n}\sum_{i=1}^n \sqrt{(XVX^\top)_{ii}} - \frac{1}{n}\operatorname{tr}(VX^\top X) - \lambda\|V\|_1. \tag{5.18}$$

The sparse regularizers considered in this chapter have a significant algorithmic appeal for certain applications in computer vision [Bojanowski et al., 2013, Alayrac et al., 2016], audio processing [Lajugie et al., 2016] and natural language processing [Grave, 2014]. They also lead to robust cluster recovery under minor assumptions as will be illustrated on a simple example in Section 5.5. The practical benefits of the sparse regularizers will be further demonstrated using empirical evaluation on synthetic and real data sets in Section 5.7.

5.4 Extension to Multiple Labels

The discussion so far has focussed on two clusters. Yet it is key in practice to tackle more clusters. It is worth noting that the discrete formulations in Eq. (5.2) and Eq. (5.5) extend directly to more than two clusters. However two different extensions of the initial problems Eq. (5.2) or Eq. (5.5) are conceivable. They lead to problems with different constraints on different optimization domains and, consequently, to different relaxations. We discuss these possibilities next.

One extension is the *multi-class* case. The multi-class problem which is dealt with by Bach and Harchaoui [2007] assumes that the data are clustered into K classes and the various partitions of the data points into clusters are represented by the K-class indicator matrices $y \in \{0,1\}^{n\times K}$ such that $y1_K = 1_n$. The constraint $y1_K = 1_n$ ensures that one data point belongs to only one cluster. However as discussed by Bach and Harchaoui [2007], by letting $Y = yy^\top$, it is possible to lift these K-class indicator matrices into the outer convex approximations $\mathcal{C}_K = \{Y \in \mathbb{R}^{n\times n} : Y = Y^\top, \operatorname{diag}(Y) = 1_n, Y \succcurlyeq 0, Y \preccurlyeq \frac{1}{K}1_n 1_n^\top\}$ [Frieze and Jerrum, 1995], which is different for all values of K. Note that letting $K = 2$ corresponds to the previous sections.

In this chapter, we consider a different novel extension for discriminative clustering to the *multi-label* case. The multi-label problem assumes that the data share k labels and the data-label membership is represented by matrices $y \in \{-1,+1\}^{n\times k}$. In other words, the multi-class problem embeds the data in the extreme points of a simplex, while the multi-label problem does so in the extreme points of the hypercube.

The discriminative clustering formulation of the multi-label problem is

$$\min_{v\in\mathbb{R}^{d\times k},\ y\in\{-1,1\}^{n\times k}} \frac{1}{n}\|\Pi_n y - Xv\|_F^2, \tag{5.19}$$

where the Frobenius norm is defined for any vector or rectangular matrix as $\|A\|_F^2 = \operatorname{tr} AA^\top = \operatorname{tr} A^\top A$. Letting $k = 1$ here corresponds to the previous sections. The discrete ensemble of matrices $y \in \{-1,+1\}^{n\times k}$ can be naturally lifted into $\mathcal{D}_k = \{Y \in \mathbb{R}^{n\times n} : Y = Y^\top, \operatorname{diag}(Y) = k1_n, Y \succcurlyeq 0\}$, since $\operatorname{diag}(Y) = \operatorname{diag}(yy^\top) = \sum_{i=1}^k y_{i,i}^2 = k$. As the optimization problems in Eq. (5.7) and Eq. (5.8) have linear objective functions, we can change the variable from Y to $\tilde{Y} = Y/k$ to change the constraint $\operatorname{diag}(Y) = k1_n$ to $\operatorname{diag}(\tilde{Y}) = 1_n$ without changing the optimizer of the problem. Thus the problems can be solved over the relaxed domain $\mathcal{D} = \{Y \in \mathbb{R}^{n\times n} : Y = Y^\top, \operatorname{diag}(Y) = 1_n, Y \succcurlyeq 0\}$ which is independent of k.

Note that the domain $\mathcal{D}$ is similar to that considered in the problems in Eq. (5.8) and Eq. (5.11) and these convex relaxations are the same regardless of the value of k. Hence the multi-label problem is a more natural extension of the discriminative framework, with a slight change in how the labels y are recovered from the solution Y (we discuss this in Section 5.5.3).

5.5 Theoretical Analysis

In this section, we provide the first theoretical analysis for the discriminative clustering framework with the square loss. We start with the 2-clusters situation: the non-sparse case is considered first and analysis is provided for both balanced and imbalanced clusters. Our study for the sparse case currently only provides results for the simple 1-sparse solution. However, the analysis also yields valuable insights on the scaling between n and d. We then derive results for multi-label situation.

For ease of analysis, we consider the constrained problem in Eq. (5.7), the penalized problem in Eq. (5.8) or their equivalent relaxations in Eq. (5.12) or Eq. (5.18) under various scenarios, for which we use the same proof technique. We first try to characterize the low-rank solutions of these relaxations and then show in certain simple situations the uniqueness of such solutions, which are then non-ambiguously found by convex optimization. Perturbation arguments could extend these results by weakening our assumptions but are not within the scope of this chapter, and hence we do not investigate them further in this section.

5.5.1 Analysis for 2 Clusters: Non-Sparse Problems

In this section, we consider several noise models for the problem, either adding irrelevant dimensions or perturbing the label vector with noise. We consider these separately for simplicity, but they could also be combined (with little extra insight).

Irrelevant Dimensions

We consider an "ideal" design matrix $X \in \mathbb{R}^{n \times d}$ such that there exists a direction v along which the projection Xv is perfectly clustered into two distinct real values c_1 and c_2. Since Eq. (5.2) is invariant by affine transformation, we can rotate the design matrix X to have $X = [y, Z]$ with $y \in \{-1, 1\}^n$, which is clustered into $+1$ or -1 along the direction $v = \binom{1}{0_{d-1}}$. Then after being centered, the design matrix is written as $X = [\Pi_n y, Z]$ with $Z = [z_1, \dots, z_{d-1}] \in \mathbb{R}^{n \times (d-1)}$. The columns of Z represent the noisy irrelevant dimensions added on top of the signal y.

Balanced Problem

When the problem is well balanced ($y^\top 1_n = 0$), y is already centered and $\Pi_n y = y$. Thus the design matrix is represented as $X = [y, Z]$. We consider here the penalized formulation in Eq. (5.8) with $\nu = 1$ which is the only scenario where we are able to provide a theoretical analysis.

Let us assume that the columns $(z_i)_{i=1,\dots,d-1}$ of Z are i.i.d. with symmetric distribution z, with $\mathbb{E}z = \mathbb{E}z^3 = 0$ and such that $\|z\|_\infty$ is almost surely bounded by $R \geq 0$. We denote by $\mathbb{E}z^2 = m$ its second moment and by $\mathbb{E}z^4/(\mathbb{E}z^2)^2 = \beta$ its (unnormalized) kurtosis.

Surprisingly the clustered vector y happens to generate a solution $yy^\top$ of the relaxation Eq. (5.8) for all possible values of Z (see Lemma 35 in Appendix 5.D.2).

However the problem in Eq. (5.8) should have a *unique* solution in order to always recover the correct assignment y. Unfortunately the semidefinite constraint $Y \succcurlyeq 0$ of the relaxation makes the second-order information arduous to study. Due to this reason, we consider the other equivalent relaxation in Eq. (5.12) for which $V_* = vv^\top$ is also solution with $v \propto (X^\top X)^{-1} X^\top y$ (see Lemma 36 in Appendix 5.D.3). Fortunately the semidefinite constraint $V \succcurlyeq 0$ of the problem in Eq. (5.12) may be ignored since the second-order information in V of the objective function already provides unicity for the unconstrained problem. Hence we are able to ensure the uniqueness of the solution with high probability.

Proposition 19. *Let us assume $d \geq 3$, $\beta > 1$ and $m^2 \geq \frac{\beta-3}{2(d+\beta-4)}$:*
(a) If $n \geq d^2 R^4 \frac{1+(d+\beta)m^2}{m^2(\beta-1)}$, V_* is the unique solution of the problem in Eq. (5.12) with high probability.
(b) If $n \geq \frac{d^2 R^4}{\min\{m^2(\beta-1), 2m^2, 2m\}}$, v is the principal eigenvector of any solution of the problem in Eq. (5.12) with high probability.

Let us make the following observations:

— **Proof technique**: The proof relies on a computation of the Hessian of $f(V) = \frac{2}{n}\sum_{i=1}^n \sqrt{(XVX^\top)_{ii}} - \frac{1}{n}\operatorname{tr} X^\top XV$ which is the objective function in Eq. (5.12). We first derive the expectation of $\nabla^2 f(V)$ with respect to the distribution of X. By the law of large numbers, it amounts to have n going to infinity in $\nabla^2 f(V)$. Then we expand the spectrum of this operator $\mathbb{E}\nabla^2 f(V)$ to lower-bound its smallest eigenvalue. Finally we use concentration theory on matrices, following Tropp [2012], to bound the Hessian $\nabla^2 f(V)$ for finite n.

— **Effect of kurtosis**: We remind that $\beta \geqslant 1$, with equality if and only if z follows a Rademacher law ($\mathbb{P}(z = +1) = \mathbb{P}(z = -1) = 1/2$). Thus, if the noisy dimensions are clustered, then unsurprisingly, our guarantee is meaningless. Note that the constant β behaves like a distance of the distribution z to the Rademacher distribution. Moreover, $\beta = 3$ if z follows a standard normal distribution.

— **Scaling between d and n**: If the noisy variables are not evenly clustered between the same clusters $\{\pm 1\}$ (i.e., $\beta > 1$), we recover a rank-one solution as long as $n = O(d^3)$; while, as long as $n = O(d^2)$, the solution is not unique but its principal eigenvector recovers the correct clustering. Moreover, as explained in the proof, its spectrum would be very spiky.

— The assumption $m^2 \geq \frac{\beta-3}{2(d+\beta-4)}$ is generally satisfied for large dimensions. Note that $m^2 d$ is the total variance of the irrelevant dimensions, and when it is small, i.e., when $m^2 \leq \frac{\beta-3}{2(d+\beta-4)}$, the problem is particularly simple, and we can also show that V_* is the unique solution of the problem in Eq. (5.12) with high probability if $n \geq \frac{d^2 R^4}{m^2}$. Finally, note that for sub-Gaussian distributions (where $\beta \leq 3$), the extra constraint is vacuous, while for super-Gaussian distributions (where $\beta \geq 3$), this extra constraint only appears for small m.

— This result provides the first guarantee for discriminative clustering. However similar theoretical results have been derived for K-means by Ostrovsky et al.

[2006] and Gaussian mixtures by Kalai et al. [2010], Moitra and Valiant [2010], where separation conditions between the two clusters are derived, under which the clustering problem is efficiently solved. It would be of great interest to relate these separation conditions to our condition on n and d but this is outside the scope of this work.

Noise Robustness for the 1-Dimensional Balanced Problem

We assume now that the data are one-dimensional and are perturbed by some noise $\varepsilon \in \mathbb{R}^n$ such that $X = y + \varepsilon$ with $y \in \{-1, 1\}^n$. The solution of the relaxation in Eq. (5.8) recovers the correct y in this setting only when each component of y and $y + \varepsilon$ have the same sign (this is shown in Appendix 5.D.5). This result comes out naturally from the information on whether the signs of y and $y + \varepsilon$ are the same or not. Further if we assume that y and ε are independent, this condition is equivalent to $\|\varepsilon\|_\infty < 1$ almost surely.

Unbalanced Problem

When the clusters are imbalanced ($y^\top 1_n \neq 0$), the natural rank-one candidates $Y_* = yy^\top$ and $V_* = vv^\top$ are no longer solutions of the relaxations in Eq. (5.8) (for $\nu = 1$) and Eq. (5.12), as proved in Appendix 5.D.6. Nevertheless we are able to characterize some solutions of the penalized relaxation in Eq. (5.8) for $\nu = 0$.

Lemma 28. *For $\nu = 0$ and for any non-negative $a, b \in \mathbb{R}$ such that $a + b = 1$,*

$$Y = ayy^\top + b1_n 1_n^\top$$

is solution of the penalized relaxation in Eq. (5.8).

Hence any eigenvector of this solution Y would be supported by the directions y and 1_n. Moreover when the value $\alpha_* = (\frac{1_n^\top y}{n})^2$ is known, it turns out that we can characterize some solutions of the constrained relaxation in Eq. (5.7), as stated in the following lemma.

Lemma 29. *For $\alpha \geq \alpha_*$,*

$$Y = \frac{1-\alpha}{1-\alpha_*} yy^\top + \Big(1 - \frac{1-\alpha}{1-\alpha_*}\Big) 1_n 1_n^\top$$

is a rank-2 solution of the constrained relaxation in Eq. (5.7) with constraint parameter α.

The eigenvectors of Y enable to recover y for $\alpha_* \leq \alpha < 1$. We conjecture (and checked empirically) that this rank-2 solution is unique under similar regimes to those considered for the balanced case. The proof would be more involved since, when $\nu \neq 1$, we are not able to derive an equivalent problem in V for the penalized relaxation in Eq. (5.8) similar to Eq. (5.12) for the balanced case. We also note that Lemmas 28 and 29 will be direct consequences of Lemma 32 in Section 5.5.3.

Thus Y being rank-2, one should really be careful and consider the first two eigenvectors when recovering y from a solution Y. This can be done by rounding the principal eigenvector of $\Pi_n Y \Pi_n = \frac{1-\alpha}{1-\alpha_*}\Pi_n y(\Pi_n y)^\top$ as discussed in the following lemma.

Lemma 30. *Let y_{ev} be the principal eigenvector of $\Pi_n Y \Pi_n$ where Y is defined in Lemma 29, then*

$$\operatorname{sign}(y_{ev}) = y.$$

Proof. By definition of Y, $y_{ev} = \sqrt{\frac{1-\alpha}{1-\alpha_*}}\Pi_n y$ thus $\operatorname{sign}(y_{ev}) = \operatorname{sign}(\Pi_n y)$ and since $\alpha \leq 1$ then $\operatorname{sign}(\Pi_n y) = \operatorname{sign}(y - \sqrt{\alpha}1_n) = y$. □

In practice, contrary to the standard procedure, we should, for any ν, solve the penalized relaxation in Eq. (5.8) and then do K-means on the principal eigenvector of the centered solution $\Pi_n Y \Pi_n$ instead of the solution Y to recover the correct y. This procedure is followed in our experiments on real-world data in Section 5.7.2.

5.5.2 Analysis for 2 Clusters: 1-Sparse Problems

We assume here that the direction of projection v (such that $Xv = y$) is l-sparse (by l-sparse we mean $\|v\|_0 = l$). The ℓ_1-norm regularized problem in Eq. (5.18) is no longer invariant by affine transformation and we cannot consider that $X = [y, Z]$ without loss of generality. Yet the relaxation Eq. (5.18) seems experimentally to only have rank-one solutions for the simple $l = 1$ situation. Hence we are able to derive some theoretical analysis only for this case. It is worth noting the $l = 1$ case is simple since it can be solved in $O(d)$ by using K-means separately on all dimensions and ranking them. Nonetheless the proposed scaling also holds in practice for $l \geqslant 1$ (see Figure 5-1b).

Thereby we consider data $X = [y, Z]$ with $y \in \{-1, 1\}^n$ and $Z \in \mathbb{R}^{n\times(d-1)}$ which are clustered in the direction $v = [1, 0, \ldots, 0]^\top \in \mathbb{R}^d$. When adding a ℓ_1-penalty, the initial problem in Eq. (5.5) for $\alpha = 0$ is

$$\min_{y\in\{-1,1\}^n,\ v\in\mathbb{R}^d} \frac{1}{n}\|y - Xv\|_2^2 + \lambda\|v\|_1^2. \tag{5.20}$$

When optimizing in v this problem is close to the Lasso [Tibshirani, 1996] and a solution is known to be $v_i^* = (y^\top y + n\lambda)^{-1} y^\top y = \frac{1}{1+\lambda},\ \forall i \in J$ and $v_i^* = 0,\ \forall i \in \{1, 2, \ldots, d\} \setminus J$, where J is the support of v^*. The candidate $V_* = v^* v^{*\top}$ is still a solution of the relaxation in Eq. (5.18) (see Lemma 39 in Appendix 5.E.1) and we will investigate under which conditions on X this solution is unique. Let us assume as before $(z_i)_{i=1,\ldots,d}$ are i.i.d. with distribution z symmetric with $\mathbb{E}z = \mathbb{E}z^3 = 0$, and denote by $\mathbb{E}z^2 = m$ and $\mathbb{E}z^4/(\mathbb{E}z^2)^2 = \beta$. We also assume that $\|z\|_\infty$ is almost surely bounded by $0 \leq R \leq 1$. We are able to ensure the uniqueness of the solution with high-probability.

Proposition 20. *Let us assume $d \geq 3$.*
(a) If $n \geq dR^2 \frac{1+(d+\beta)m^2}{m^2(\beta-1)}$, V_* is the unique solution of the problem Eq. (5.12) with high probability.
(b) If $n \geq \frac{dR^2}{m^2(\beta-1)}$, v^* is the principal eigenvector of any solution of the problem Eq. (5.12) with high probability.

The proof technique is very similar to the one of Proposition 19. With the function $g(V) = \frac{2}{n}\sum_{i=1}^n \sqrt{(XVX^\top)_{ii}} - \lambda\|V\|_1 - \frac{1}{n}\operatorname{tr} X^\top XV$, we can certify that g will decrease around the solution V_* by analyzing the eigenvalues of its Hessian.

The rank-one solution V_* is recovered by the principal eigenvector of the solution of the relaxation Eq. (5.18) as long as $n = O(d)$. Thus we have a much better scaling when compared to the non-sparse setting where $n = O(d^2)$. We also conjecture a scaling of order $n = O(ld)$ for a projection in a l-sparse direction (see Figure 5-1b for empirical results).

The proposition does not state any particular value for the regularizer parameter λ. This makes sense since the proposition only holds for the simple situation when $l = 1$. We propose to use $\lambda = 1/\sqrt{n}$ by analogy with the Lasso.

5.5.3 Analysis for the Multi-Label Extension

In this section, the signals share k labels which are corrupted by some extra noisy dimensions. We assume the centered design matrix to be $X = [\Pi_n y, Z]$ where $y \in \{-1, +1\}^{n\times k}$ and $Z \in \mathbb{R}^{n\times(d-k)}$. We also assume that y is full-rank [1]. We denote by $y = [y_1, \ldots, y_k]$ and $\alpha_i = \left(\frac{y_i^\top 1_n}{n}\right)^2$ for $i = 1, \ldots, k$. We consider the discrete constrained problem

$$\min_{v\in\mathbb{R}^{d\times k},\ y\in\{-1,1\}^{n\times k}} \frac{1}{n}\|\Pi_n y - Xv\|_F^2 \text{ such that } \frac{1_n^\top yy^\top 1_n}{n^2} = \alpha^2, \tag{5.21}$$

and the discrete penalized problem for $\nu = 0$

$$\min_{v\in\mathbb{R}^{d\times k},\ y\in\{-1,1\}^{n\times k}} \frac{1}{n}\|\Pi_n y - Xv\|_F^2. \tag{5.22}$$

As explained in Section 5.4, these two discrete problems admit the same relaxations in Eq. (5.7) and Eq. (5.8) we have studied for one label. We now investigate when the solution of the problems in Eq. (5.21) and in Eq. (5.22) generate solutions of the relaxations in Eq. (5.7) and Eq. (5.8).

By analogy with Lemma 28, we want to characterize the solutions of these relaxations which are supported by the constant vector 1_n and the labels $(y_1, \ldots, y_k)$. Their general form is $Y = \tilde{y}A\tilde{y}^\top$ where $A \in \mathbb{R}^{k\times k}$ is symmetric semi-definite positive and $\tilde{y} = [1_n, y]$. However the initial y is easily recovered from the solution Y only

1. This assumption is fairly reasonable since the probability of a matrix whose entries are i.i.d. Rademacher random variables to be singular is conjectured to be $1/2 + o(1)$ [Bourgain et al., 2010].

when A is diagonal. To that end the following lemma derives some condition under which the only matrix A such that the corresponding Y satisfies the constraint of the relaxations in Eq. (5.7) and Eq. (5.8) is diagonal.

Lemma 31. *The solutions of the matrix equation* $\operatorname{diag}(\tilde{y}A\tilde{y}^\top) = 1_n$ *with unknown variable* A *are diagonal if and only if the family* $\{1_n, (y_i)_{1\leq i\leq k}, (y_i \odot y_j)_{1\leq i<j\leq k}\}$ *is linearly independent where we denoted by* $\odot$ *the Hadamard (i.e., pointwise) product between matrices.*

In this way we are able to characterize the solution of relaxations in Eq. (5.7) and Eq. (5.8) with the following result:

Lemma 32. *Let us assume that the family* $\{1_n, (y_i)_{1\leq i\leq k}, (y_i \odot y_j)_{1\leq i<j\leq k}\}$ *is linearly independent. If* $\alpha \geq \alpha_{\min} = \min_{1\leq i\leq k}\{\alpha_i\}$ *with* $(\alpha_i)_{1\leq i\leq k}$ *defined above Eq. (5.21), the solutions of the constrained relaxation in Eq. (5.7) supported by the vectors* $(1_n, y_1, \ldots, y_k)$ *are of the form:*

$$Y = a_0^2 1_n 1_n^\top + \sum_{i=1}^k a_i^2 y_i y_i^\top,$$

where $(a_i)_{0\leq i\leq k}$ *satisfies* $\sum_{i=0}^k a_i^2 = 1$ *and* $a_0^2 + \sum_{i=1}^k a_i^2 \alpha_i = \alpha$.

Moreover the solutions of the penalized relaxation in Eq. (5.8) for $\nu = 0$ *which are supported by the vectors* $(1_n, y_1, \ldots, y_k)$ *are of the form:*

$$Y = a_0^2 1_n 1_n^\top + \sum_{i=1}^k a_i^2 y_i y_i^\top,$$

where $(a_i)_{0\leq i\leq k}$ *satisfies* $\sum_{i=0}^k a_i^2 = 1$.

In the *multi-label* case, some combinations of the constant matrix $1_n 1_n^\top$ and the rank-one matrices $y_i y_i^\top$ are solutions of constrained or penalized relaxations. Furthermore, under some assumptions on the labels $(y_i)_{1\leq i\leq k}$, these combinations are the only solutions which are supported by the vectors $(1_n, y_1, \ldots, y_k)$. And we conjecture (and checked empirically) that under assumptions similar to those made for the balanced one-label case, all the solutions of the relaxation are supported by the family $(1_n, y_1, \ldots, y_k)$ and consequently share the same form as in Lemma 32. Thus the eigenvector of the solution Y would be in the span of the directions $(1_n, y_1, \ldots, y_k)$.

Let us consider an eigenvalue decomposition of $Y = FF^\top = \sum_{i=0}^k \lambda_i e_i e_i^\top$ and denote by $M = [a_0 1_n, a_1 y_1, \ldots, a_k y_k]$ where $(a_i)_{0\leq i\leq k}$ are defined in Lemma 32. Since $MM^\top = FF^\top$, there is an orthogonal transformation R such that $FR = M$. We also denote the product FR by $FR = [\xi_0, \ldots, \xi_K]$. We propose now an alternating minimization procedure to recover the labels $(y_1, \ldots, y_k)$ from M.

Lemma 33. *Consider the optimization problem*

$$\min_{M\in\mathcal{M},\ R\in\mathbb{R}^{k\times k}:\ R^\top R = I_k} \|FR - M\|_F^2,$$

where $\mathcal{M} = \{[a_0 1_n, a_1 y_1, \ldots, a_k y_k], a \in \mathbb{R}^{k+1} : \|a\|_2 = 1, y_i \in \{\pm 1\}^n\}$.

Given M, the problem is equivalent to the orthogonal Procrustes problem [Schönemann, 1966]. Denote by $U\Delta V^\top$ a singular value decomposition of $F^\top M$. The optimal R is obtained as $R = UV^\top$. While given R, the optimal M is obtained as

$$M = \frac{1}{\sqrt{\|\xi_1\|_1^2 + \|\xi_2\|_1^2 + \ldots + \|\xi_k\|_1^2}} [\|\xi_0\|_1 \operatorname{sign}(\xi_0), \ldots, \|\xi_k\|_1 \operatorname{sign}(\xi_k)].$$

Proof. We give only the argument for the optimization problem with respect to M. Given R, the optimization problem in M is equivalent to $\max_{a \in \mathbb{R}^{k+1}} \operatorname{tr}(FR)^\top M$ s.t. $\|a\|_2 = 1,\ y \in \{-1, 1\}^{n \times k}$ and $\operatorname{tr}(FR)^\top M = a_0 \xi_0^\top 1_n + \sum_{i=1}^k a_i \xi_i^\top y_i$. Thus by property of the dual norms the solution is given by $y_i = \operatorname{sign}(\xi_i)$ and $a_i = \frac{\|\xi_i\|_1}{\sqrt{\|\xi_1\|_1^2 + \|\xi_2\|_1^2 + \ldots + \|\xi_k\|_1^2}}$. □

The minimization problem in Lemma 33 is non-convex; however we observe that performing few alternating optimizations is sufficient to recover the correct $(y_1, \ldots, y_k)$ from M.

5.5.4 Discussion

In this section we studied the tightness of convex relaxations under simple scenarios where the relaxed problem admits low-rank solutions generated by the solution of the original non-convex problem. Unfortunately the solutions lose the characterized rank when the initial problem is slightly perturbed since the rank of a matrix is not a continuous function. Nevertheless, the spectrum of the new solution is really spiked, and thus these results are quite conservative. We empirically observe that the principal eigenvectors keep recovering the correct information outside these scenarios. However this simple proof mechanism is not easily adaptable to handle perturbed problems in a straightforward way since it is difficult to characterize the properties of eigenvectors of the solution of a semi-definite program. Hence we are able to derive a proper theoretical study only for these simple models.

5.6 Algorithms

In this section, we present an optimization algorithm which is adapted to large n settings, and avoids the n-dimensional semidefinite constraint.

5.6.1 Reformulation

We aim to solve the general regularized problem which correponds to Eq. (5.15)

$$\max_{V \succcurlyeq 0} \frac{2}{n} \sum_{i=1}^n \sqrt{(XVX^\top)_{ii}} - \frac{1}{n} \operatorname{tr} V(X^\top X + n \operatorname{Diag}(a)^2) - \|\operatorname{Diag}(c) V \operatorname{Diag}(c)\|_1. \quad (5.23)$$

We consider a slightly different optimization problem:

$$\begin{aligned} \max_{V \succcurlyeq 0} \quad & \frac{1}{n} \sum_{i=1}^{n} \sqrt{(XVX^\top)_{ii}} - \| \operatorname{Diag}(c) V \operatorname{Diag}(c) \|_1 \\ \text{s.t.} \quad & \operatorname{tr} V (\frac{1}{n} X^\top X + \operatorname{Diag}(a)^2) = 1. \end{aligned} \tag{5.24}$$

When c is equal to zero, then Eq. (5.24) is exactly equivalent to Eq. (5.23); when c is small (as will typically be the case in our experiments), the solutions are very similar—in fact, one can show by Lagrangian duality that by a sequence of problems in Eq. (5.24), one may obtain the solution to Eq. (5.23).

5.6.2 Smoothing

By letting $A = \frac{X^\top X}{n} + \operatorname{Diag}(a)^2$, we consider a strongly convex approximation of Eq. (5.24) as:

$$\begin{aligned} \max_{V \succcurlyeq 0} \quad & \frac{1}{n} \sum_{i=1}^{n} \sqrt{(XVX^\top)_{ii}} - \| \operatorname{Diag}(c) V \operatorname{Diag}(c) \|_1 - \varepsilon \operatorname{tr}[(A^{\frac{1}{2}} V A^{\frac{1}{2}}) \log(A^{\frac{1}{2}} V A^{\frac{1}{2}})] \\ \text{s.t.} \quad & \operatorname{tr}(A^{\frac{1}{2}} V A^{\frac{1}{2}}) = 1, \end{aligned} \tag{5.25}$$

where $-\operatorname{tr} M \log(M)$ is a spectral convex function called the von Neumann entropy [von Neumann, 1927]. The difference in the two problems is known to be $\varepsilon \log(d)$ [Nesterov, 2007]. As shown in Appendix 5.G.1, the dual problem is

$$\min_{u \in \mathbb{R}^n_+, C \in \mathbb{R}^{d \times d} : |C_{ij}| \leqslant c_i c_j} \frac{1}{2n} \sum_{i=1}^{n} \frac{1}{u_i} + \phi^\varepsilon\big(A^{-\frac{1}{2}} \big(\frac{1}{2n} X^\top \operatorname{Diag}(u) X - C\big) A^{-\frac{1}{2}}\big), \tag{5.26}$$

where $\phi^\varepsilon(M)$ is an ε-smooth approximation to the maximal eigenvalue of the matrix M.

5.6.3 Optimization Algorithm

In order to solve Eq. (5.26), we split the objective function into a smooth part $F(u, C) = \phi^\varepsilon\big(A^{-\frac{1}{2}} \big(\frac{1}{2n} X^\top \operatorname{Diag}(u) X - C\big) A^{-\frac{1}{2}}\big)$ and a non-smooth part $H(u, C) = \mathbb{I}_{|C_{ij}| \leqslant c_i c_j} + \frac{1}{2n} \sum_{i=1}^{n} \frac{1}{u_i}$. We may then apply FISTA [Beck and Teboulle, 2009] updates to the smooth function $\phi^\varepsilon(A^{-\frac{1}{2}}(\frac{1}{2n} X^\top \operatorname{Diag}(u) X - C) A^{-\frac{1}{2}})$, along with a proximal operator for the non-smooth terms $\mathbb{I}_{|C_{ij}| \leqslant c_i c_j}$ and $\frac{1}{2n} \sum_{i=1}^{n} \frac{1}{u_i}$, which may be computed efficiently. See details in Appendix 5.G.2.

Running-time complexity. Since we need to project on the SDP cone of size d at each iteration, the running-time complexity per iteration is $O(d^3 + d^2 n)$; given that often $n \geqslant d$, the dominating term is $O(d^2 n)$. It is still an open problem to make this linear in d. Our function being $O(1/\varepsilon)$-smooth, the convergence rate is of the form

$O(1/(\varepsilon t^2))$. Since we stop when the duality gap is $\varepsilon \log(d)$ (as we use smoothing, it is not useful to go lower), the number of iterations is of order $1/(\varepsilon\sqrt{\log(d)})$. The proposed algorithm is a clear improvement over the existing approach by Bach and Harchaoui [2007] which is quadratic in n.

5.7 Experiments

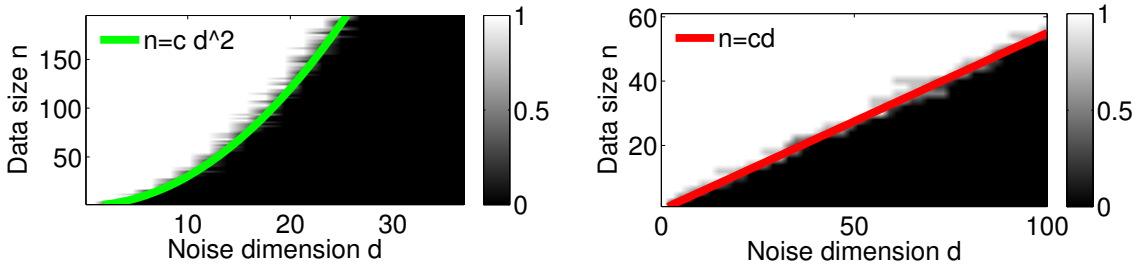
We implemented the proposed algorithm in Matlab. The code is available in `https://drive.google.com/uc?export=download&id=0B5Bx9jrp7celMk5pOFI4UGt0ZEk`. Two sets of experiments were performed: one on synthetically generated data sets and the other on real-world data sets. The details about experiments follow.

5.7.1 Experiments on Synthetic Data

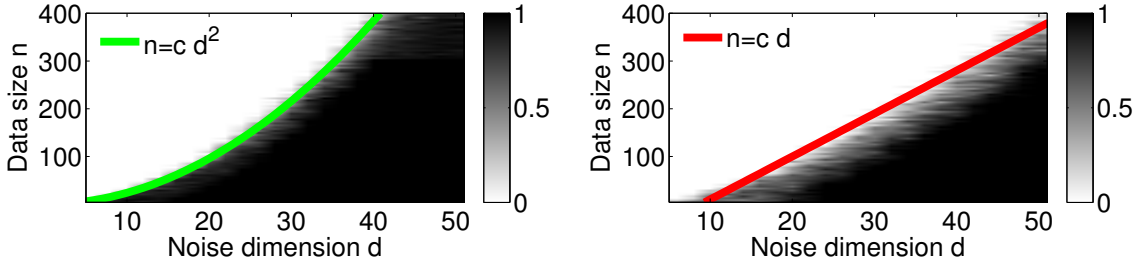
In this section, we illustrate our theoretical results and algorithms on synthetic examples. The synthetic data were generated by assuming a fixed clustering with $\alpha_* \in [0, 1]$, along a single direction and the remaining variables were whitened. We consider clustering error defined for a predictor $\bar{y}$ as $1-(\bar{y}^\top y/n)^2$, with values in $[0, 1]$ and equal to zero if and only if $y = \bar{y}$.

Phase transition. We first illustrate our theoretical results for the balanced case in Figure 5-1. We solve the relaxation in Eq. (5.12) and Eq. (5.18) for a large range of d and n using the `cvx` solver [Grant and Boyd, 2008, 2014]. We show the results averaged over 4 replications and take $\lambda = 1/\sqrt{n}$ for the sparse problems. In Figure 5-1a we investigate whether `cvx` finds a rank-one solution for a problem of size (n, d) (the value is 1 if the solution is rank-one and 0 otherwise). We compare the performance of the algorithms without ℓ_1-regularization in the affine invariant case and with ℓ_1-regularization in the 1-sparse case. We observe a phase transition with a scaling over the form $n = O(d^2)$ for the affine invariant case and $n = O(d)$ for the 1-sparse case. This is better than what is expected by the theory and corresponds rather to the performance of the principal eigenvector of the solution. It is worth noting that it may be uncertain to really distinguish between a rank-one solution and a spiked solution.

We also solve the relaxation for 4-sparse problems of different sizes d and n and plot the clustering error. We compare, in Figure 5-1b, the performance of the formulation in Eq. (5.12) (without ℓ_1-regularization) which corresponds to the affine invariant case, against the ℓ_1-regularized formulation in Eq. (5.18). We notice a phase transition of the clustering error with a scaling over the form $n = O(d^2)$ for the affine invariant case and $n = O(d)$ for the 4-sparse case. It supports our conjecture on the scaling of order $n = O(ld)$ for l-sparse problems. Comparing left plots of Figure 5-1a and Figure 5-1b, we observe that the two phase-transitions occur at the same scaling between n and d. Thus there are few values of (n, d) for which the `cvx` solver finds a solution whose rank is stricly larger than one and whose principal eigenvector has a low clustering



(a) Phase transition for rank-one solution. Left: affine invariant case. Right: 1-sparse case.



(b) Phase transition for clustering error. Left: affine invariant case . Right: 4-sparse case.

Figure 5-1 – Phase transition plots.

error. This illustrates, in practice, this solver aims to find a rank-one solution under the improved scaling $n = O(d^2)$.

Unbalanced case. We generate an unbalanced problem for $d = 10$, $n = 80$ and $\alpha_* = 0.25$ and we average the results over 10 replications. We compare the clustering error for the constrained and the penalized relaxations in Eq. (5.7) and Eq. (5.8) when we consider the sign of the first or second eigenvector and when we use projection technique defined as $(\Pi_n Y_{(2)} \Pi_n)_{(1)}$ where $Y_{(k)}$ is the best rank-k approximation of Y, to extract the information of y. We see in Figure 5-2 that (a) for the constrained case, the range of α such that the sign of y is recovered is cut in two parts where one eigenvector is correct, whereas the projection method performs well on the whole set. (b) For the penalized case, the correct sign is recovered for ν close to 0 by the first eigenvector and the projection method whereas the second one performs always badly. (c) When there is zero noise the rank of the solution is one for $\alpha \in \{\alpha_*, 1\}$, two for $\alpha \in (\alpha_*, 1)$ and greater otherwise. These findings confirm our analysis. However, when y is corrupted by some noise this result is no longer true.

Runtime experiments. We generated data with a k-sparse direction of projection v by adding $d-k$ noise variables to a randomly generated and rotated k-dimension data. The scalability of the FISTA based optimization algorithm illustrated in Sec-

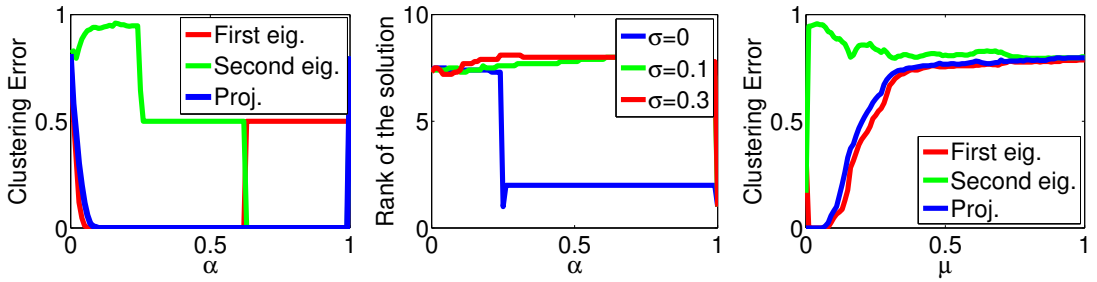


Figure 5-2 – Unbalanced problem for $n = 80$, $d = 10$ and $\alpha_* = 0.25$. Left: Clustering error for the constrained relaxation. Middle: Rank of the solution for different level of noise σ. Right: Clustering error for the penalized relaxation.

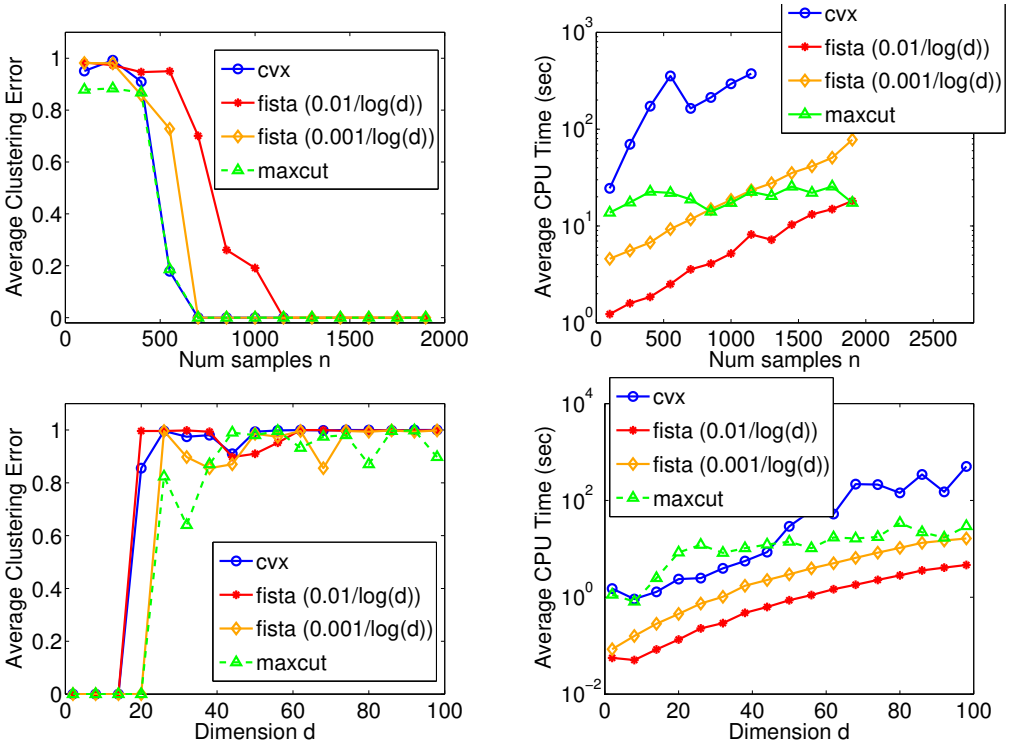
tion 5.6.3 to solve Eq. (5.24) (with $c = \sqrt{\lambda}1_d$, $a = 0_d$) was compared against a benchmark `cvx` solver (which solves Eq. (5.18)). Experiments were performed for $\lambda = 0$ and $\lambda = 0.001$, the coefficient associated with the sparse $\|V\|_1$ term. For a fixed d, `cvx` breaks down for large n values (typically $n \geqslant 1000$). Similarly, the runtime required by `cvx` is generally high for $\lambda = 0$ and is comparable to our method for $\lambda = 0.001$. This behavior is illustrated in Figure 5-3.

When $\lambda = 0$, the problem reduces exactly to the original Diffrac problem [Bach and Harchaoui, 2007]. In the plots in Figure 5-3a our implementation using FISTA is compared to the baseline Diffrac which is solved with max-cut SDP [Boumal et al., 2014]. We observed that our method is comparable in terms of runtime and clustering performance of low-rank methods for max-cut. However, for $\lambda > 0$, the equivalence with max-cut disappears.

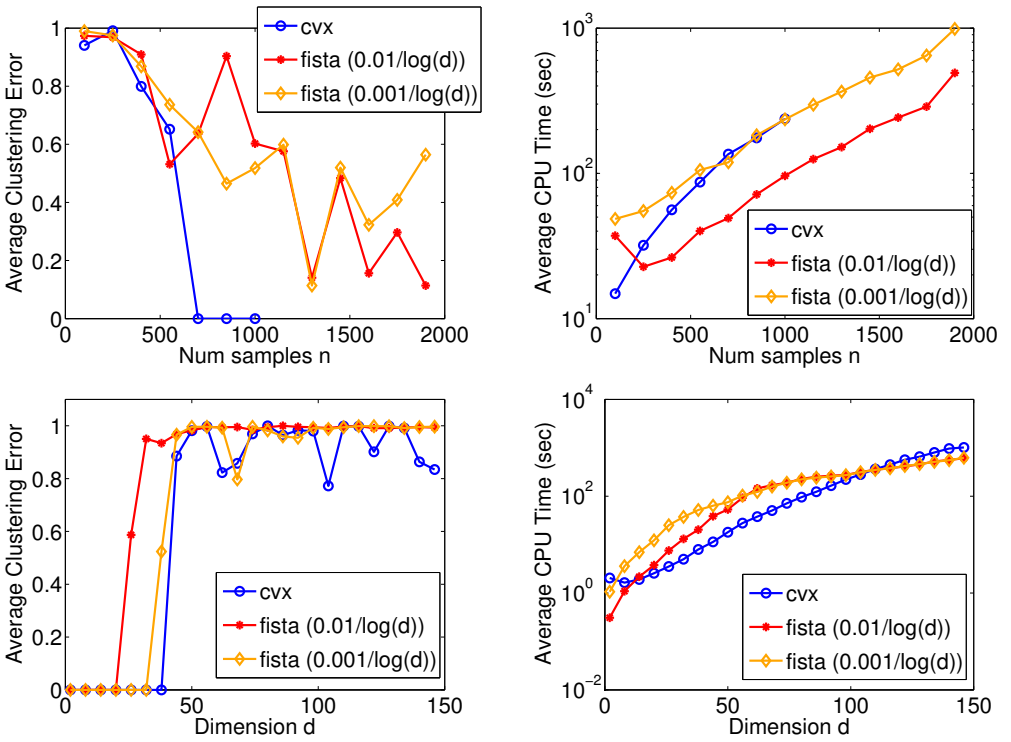
The plots in these figures show the behavior of FISTA for two different stopping criteria: $\varepsilon = 10^{-2}/\log(d)$ and $\varepsilon = 10^{-3}/\log(d)$. It is observed that the choice $10^{-3}/\log(d)$ gives a better accurate solution at the cost of more number of iterations (and hence higher runtime). For sparse problems in Figure 5-3b, we see that cvx gets a better clustering performance (while crashing for large n); the difference would be reduced with a smaller duality gap for FISTA.

Clustering performance. Experiments comparing the proposed method (Eq. (5.24) with $c = \sqrt{\lambda}1_d$ and $a = 0_d$ solved using FISTA based optimization algorithm, and Eq. (5.18) solved using benchmark `cvx` solver) with K-means and alternating optimization are given in Figure 5-4. K-means is run on the whitened variables in $\mathbb{R}^d$. Alternating optimization is another popular method proposed by Ye et al. [2008] for dimensionality reduction with clustering (where alternating optimization of w and y is performed to solve the non-convex formulation (5.2)). The plots show that both K-means and alternating optimization fail when only a few dimensions of noise variables are present. The plots also show that with the introduction of a sparse regularizer (corresponding to the non-zero λ) the proposed method becomes more robust to noisy dimensions. As observed earlier, the performance of FISTA is also sensitive to the choice of ε.

Finally we give a comparison of sparse discriminative clustering (cvx and FISTA) with max-margin clustering [Li et al., 2009] in Figure 5-5. While square loss is used

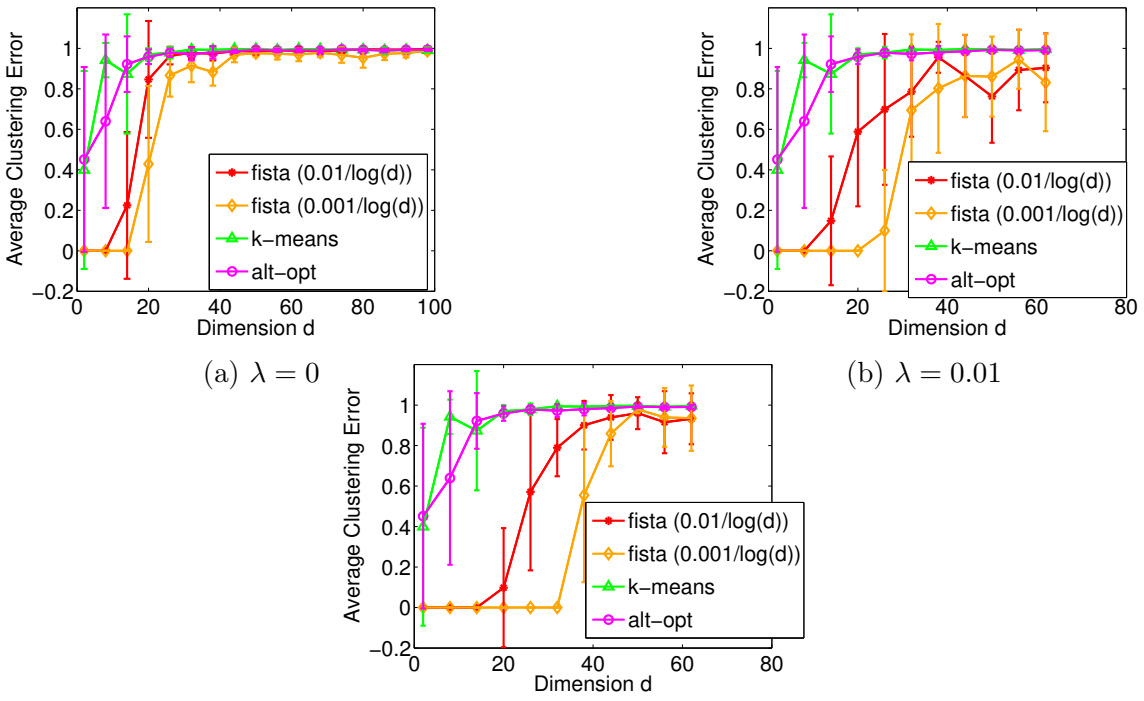


(a) `cvx`, max-cut comparison with $\lambda = 0$. Top: n varied with $d = 50$, $k = 6$. `cvx` crashed for $n \approx 1000$. Bottom: d varied with $n = 100$, $k = 2$.



(b) `cvx` comparison with $\lambda = 0.001$. Top: n varied with $d = 50$, $k = 6$. `cvx` crashed for $n \approx 1000$. Bottom: d varied with $n = 100$, $k = 2$.

Figure 5-3 – Scalability experiments.



(a) $\lambda = 0$ (b) $\lambda = 0.01$

(c) $\lambda = 0.001$

Figure 5-4 – Comparison with k-means and alternating optimization, $n = 100$.

in our framework, hinge loss is used in max-margin clustering. We have also included the behavior of K-means and alternating optimization methods in Figure 5-5 for completeness. From this plot, it is clear that the max-margin clustering is sensitive to noisy dimensions present in the data. Sparse discriminative clustering with square loss is able to maintain zero cluster error for a large number of noisy dimensions, while the performance of max-margin clustering starts deteriorating after adding a few noisy dimensions. However, we note from Figure 5-5 that for large dimensions, the hinge loss used in max-margin clustering is observed to provide a better solution than the square loss used in our framework.

5.7.2 Experiments on Real-World Data

Experiments on two-class data. Experiments were conducted on real two-class classification datasets [2] to compare the performance of sparse discriminative clustering against non-sparse discriminative clustering, alternating optimization, K-means and max-margin clustering algorithms. For sparse and non-sparse discriminative clustering, we consider the problem in Eq. (5.24) and the algorithm detailed in Section 5.6.3 (with the regularization $c = 0$ for the non-sparse case). The alternating optimization method is described in Proposition 18. For the two-class datasets, the clustering performance for a cluster $\bar{y} \in \{+1, -1\}^n$ obtained from an algorithm under comparison,

2. The data sets were obtained from `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`

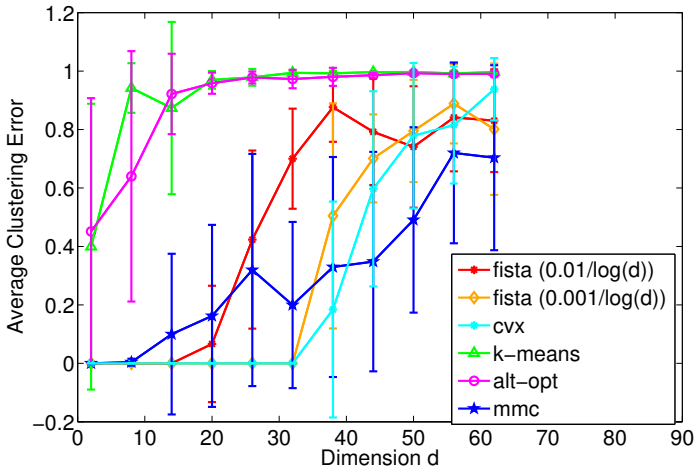


Figure 5-5 – Comparison with k-means, alternating optimization and max-margin clustering (mmc), $n = 100$. The plots for FISTA, `cvx` and mmc correspond to the best choice of regularization parameters.

was computed as $1 - (\bar{y}^\top y/n)^2$, where y is the original labeling. Here we explicitly compare the output of clustering with the original labels of the data points.

The dataset details and clustering performance results are summarized in Table 5.1. The experiments for discriminative clustering were conducted for different values of $a, c \in \{10^{-3}, 10^{-2}, 10^{-1}\}1_d$ associated with the ℓ_2-regularizer and ℓ_1-regularizer respectively. The range of cluster imbalance parameter was chosen to be $\nu \in \{0.01, 0.25, 0.5, 0.75, 1\}$. Note that for $\nu \neq 1$, the reformulation given in Eq. (5.14) was used, as explained in Section 5.3 after Eq. (5.17). The results given in Table 5.1 pertain to the best choices of these parameters. Similarly, the values of regularization parameter for max-margin clustering [Li et al., 2009] were chosen from the set $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 0.1, 1, 10\}$ and the cluster balance parameter was chosen from $\{0.1, 0.2, \ldots, 0.9\}$. The results for alternating optimization and K-means show the average cluster error (and standard deviation) over 10 different runs. These results show that the cluster error is quite high for many datasets. This is primarily due to the absence of an ambient low-dimensional clustering of the two-class data, which can be identified by the simple linear model presented in this chapter. Since K-means does not provide explicit dimensionality reduction, it might not be able to take advantage of the existence of an ambient low-dimensional clustering of the two-class data and its performance is poor. The results show that max-margin clustering achieves best clustering performance on most datasets. This improved performance of max-margin clustering may be due to the use of hinge loss, as opposed to square loss used in discriminative clustering in this chapter. However for the heart dataset, we note that the sparse version with the square loss performs significantly better than the non-sparse version with the hinge loss (see additional experiments in Figure 5-5). The results also show that adding sparse regularizers to discriminative clustering helps in a better cluster identification when compared to the non-sparse case.

Dataset	n	d	Cluster Error				
			S. D.C.	Non-S. D.C.	Alternating Optimization	K-means	Max-margin Clustering
Heart	270	3	**0.52**	0.61	0.97 ± 0.03	0.91 ± 0.09	0.93
Diabetes	768	8	**0.88**	**0.88**	0.91 ± 0.05	0.93 ± 0.06	**0.88**
Breast-cancer	683	10	**0.15**	**0.15**	0.48 ± 0.17	0.68 ± 0.24	**0.15**
Australian	690	14	**0.5**	**0.5**	0.88 ± 0.17	0.87 ± 0.21	**0.5**
Liver-disorder	345	6	0.97	0.97	0.99 ± 0.01	0.99 ± 0.01	**0.73**
Sonar	208	60	**0.92**	0.95	0.98 ± 0.02	0.99 ± 0.01	**0.92**
DNA(1 vs 2,3)	1400	180	0.75	0.83	0.99 ± 0.01	0.98 ± 0.02	**0.71**
a1a	1605	113	0.74	0.75	0.98 ± 0.02	0.8 ± 0.08	**0.69**
w1a	2270	290	**0.11**	**0.11**	0.92 ± 0.08	0.16 ± 0.06	**0.11**

Table 5.1 – Experiments on two-class datasets (S.D.C. means sparse discriminative clustering, and Non-S. D.C. non sparse discriminative clustering.)

Experiments on real multi-label data. Experiments were also conducted on the Microsoft COCO dataset [3] to demonstrate the effectiveness of the proposed method in discovering multiple labels. We considered $n = 2000$ images from the dataset, each of which was labeled with a subset of $K = 80$ labels. The labels identified the objects in the images like person, car, chair, table, etc. and the corresponding features for each image were extracted from the last layer of a conventional convolutional neural network (CNN). The CNN was originally trained over the imagenet data [Krizhevsky et al., 2012].

For each image in the dataset, we obtained $d = 1000$ features. We then performed discriminative clustering on the 2000×1000 data matrix X and obtained the label matrix Y which was then subjected to the alternating optimization procedure (see Section 5.5.3).

It is clearly unlikely to recover perfect labels; therefore we now describe a way of measuring the amount of information which is recovered. In order to extract meaningful cluster information from the result so-obtained, we computed the correlation matrix $Y_k \Pi_n Y_{true}$ where Y_{true} is the $n \times K$ label matrix containing actual labels and Π_n is the $n \times n$ centering matrix $I_n - \frac{1}{n} 1_n 1_n^\top$. The k predicted labels are present in the Y_k matrix. In order to choose an appropriate value of k, we plotted $\mathrm{Tr}(\Phi_{Y_{true}} \Phi_{Y_k})$ (shown in Figure 5-6 along with a K-means baseline), where $\Phi_{Y_k} = Y_k ({Y_k}^\top Y_k)^{-1} {Y_k}^\top$. From these plots, we chose $k = 30$ to be a suitable value for our interpretation purposes.

After choosing an arbitrary value of $k = 30$, we plotted the correlations between the actual and predicted labels. The heatmap of the normalized absolute correlations is given in Figure 5-7, where the columns and rows corresponding to the 80 true labels and 30 predicted labels respectively, are ordered according to the sum of squared correlations (the top-scoring labels appear to the left-bottom). From this plot, we extract following highly correlated labels: person, dining table, car, chair, cup, tennis racket, bowl, truck, fork, pizza, showing that these labels were partially recovered by

3. Dataset obtained from `http://mscoco.org/dataset`

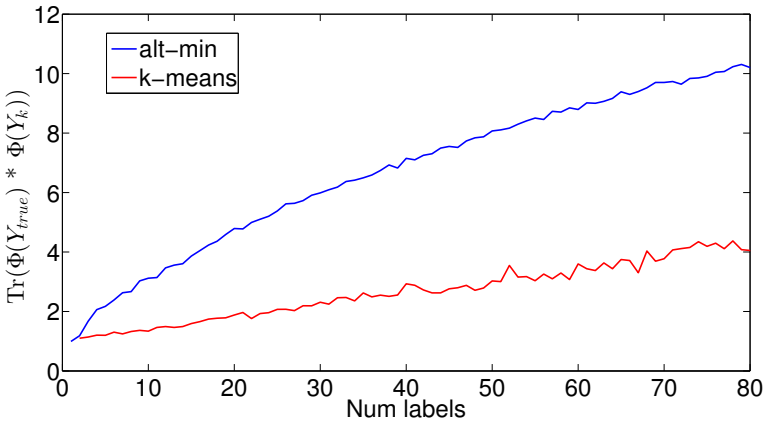


Figure 5-6 – Plot of $\mathrm{Tr}(\Phi_{Y_{true}}\Phi_{Y_k})$.

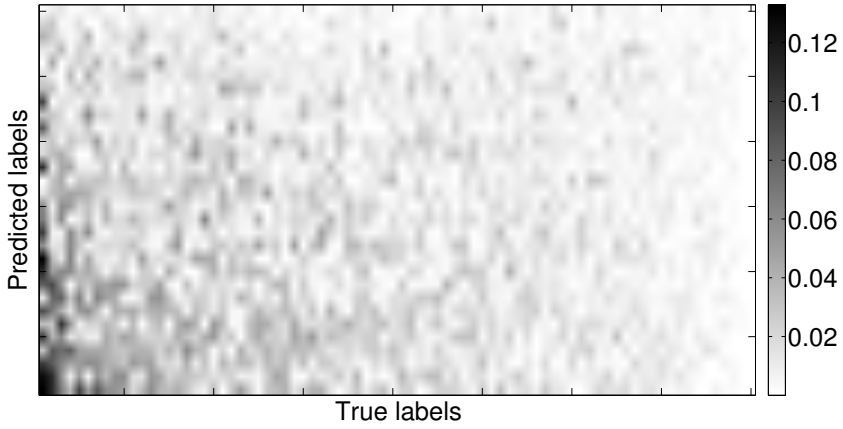


Figure 5-7 – Heatmap of correlations, $Y_k \Pi_n Y_{true}$ with $k = 30$, with columns and rows ordered according to the sum of squared correlations.

our unsupervised technique (note that the CNN features are learned with supervision on the different dataset Imagenet, hence there is still some partial supervision).

5.8 Conclusion

In this chapter, we provided a sparse extension of the discriminative clustering framework, and gave a first analysis of its theoretical performance in the totally unsupervised situation, highlighting provable scalings between ambient dimension d, number of observations and "clusterability" of irrelevant variables. We also proposed an efficient algorithm which is the first of its kind to be linear in the number of observations for discriminative clustering with the square loss. Our work could be extended in a number of ways, e.g., extending the sparse analysis to l-sparse case with higher l,extending the framework to nonlinear clustering using kernels, considering related weakly supervised learning extensions [Joulin and Bach, 2012], going beyond uniqueness of rank-one solutions, and improving the complexity of our algorithm to $O(nd)$, for example using stochastic gradient techniques.

Appendix

5.A Joint Clustering and Dimension Reduction

Given y, we need to optimize the Rayleigh quotient $\frac{w^\top X^\top y y^\top X w}{w^\top X^\top X w}$ with a rank-one matrix in the numerator, which leads to $w = (X^\top X)^{-1} X^\top y$. Given w, we will show that the averaged distortion measure of K-means once the means have been optimized is exactly equal to $(y^\top \Pi_n X w)^2 / \|\Pi_n y\|_2^2$. Given the data matrix $X \in \mathbb{R}^{n \times d}$, K-means to cluster the data into two components will tend to approximate the data points in X by the centroids $c_+ \in \mathbb{R}^d$ and $c_- \in \mathbb{R}^d$ such that

$$\begin{aligned} X &\approx \frac{(y + 1_n)}{2} c_+^\top - \frac{(y - 1_n)}{2} c_-^\top \text{ (since } y \in \{-1, 1\}^n) \\ &= \frac{y}{2}(c_+^\top - c_-^\top) + \frac{1}{2} 1_n (c_+^\top + c_-^\top). \end{aligned}$$

The objective of K-means can now be written as problem $\mathcal{KM}$:

$$\begin{aligned} &\min_{y, c_+, c_-} \left\| X - \frac{y}{2}(c_+^\top - c_-^\top) - \frac{1}{2} 1_n (c_+^\top + c_-^\top) \right\|_F^2 \\ &= \min_{y, c_+, c_-} \left\| X - \frac{(y + 1_n)}{2} c_+^\top - \frac{(1_n - y)}{2} c_-^\top \right\|_F^2 \\ &= \min_{y, c_+, c_-} \|X\|_F^2 + \|c_+^\top\|_F^2 \left\| \frac{(y + 1_n)}{2} \right\|^2 + \|c_-^\top\|_F^2 \left\| \frac{(1 - y_n)}{2} \right\|^2 \\ &\qquad + 2 c_-^\top c_+ \frac{(y + 1_n)}{2}^\top \frac{(1_n - y)}{2} - 2 \operatorname{tr} X^\top \left(\frac{(y + 1_n)}{2} c_+^\top + \frac{(1_n - y)}{2} c_-^\top \right) \\ &= \min_{y, c_+, c_-} \|X\|_F^2 + \|c_+^\top\|_F^2 \frac{1}{2}(n + 1_n^\top y) + \|c_-^\top\|_F^2 \frac{1}{2}(n - 1_n^\top y) - 2 c_+^\top X^\top \left(\frac{y + 1_n}{2} \right) \\ &\qquad - 2 c_-^\top X^\top \left(\frac{1_n - y}{2} \right). \end{aligned}$$

Fixing y and minimizing with respect to c_+ and c_-, we get closed-form expressions for c_+ and c_- as

$$c_+ = \frac{X^\top (y + 1_n)}{(n + 1_n^\top y)} \quad \text{and} \quad c_- = \frac{X^\top (1_n - y)}{(n - 1_n^\top y)}.$$

Substituting these expressions in $\mathcal{KM}$, we have the following optimization problem in y:

$$\begin{aligned}
&\min_y \|X\|_F^2 - \frac{1}{2}\frac{\|X^\top(y+1_n)\|_F^2}{(n+1_n^\top y)} - \frac{1}{2}\frac{\|X^\top(1_n-y)\|_F^2}{(n-1_n^\top y)} \\
&= \min_y \|X\|_F^2 - \frac{1}{2}\frac{\operatorname{tr} XX^\top(y+1_n)(y+1_n)^\top}{(n+1_n^\top y)} - \frac{1}{2}\frac{\operatorname{tr} XX^\top(1_n-y)(1_n-y)^\top}{(n-1_n^\top y)} \\
&= \min_y \|X\|_F^2 - \frac{2}{(n+1_n^\top y)}\operatorname{tr} XX^\top\left(\frac{y+1_n}{2}\right)\left(\frac{y+1_n}{2}\right)^\top \\
&\qquad - \frac{2}{(n-1_n^\top y)}\operatorname{tr} XX^\top\left(\frac{1_n-y}{2}\right)\left(\frac{1_n-y}{2}\right)^\top \\
&= \min_y \ \operatorname{tr} XX^\top - \frac{2}{(n+1_n^\top y)}\operatorname{tr} XX^\top\left(\frac{y+1_n}{2}\right)\left(\frac{y+1_n}{2}\right)^\top \\
&\qquad - \frac{2}{(n-1_n^\top y)}\operatorname{tr} XX^\top\left(\frac{1_n-y}{2}\right)\left(\frac{1_n-y}{2}\right)^\top \\
&= \min_y \ \operatorname{tr} XX^\top\left(I - \frac{1}{2(n+1_n^\top y)}(yy^\top + 1_n1_n^\top + y1_n^\top + 1_ny^\top)\right. \\
&\qquad \left. - \frac{1}{2(n-1_n^\top y)}(1_n1_n^\top + yy^\top - 1_ny^\top - y1_n^\top)\right).
\end{aligned}$$

By the centering of X, we have $1_n^\top X = 0$ and hence $\operatorname{tr} XX^\top 1_n1_n^\top = \operatorname{tr} XX^\top 1_ny^\top = \operatorname{tr} XX^\top y1_n^\top = 0$. Therefore, we obtain

$$\begin{aligned}
&\min_y \ \operatorname{tr} XX^\top\left(I - \frac{1}{2(n+1_n^\top y)}(yy^\top) - \frac{1}{2(n-1_n^\top y)}(yy^\top)\right) \\
&= \min_y \ \operatorname{tr} XX^\top\left(I - (yy^\top)\left(\frac{1}{2(n+1_n^\top y)} + \frac{1}{2(n-1_n^\top y)}\right)\right) \\
&= \min_y \ \operatorname{tr} XX^\top\left(I - (yy^\top)\left(\frac{n}{n^2-(1_n^\top y)^2)}\right)\right) \\
&= \min_y \ \operatorname{tr} XX^\top\left(I - \frac{nyy^\top}{n^2-(1_n^\top y)^2}\right).
\end{aligned}$$

Thus we have the equivalent K-means problem as

$$\min_{y\in\{-1,1\}^n} \frac{1}{n}\operatorname{tr} Xww^\top X^\top\left(I - \frac{n}{n^2-(y^\top 1)^2}yy^\top\right) = 1 - \max_{y\in\{-1,1\}^n}\frac{(w^\top X^\top y)^2}{n^2-(y^\top 1)^2}.$$

Thus the averaged distortion measure of K-means with the optimized means is $\frac{(y^\top \Pi_n Xw)^2}{\|\Pi_n y\|_2^2}$.

5.B Full (Unsuccessful) Relaxation

It is tempting to find a direct relaxation of Eq. (5.2). It turns out to lead to a trivial relaxation, which we outline in this section. When optimizing Eq. (5.2) with respect to w, we obtain $\max_{y\in\{-1,1\}^n} \frac{y^\top X(X^\top X)^{-1}X^\top y}{y^\top \Pi_n y}$, leading to a quasi-convex relaxation as $\max_{\substack{Y\succcurlyeq 0,\\ \mathrm{diag}(Y)=1}} \frac{\mathrm{tr}\, YX(X^\top X)^{-1}X^\top}{\mathrm{tr}\,\Pi_n Y}$. Unfortunately, this relaxation always leads to trivial solutions as described below.

Consider the quasi-convex relaxation

$$\max_{Y\succcurlyeq 0,\mathrm{diag}(Y)=1} \frac{\mathrm{tr}\, YX(X^\top X)^{-1}X^\top}{\mathrm{tr}\,\Pi_n Y}. \tag{5.27}$$

By definition of Π_n this relaxation is equal to:

$$\max_{Y\succcurlyeq 0,\mathrm{diag}(Y)=1} \frac{1}{n}\frac{\mathrm{tr}\, YX(X^\top X)^{-1}X^\top}{1-\frac{1_n^\top Y 1_n}{n^2}}.$$

Let $\mathcal{A} = \{Y \succcurlyeq 0,\ \mathrm{diag}(Y) = 1\}$ the feasible set of this problem and define $\mathcal{B} = \{M \succcurlyeq 0,\ \mathrm{diag}(M) = 1 + \frac{1_n^\top M 1_n}{n^2}\}$. Let $Y \in \mathcal{A}$, then M defined by $M = \frac{Y}{1-\frac{1_n^\top Y 1_n}{n^2}}$ belongs to $\mathcal{B}$ since $1 + \frac{1_n^\top M 1_n}{n^2} = 1 + \frac{1_n^\top Y 1_n}{n^2 - 1_n^\top Y 1_n} = \frac{1}{1-\frac{1_n^\top Y 1_n}{n^2}} = \mathrm{diag}(M)$. Reciprocally for $M \in \mathcal{B}$, we can define $Y = \frac{M}{1+\frac{1_n^\top M 1_n}{n^2}}$, such that $\mathrm{diag}(Y) = 1$ and $Y \in \mathcal{A}$ and then verify that $M = \frac{Y}{1-\frac{1_n^\top Y 1_n}{n^2}}$. Thus the problem Eq. (5.27) is equivalent to the relaxation

$$\max_{M\succcurlyeq 0,\mathrm{diag}(M)=1+\frac{1_n^\top M 1_n}{n^2}} \frac{1}{n}\,\mathrm{tr}\, MX(X^\top X)^{-1}X^\top. \tag{5.28}$$

The Lagrangian function of this problem can be written as:

$$\begin{aligned} L(\mu) &= \mathrm{tr}\, MX(X^\top X)^{-1}X^\top - \frac{\mu}{n}^\top[\mathrm{diag}(M) - 1_n - \frac{1_n^\top M 1_n}{n^2}1_n] \\ &= \mathrm{tr}\, M[X(X^\top X)^{-1}X^\top - \mathrm{Diag}(\mu) + \frac{1_n^\top \mu}{n^2}1_n 1_n^\top] + \frac{1}{n}\mu^\top 1_n. \end{aligned}$$

Using $L(\mu)$ and the PSD constraint $M \succcurlyeq 0$, the dual problem is given by

$$\min_\mu \frac{\mu^\top 1_n}{n} \quad \text{s.t. } \mathrm{Diag}(\mu) - \frac{1_n^\top \mu}{n^2}1_n 1_n^\top \succcurlyeq X(X^\top X)^{-1}X^\top.$$

Since $X(X^\top X)^{-1}X^\top \succcurlyeq 0$, this implies for the dual variable μ:

$$\mathrm{Diag}(\mu) - \frac{1_n^\top \mu}{n^2}1_n 1_n^\top \succcurlyeq 0 \quad \Leftrightarrow \quad 1_n^\top \mathrm{Diag}(\mu)^{-1}1_n \le \frac{n^2}{\mu^\top 1_n}$$

$$\Leftrightarrow \quad \sum_{i=1}^{n} \frac{1}{\mu_i} \leq \frac{n^2}{\sum_{i=1}^{n} \mu_i}$$
$$\Leftrightarrow \quad \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\mu_i} \leq \frac{1}{\frac{1}{n} \sum_{i=1}^{n} \mu_i}.$$

However for $\nu \in \mathbb{R}^n$, the harmonic mean $\left[\frac{1}{n} \sum_{i=1}^{n} \frac{1}{\nu_i}\right]^{-1}$ is always smaller than the arithmetic mean $\frac{1}{n} \sum_{i=1}^{n} \nu_i$ with equality if and only if $\nu = c1_n$ for $c \in \mathbb{R}$.

Thus the dual variable μ is constant and the diagonal constraint simplifies itself as a trace constraint. Therefore the problem is equivalent to the trivial relaxation whose each eigenvector of $X(X^\top X)^{-1}X^\top$ is solution

$$\max_{M \succcurlyeq 0,\ \operatorname{tr}(M) = n + \frac{1_n^\top M 1_n}{n}} \operatorname{tr} MX(X^\top X)^{-1}X^\top.$$

5.C Equivalent Relaxation

5.C.1 First Equivalent Relaxation

We start from the penalized version of Eq. (5.5),

$$\min_{y \in \{-1,1\}^n,\ v \in \mathbb{R}^d} \frac{1}{n} \|\Pi_n y - Xv\|_2^2 + \nu \frac{(y^\top 1_n)^2}{n^2}, \tag{5.29}$$

which we expand as:

$$\min_{y \in \{-1,1\}^n,\ v \in \mathbb{R}^d} \frac{1}{n} \operatorname{tr} \Pi_n yy^\top - \frac{2}{n} \operatorname{tr} Xvy^\top + \frac{1}{n} \operatorname{tr} X^\top Xvv^\top + \nu \frac{(y^\top 1_n)^2}{n^2}, \tag{5.30}$$

and relax as, using $Y = yy^\top$, $P = yv^\top$ and $V = vv^\top$,

$$\begin{array}{ll} \min\limits_{V,P,Y} & \frac{1}{n} \operatorname{tr} \Pi_n Y - \frac{2}{n} \operatorname{tr} P^\top X + \frac{1}{n} \operatorname{tr} X^\top XV + \nu \frac{1_n^\top Y 1_n}{n^2} \\ \text{s.t.} & \begin{pmatrix} Y & P \\ P^\top & V \end{pmatrix} \succcurlyeq 0,\ \operatorname{diag}(Y) = 1. \end{array} \tag{5.31}$$

When optimizing Eq. (5.31) with respect to V and P, we get exactly Eq. (5.8). Indeed we solve this problem by fixing the matrix Y such that $Y = Y_0$ and $\operatorname{diag}(Y_0) = 1_n$. Then the Lagrangian function of the problem in Eq. (5.31) can be written as

$$\begin{aligned} L(A) &= \frac{1}{n} \operatorname{tr} \Pi_n Y - \frac{2}{n} \operatorname{tr} P^\top X + \frac{1}{n} \operatorname{tr} X^\top XV + \nu \frac{1_n^\top Y 1_n}{n^2} + \operatorname{tr} A(Y - Y_0) \\ &= \begin{pmatrix} Y & P \\ P^\top & V \end{pmatrix} \begin{pmatrix} \frac{1}{n}\Pi_n + \frac{\nu}{n^2} 1_n 1_n^\top + A & \frac{-1}{n} X \\ \frac{-1}{n} X^\top & \frac{1}{n} X^\top X \end{pmatrix} - \operatorname{tr} AY_0. \end{aligned}$$

Using $L(A)$ and the psd constraint $\begin{pmatrix} Y & P \\ P^\top & V \end{pmatrix} \succcurlyeq 0$, we write the dual problem as

$$\min_A \operatorname{tr} AY_0 \text{ s.t. } \begin{pmatrix} \frac{1}{n}\Pi_n + \frac{\nu}{n^2}1_n 1_n^\top + A & \frac{-1}{n}X \\ \frac{-1}{n}X^\top & \frac{1}{n}X^\top X \end{pmatrix} \succcurlyeq 0.$$

From the Schur's complement condition of $\begin{pmatrix} \frac{1}{n}\Pi_n + \frac{\nu}{n^2}1_n 1_n^\top + A & \frac{-1}{n}X \\ \frac{-1}{n}X^\top & \frac{1}{n}X^\top X \end{pmatrix} \succcurlyeq 0$, we obtain $\frac{1}{n}\Pi_n + \frac{\nu}{n^2}1_n 1_n^\top + A \succcurlyeq \frac{1}{n}X(X^\top X)^{-1}X^\top$. Substituting the bound for A we get the optimal objective function value

$$\mathcal{D}^* = \frac{1}{n}\operatorname{tr} X(X^\top X)^{-1}X^\top Y_0 - \frac{1}{n}\operatorname{tr}\Pi_n Y_0 - \frac{\nu}{n^2}1_n^\top Y_0 1_n.$$

Note that the optimal dual objective value $\mathcal{D}^*$ corresponds to a fixed Y_0. Hence by maximizing with respect to Y we obtain exactly Eq. (5.8) and therefore, the convex relaxation in Eq. (5.11) is equivalent to Eq. (5.8). Moreover the Karush-Kuhn-Tucker (KKT) conditions gives

$$P^\top - X + VX^\top X = 0 \text{ and } -YX + PX^\top X = 0$$

Thus the optimum is attained for $P{=}YX(X^\top X)^{-1}$ and $V{=}(X^\top X)^{-1}X^\top YX(X^\top X)^{-1}$.

5.C.2 Second Equivalent Relaxation

For $\nu = 1$, we solve the problem in Eq. (5.31) by fixing the matrix $V = V_0$. Then the Lagrangian function of this problem can be written as

$$\begin{aligned} \hat{L}(\mu, B) &= \frac{1}{n}\operatorname{tr}\Pi_n Y - \frac{2}{n}\operatorname{tr} P^\top X + \frac{1}{n}\operatorname{tr} X^\top XV + \nu\frac{1_n^\top Y 1_n}{n^2} \\ &\quad +\mu^\top(\operatorname{diag}(Y) - 1_n) + \operatorname{tr} B(V - V_0) \\ &= \begin{pmatrix} Y & P \\ P^\top & V \end{pmatrix}\begin{pmatrix} \frac{1}{n}I_n + \operatorname{diag}(\mu) & \frac{-1}{n}X \\ \frac{-1}{n}X^\top & \frac{1}{n}X^\top X + B \end{pmatrix} - \mu^\top 1_n - \operatorname{tr} BV_0. \end{aligned}$$

Using $\hat{L}(\mu, B)$ and the psd constraint $\begin{pmatrix} Y & P \\ P^\top & V \end{pmatrix} \succcurlyeq 0$, the dual problem is given by

$$\min_{\mu,B} \mu^\top 1_n + \operatorname{tr} BV_0 \text{ s.t. } \begin{pmatrix} \frac{1}{n}I_n + \operatorname{diag}(\mu) & \frac{-1}{n}X \\ \frac{-1}{n}X^\top & \frac{1}{n}X^\top X + B \end{pmatrix} \succcurlyeq 0.$$

From the Schur's complement condition of $\begin{pmatrix} \frac{1}{n}I_n + \operatorname{diag}(\mu) & \frac{-1}{n}X \\ \frac{-1}{n}X^\top & \frac{1}{n}X^\top X + B \end{pmatrix} \succcurlyeq 0$, we obtain $B \succcurlyeq \frac{1}{n^2}X^\top \operatorname{diag}(\mu + 1_n/n)^{-1}X - \frac{1}{n}X^\top X$. Substituting the bound for B we get the dual problem as

$$\min_\mu \mu^\top 1_n + \frac{1}{n^2}\operatorname{tr} V_0 X^\top \operatorname{diag}(\mu + 1_n/n)^{-1}X - \frac{1}{n}\operatorname{tr} V_0 X^\top X$$

$$\min_{\mu} \sum_{i=1}^{n} \left(\mu_i + \frac{1}{n^2 \mu_i + n} x_i^\top V_0 x_i \right) - \frac{1}{n} \operatorname{tr} V_0 X^\top X.$$

Solving for μ_i, we get

$$\mu_i^* = \frac{1}{n} \sqrt{x_i^\top V_0 x_i} - \frac{1}{n}.$$

Substituting μ_i^* into the dual obkective function, we get the optimal objective function value

$$\hat{D} = \frac{2}{n} \sum_{i=1}^{n} \sqrt{(XVX^\top)_{ii}} - 1 - \frac{1}{n} \operatorname{tr} V_0 X^\top X.$$

Furthermore the KKT conditions gives

$$Y \operatorname{diag}(\nu + 1_n/n) - \frac{1}{n} P X^\top = 0 \text{ and } P^\top \operatorname{diag}(\nu + 1_n/n) - \frac{1}{n} V X^\top = 0.$$

Thus we obtain the following closed form expressions:

$$\begin{aligned} P &= \operatorname{Diag}(\operatorname{diag}(XVX^\top))^{-1/2} XV \\ Y &= \operatorname{Diag}(\operatorname{diag}(XVX^\top))^{-1/2} XVX^\top \operatorname{Diag}(\operatorname{diag}(XVX^\top))^{-1/2}. \end{aligned}$$

The optimal dual objective value $\hat{D}$ corresponds to a fixed V_0. Therefore, maximizing with respect to V leads to the problem:

$$\min_{V \succcurlyeq 0} \quad 1 - \frac{2}{n} \sum_{i=1}^{n} \sqrt{(XVX^\top)_{ii}} + \frac{1}{n} \operatorname{tr}(VX^\top X). \tag{5.32}$$

5.D Auxiliary Results for Section 5.5.1

5.D.1 Auxiliary Lemma

The matrix $X(X^\top X)^{-1} X^\top$ has the following properties [see, e.g., Freedman, 2009].

Lemma 34. *The matrix $H = X(X^\top X)^{-1} X^\top$ is the orthogonal projection onto the column space of the design matrix X since:*

- *H is symmetric.*
- *H is idempotent $(H^2) = H$.*
- *X is invariant under H, that is $HX = X$.*

5.D.2 Rank-One Solution of the Relaxation Eq. (5.8)

We denote by $(x_i)_{i=1\dots n}$ the lines of X.

Lemma 35. *The rank-one solution $Y_* = yy^\top$ is always solution of the relaxation Eq. (5.8).*

Proof. We give an elementary proof of this result without using convex optimization tools. Using lemma 34 we have $Hy = y$, thus

$$\operatorname{tr} HY_* = \operatorname{tr} Hyy^\top = \operatorname{tr} yy^\top = n.$$

Moreover all $M \succcurlyeq 0$ can always be decomposed as $\sum_{i=1}^n \lambda_i u_i u_i^\top$ with $\lambda_i \geq 0$ and $(u_i)_{i=1,\dots,n}$ an orthonormal family. Since H is an orthogonal projection $(u_i)^\top H u_i = (Hu_i)^\top H u_i = \|Hu_i\|^2 \leq \|u_i\|^2 \leq 1$. Thus

$$\begin{aligned} \operatorname{tr} HM &= \sum_{i=1}^n \lambda_i \operatorname{tr} Hu_i(u_i)^\top = \sum_{i=1}^n \lambda_i (u_i)^\top H u_i \\ &\leq \sum_{i=1}^n \lambda_i = \operatorname{tr} M. \end{aligned}$$

Then for all matrix M feasible we have $\operatorname{tr} HM \leq n$ since $\operatorname{diag}(M) = 1_n$ and $\operatorname{tr} HY_* = n$ which conclude the lemma. □

5.D.3 Rank-One Solution of the Relaxation Eq. (5.12)

Lemma 36. *The rank-one solution $V_* = vv^\top$ is always solution of the relaxation Eq. (5.12).*

Proof. The Karush-Kuhn-Tucker (KKT) optimality conditions for the problem are for the dual variable $A \preccurlyeq 0$:

$$\frac{1}{n}\sum_{i=1}^n \frac{x_i x_i^\top}{\sqrt{x_i^\top V x_i}} - \frac{1}{n} X X^\top = A \text{ and } AV = 0 \quad \text{(Complementary Slackness)}.$$

Since $x_i^\top w = y_i$, $\sqrt{x_i^\top V_* x_i} = |y_i| = 1$, V_* and the dual variable $A = 0$ satisfy the KKT conditions and then V_* is solution of this problem. □

5.D.4 Proof of Proposition 19

In the following lemma, we use a Taylor expansion to lower-bound f around its minimum.

Lemma 37. *For $d \geq 3$ and $\delta \in [0, 1)$.*

If $\beta \geq 3$ and $m^2 \leq \frac{\beta-3}{2(d+\beta-4)}$, then with probability at least $1 - d\exp\big(-\frac{\delta^2 n m^2}{2R^4 d^2}\big)$, for any symmetric matrix Δ:

$$f(V_*) - f(V_* + \Delta) > 2(1-\delta)m^2\|\Delta\|_F^2 + o(\|\Delta\|^2) \geq 0.$$

Otherwise with probability at least $1 - d\exp\big(-\frac{\delta^2 n \mu_1}{4R^4 d^2}\big)$, for any symmetric matrix Δ:

$$f(V_*) - f(V_* + \Delta) > (1-\delta)\mu_1\|\Delta\|_F^2 + o(\|\Delta\|^2) \geq 0,$$

with $\mu_1 \geq \frac{m^2(\beta-1)}{1+(d+\beta-2)m^2}$. *Moreover we also have with probability at least* $1 - d\exp\big(-\frac{\delta^2 n\mu_2}{4R^4d^2}\big)$, *for any symmetric matrix* $\Delta \in \Delta_{\min}^{\perp}$:

$$f(V_*) - f(V_* + \Delta) > (1-\delta)\mu_2\|\Delta\|_F^2 + o(\|\Delta\|^2) \geq 0,$$

where $\mu_2 = \min\{2m^2, m^2(\beta-1), 2m\}$ *and* $\Delta_{\min} = \begin{pmatrix} 1 & 0 \\ 0 & c_{\min} I_{d-1} \end{pmatrix}$ *is defined in the proof and satisfies*

$$|c_{\min}| \leq \frac{m}{|(d+\beta-2)m^2 - 1|}.$$

This lemma directly implies Proposition 19.

Proof. For $\Delta \in \mathcal{S}(d)$ and $\delta \in \mathbb{R}$ we compute for $f(V) = \frac{1}{n}\sum_{i=1}^n \sqrt{x_i^\top V x_i}$,

$$\frac{d^2}{d\delta^2} f(V + \delta\Delta) = -\frac{1}{4n}\sum_{i=1}^n \frac{(x_i^\top \Delta x_i)^2}{\sqrt{x_i^\top (V+\delta\Delta)x_i}^3}.$$

Thus the second directional derivative in $V = V_*$ along Δ is

$$\nabla_\Delta^2 f(V_*) = \lim_{\delta\to 0} \frac{d^2}{d\delta^2} f(V+\delta\Delta) = -\frac{1}{4n}\sum_{i=1}^n (x_i^\top \Delta x_i)^2.$$

Let $\mathcal{T}_x$ be the semidefinite positive quadratic form of $\mathcal{S}(d)$ defined for $\Delta \in \mathcal{S}(d)$, by

$$\mathcal{T}_x : \Delta \mapsto (x^\top \Delta x)^2. \tag{5.33}$$

Then it exists a positive linear operator T_x from $\mathcal{S}(d)$ to $\mathcal{S}(d)$ such that $\mathcal{T}_x(\Delta) = \langle \Delta, T_x\Delta\rangle$.

Therefore the function f will be stricly concave if for all directions $\Delta \in \mathcal{S}(d)$

$$\frac{1}{n}\sum_{i=1}^n \mathcal{T}_{x_i}(\Delta) > 0. \tag{5.34}$$

We will bound the empirical expectation in Eq. (5.34) by first showing that its expectation remains away from 0. Then we will use a concentration inequality for matrices to control the distance between the sum in Eq. (5.34) and its expectation.

We first derive conditions so that the result is true in expectation, i.e. for the operator $\mathcal{T}$ defined by $\mathcal{T} = \mathbb{E}\mathcal{T}_x$ for x following the same law as $(y, z^\top)^\top$. We denote by $m = \mathbb{E}z^2$ and by $\beta = \mathbb{E}z^4/m^2$ its kurtosis.

We let $\Delta = \begin{pmatrix} a & b^\top \\ b & C \end{pmatrix}$ and then have $x^\top\Delta x = a + 2yb^\top z + z^\top C z$. Thus

$$\mathcal{T}_x(\Delta) = a^2 + 4ayb^\top z + 2az^\top Cz + 4b^\top(zz^\top)b + (z^\top Cz)^2 + 4yb^\top z(z^\top Cz).$$

Therefore we can express the value of the operator $\mathcal{T}$ only in function of the elements

of Δ:

$$\mathcal{T}(\Delta) = (a + m \operatorname{tr} C)^2 + 4m\|b\|_2^2 + 2m^2\|C - \operatorname{Diag}(\operatorname{diag}(C))\|_F^2 + m^2(\beta - 1)\|\operatorname{diag}(C)\|^2,$$

where we have used

$$\begin{aligned}
\mathbb{E}(z^\top C z)^2 &= \mathbb{E}\sum_{i,j,k,l} z_i z_j z_k z_l c_{i,j} c_{k,l} \\
&= \mathbb{E}\sum_i (z_i)^4 c_{i,i}^2 + \mathbb{E}\sum_{i,k\neq i} z_i^2 z_k^2 c_{i,i} c_{k,k} + 2\mathbb{E}\sum_{i,j\neq i} z_i^2 z_j^2 c_{i,j}^2 \\
&= \beta m^2 \sum_i c_{i,i}^2 + m^2 \sum_{i,k\neq i} c_{i,i} c_{k,k} + 2m^2 \sum_{i,j\neq i} c_{i,j}^2 \\
&= m^2(\beta - 3)\sum_i c_{i,i}^2 + m^2 \sum_{i,k} c_{i,i} c_{k,k} + 2m^2 \sum_{i,j} c_{i,j}^2 \\
&= m^2(\beta - 3)\|\operatorname{diag}(C)\|^2 + m^2\big(2\|C\|_F^2 + \operatorname{tr}(C)^2\big) \\
&= m^2(\beta - 3)\|\operatorname{diag}(C)\|^2 + m^2\big(2\|C - \operatorname{Diag}(\operatorname{diag}(C))\|_F^2 + \operatorname{tr}(C)^2\big).
\end{aligned}$$

Since $\beta \geq 1$, we get

$$\mathcal{T}(\Delta) \geq (a + m \operatorname{tr} C)^2 + 4m\|b\|_2^2 + 2m^2(\|C\|_F^2 - \|\operatorname{diag}(C)\|^2).$$

Thus $\mathcal{T}(\Delta) = 0$ if and only if $\beta = 1$ with $b = 0_{d-1}$ and $C = \operatorname{diag}(c)$ with $c^\top 1_d = -\frac{a}{m_2}$. With the condition $\beta = 1$ meaning that $\operatorname{var}(z^2) = 0$ and thus z^2 is constant a.s., i.e. z follows a Rademacher law.

However we would like to bound $\mathcal{T}(\Delta)$ away from zero by some constant and for that we are looking for the smallest eigenvalue of the operator $\mathbb{E}T_x$. Unfortunately we are not able to solve the optimization problem

$$\min_{\Delta \in \mathcal{S}(d), \|\Delta\|_F^2 = 1} \mathcal{T}(\Delta),$$

and we have to compute all the spectrum of this operator to be able to find the smallest using $\mathbb{E}T_x\Delta = 1/2\nabla\mathcal{T}(\Delta)$.

We have

$$1/2\nabla\mathcal{T}(\Delta) = \begin{pmatrix} a + m\operatorname{tr}(C) & 2mb^\top \\ 2mb & (a + m\operatorname{tr}(C))m_2 I_{d-1} + 2m^2 C \\ & +m^2(\beta - 3)\operatorname{Diag}(\operatorname{diag}(C)) \end{pmatrix}.$$

— For all $b \in \mathbb{R}^{d-1}$ we have for $\Delta = \begin{pmatrix} 0 & b^\top \\ b & 0 \end{pmatrix}$, $1/2\nabla\mathcal{T}(\Delta) = 2m\Delta$. Thus $2m$ is an eigenvalue of multiplicity $d - 1$.

— For all $C \in \mathbb{R}^{(d-1)\times(d-1)}$ with $\operatorname{diag}(C) = 0_{d-1}$ we have for $\Delta = \begin{pmatrix} 0 & 0 \\ 0 & C \end{pmatrix}$, $1/2\nabla\mathcal{T}(\Delta) = 2m^2\Delta$. Thus $2m^2$ is an eigenvalue of multiplicity $\frac{(d-1)(d-2)}{2}$.

— For all $c \in \mathbb{R}^{d-1}$ with $c^\top 1_{d-1} = 0$ we have for $\Delta = \begin{pmatrix} 0 & 0 \\ 0 & \mathrm{diag}(C) \end{pmatrix}$, $1/2\nabla\mathcal{T}(\Delta) = m^2(\beta-1)\Delta$. Thus $m^2(\beta-1)$ is an eigenvalue of multiplicity $d-2$.

— For all $a, c \in \mathbb{R}^2$ we have for $\Delta = \begin{pmatrix} a & 0 \\ 0 & cI_{d-1} \end{pmatrix}$,

$$\begin{aligned} 1/2\nabla\mathcal{T}(\Delta) &= \begin{pmatrix} a + m(d-1)c & 0 \\ 0 & [ma + m^2(d+\beta-2)c]I_{d-1} \end{pmatrix} \\ &= \mathrm{Diag}\Big[\begin{pmatrix} 1 & m1_{d-1}^\top \\ m1_{d-1} & (d+\beta-2)m^2 I_{d-1} \end{pmatrix} \begin{pmatrix} a \\ c1_{d-1} \end{pmatrix} \Big]. \end{aligned}$$

Thus an eigenvalue of $\begin{pmatrix} 1 & (d-1)m \\ m & (d+\beta-2)m^2 \end{pmatrix}$ with an eigenvector $[a, c]^\top$ would be an eigenvalue of the operator $\mathbb{E}T_x$ with a corresponding eigenvector $\begin{pmatrix} a & 0 \\ 0 & cI_{d-1} \end{pmatrix}$. This matrix has two simple eigenvalues

$$\mu_\pm = \frac{1 + (d+\beta-2)m^2 \pm \sqrt{(1+(d+\beta-2)m^2)^2 - 4m^2(\beta-1)}}{2}. \tag{5.35}$$

Moreover when we add all the multiplicity of the found eigenvalues we get $d - 1 + \frac{(d-1)(d-2)}{2} + d - 2 + 2 = \frac{d(d+1)}{2}$ which is the dimension of $S(d)$, therefore we have found all the eigenvalues of the linear operator $\mathbb{E}T_x$.

We will prove now than the smallest eigenvalue is μ_- when the dimension d is large enough with regards to m^2 and $2m^2$ otherwise.

Lemma 38. *Let μ_1 and μ_2 be the two smallest eigenvalues of the operator $\mathbb{E}T_x$. Let us assume that $d \geq 3$ (the case $d = 2$ will also be done in the proof).*

If $\beta \geq 3$ and $m^2 \leq \frac{\beta-3}{2(d+\beta-4)}$ then

$$\mu_1 = 2m^2.$$

Otherwise

$$\mu_1 = \mu_- \geq \frac{m^2(\beta-1)}{1+(d+\beta-2)m^2} \text{ and } \mu_2 = \min\{2m^2, m^2(\beta-1), 2m\}.$$

Moreover we denote by $\Delta_{\min} = \begin{pmatrix} 1 & 0 \\ 0 & c_{\min} I_{d-1} \end{pmatrix}$ the eigenvector associated to μ_- for which we have set without loss of generality the first component $a = 1$. Then

$$|c_{\min}| \leq \frac{m}{|(d+\beta-2)m^2 - 1|}.$$

Unfortunately μ_- can become small when the dimension increases as explained by the tight bound $\mu_- \geq \frac{m^2(\beta-1)}{1+(d+\beta-2)m^2}$. However the corresponding eigenvector have a particular structure we will be able to exploit.

Proof. First we note that $\mu_- \leq m^2(\beta-1)$ and compute

$$\begin{aligned}
\mu_- \geq 2m^2 &\Leftrightarrow 1+(d+\beta-2)m^2 - \sqrt{(1+(d+\beta-2)m^2)^2 - 4m^2(\beta-1)} - 4m^2 \geq 0 \\
&\Leftrightarrow 1+(d+\beta-2)m^2 - 4m^2 \geq \sqrt{(1+(d+\beta-2)m^2)^2 - 4m^2(\beta-1)} \\
&\Leftrightarrow (1+(d+\beta-2)m^2-4m^2)^2 \geq (1+(d+\beta-2)m^2)^2 - 4m^2(\beta-1) \\
&\quad \text{and } 1+(d+\beta-6)m^2 \geq 0 \\
&\Leftrightarrow 16m^4 - 8m^2(1+(d+\beta-2)m^2) \geq -4m^2(\beta-1) \\
&\quad \text{and } 1+(d+\beta-6)m^2 \geq 0 \\
&\Leftrightarrow 2(d+\beta-4)m^2 \leq \beta-3 \text{ and } 1+(d+\beta-6)m^2 \geq 0.
\end{aligned}$$

— If $d=2$,
  — If $\beta \leq 3$ we have necessary that $\beta \leq 2$ and the first equation gives $m^2 \geq \frac{3-\beta}{2(2-\beta)}$ and the second $m^2 \leq 1/(4-\beta)$. Thus we should have $(4-\beta)(3-\beta) \leq 2(2-\beta)$ which is not possible since the polynomial $\beta^2 - 5\beta + 8 \geq 0$.
  — If $\beta \geq 3$, the first equation gives $m^2 \leq \frac{\beta-3}{2(\beta-2)} \leq 1$ and the second $m^2 \leq 1/(4-\beta) \leq \frac{\beta-3}{2(\beta-2)} \leq 1$ for $\beta \leq 4$ and is always satisfied otherwise.
— If $d \geq 3$, the first equation implies that $\beta \geq 3$ for which the second equation is always satisfied. It also implies that $m^2 \leq \frac{\beta-3}{2(d+\beta-4)} \leq 1$.

We denote by $\Delta_{\min} = \begin{pmatrix} 1 & 0 \\ 0 & c_{\min} I_{d-1} \end{pmatrix}$ the eigenvector for which we have set without loss of generality $a=1$ and

$$c_{\min} = \frac{-1}{2(d-1)m}\Big[\sqrt{((d+\beta-2)m^2-1)^2 + 4(d-1)m^2} - (d+\beta-2)m^2 + 1\Big].$$

Consequently $c_{\min} \leq 0$ and by convexity of the square root we have

$$\sqrt{((d+\beta-2)m^2-1)^2 + 4(d-1)m^2} \leq ((d+\beta-2)m^2-1) + \frac{2(d-1)m^2}{|(d+\beta-2)m^2-1|}.$$

Therefore

$$|c_{\min}| \leq \frac{m}{|(d+\beta-2)m^2-1|}.$$

□

We will control now the behavior of the empirical expection by its expectation thanks to concentration theory. By definition T_x is a symmetric positive linear operator as its projection $T_x^\perp$ onto the orthogonal space of $\Delta_{\min}$. We can thus apply the Matrix Chernoff inequality from Tropp [2012, Theorem 5.1.1] to these two operators using $\|T_x\|_{op} \leq \|xx^\top\|^2 \leq \mathrm{tr}(xx^\top)^2 \leq \|x\|_2^4 \leq R^4 d^2$ Then:

$$\mathbb{P}\left(\lambda_{\min}\Big(\sum_{k=1} T_{x_k}\Big) \leq n\delta\mu_1\right) \leq d\Big[\frac{e^{-(1-\delta)}}{\delta^\delta}\Big]^{n\mu_1/(2R^4d^2)} \leq de^{-(1-\delta)^2 n\mu_1/(4R^4d^2)},$$

$$\mathbb{P}\left(\lambda_{\min}\Big(\sum_{k=1} T_{x_k}^{\perp}\Big) \leq n\delta\mu_2\right) \leq d\Big[\frac{e^{-(1-\delta)}}{\delta^{\delta}}\Big]^{n\mu_2/(2R^4d^2)} \leq de^{-(1-\delta)^2 n\mu_2/(4R^4d^2)},$$

For $m = 1$ and $d \geq 3$ we have $\mu_1 = \mu_- \geq \frac{\beta-1}{\beta+d} \geq \min\{\frac{\beta-1}{2\beta}, \frac{\beta-1}{2d}\} \geq \min\{1/3, \frac{\beta-1}{2d}\}$. □

5.D.5 Noise Robustness for the 1-Dimensional Balanced Problem

We want a condition on ε such that the solution of the relaxation recovers the right y. We recall the dual problem of the relaxation Eq. (5.8)

$$\min \mu^\top 1_n \text{ s.t. } \operatorname{Diag}(\mu) \succcurlyeq X(X^\top X)^{-1}X^\top.$$

The KKT conditions are:

— Dual feasibility: $\operatorname{Diag}(\mu) \succcurlyeq X(X^\top X)^{-1}X^\top$.

— Primal feasibility: $\operatorname{Diag}(Y) = 1_n$ and $Y \succcurlyeq 0$.

— Complimentary slackness : $Y[\operatorname{Diag}(\mu) - X(X^\top X)^{-1}X^\top] = 0$

For $Y = yy^\top$ a rank one matrix, the last condition implies $\operatorname{Diag}(\mu)y = Hy$ and

$$\mu_i = \frac{(X(X^\top X)^{-1}X^\top y)_i}{y_i}.$$

For $X = y+\varepsilon$, we denote by $\tilde{y} = y+\varepsilon$, then $X(X^\top X)^{-1}X^\top = \frac{\tilde{y}\tilde{y}^\top}{\|\tilde{y}\|^2}$ and $X(X^\top X)^{-1}X^\top y = \frac{\tilde{y}^\top y}{\|\tilde{y}\|^2}\tilde{y}$. Thus

$$\mu_i = \frac{\tilde{y}^\top y}{\|\tilde{y}\|^2}\frac{\tilde{y}_i}{y_i}.$$

Assume that all $\tilde{y}_i y_i$ have the same sign, without loss of generality we assume $\tilde{y}_i y_i > 0$. By definition of μ, $\mu \geq 0$. To show the dual feasibility we have to show that $\operatorname{Diag}(\mu) \succcurlyeq H$ which is equivalent to $\operatorname{Diag}(\frac{\tilde{y}_i}{y_i}) \succcurlyeq \frac{\tilde{y}\tilde{y}^\top}{\tilde{y}^\top y}$, to $I_n - \operatorname{Diag}(\sqrt{\frac{y_i}{\tilde{y}_i}})\frac{\tilde{y}\tilde{y}^\top}{\tilde{y}^\top y}\operatorname{Diag}(\sqrt{\frac{y_i}{\tilde{y}_i}}) \succcurlyeq 0$ and to $\sum y_i\tilde{y}_i \leq \tilde{y}^\top y$ which is obviously true. Reciprocally if μ is dual feasible then $\operatorname{Diag}(\mu) \succcurlyeq 0$ and all the $\tilde{y}_i y_i$ have the same sign.

Therefore we have shown that y is solution of the relaxation Eq. (5.8) if and only if all the $\tilde{y}_i y_i$ have the same sign. If ε and y are independent this is equivalent to $\|\varepsilon\|_\infty \leq 1$ a.s.

5.D.6 The Rank-One Candidates Are Not Solutions of the Relaxation

We assume now that $1_n^\top y \neq 0$ thus $y \neq \Pi_n y$, which means we do not have the same proportion in the two clusters. Let us assume that $\Pi_n y$ takes two values $\{\pi y_-, \pi y_+\}$ that is by definition of Π_n $\pi y_+ = 1 - \frac{1_n^\top y}{n}$ and $\pi y_- = -1 - \frac{1_n^\top y}{n}$. For V_* defined as

before, we get $x_i^\top V_* x_i = (\pi y_i)^2$ and with $I_\pm$ the set of indices such that $\Pi_n y_i = \pi y_\pm$, the KKT conditions for $V = V_*$ can be written as

$$\frac{1}{n}\Big[\sum_{i\in I_+}\Big(\frac{1}{\pi y_+}-1\Big)x_i x_i^\top + \sum_{i\in I_-}\Big(\frac{1}{-\pi y_-}-1\Big)x_i x_i^\top\Big] = A_n \preccurlyeq 0 \text{ and } A_n V_* = 0.$$

We check that with $n_\pm = \#\{I_\pm\}$:

$$\begin{aligned} w^\top A_n w = 0 &= \sum_{i\in I_+}\Big(\frac{1}{\pi y_+}-1\Big)(\pi y_+)^2 + \sum_{i\in I_-}\Big(\frac{1}{-\pi y_-}-1\Big)(\pi y_-)^2 \\ &= n_+\Big(\frac{1}{\pi y_+}-1\Big)(\pi y_+)^2 + n_-\Big(\frac{1}{-\pi y_-}-1\Big)(\pi y_-)^2 \\ &= n_+\pi y_+ - n_-\pi y_- - \big(n_+(\pi y_+)^2 + n_-(\pi y_-)^2\big) \\ &= y^\top \Pi_n y - (\Pi_n y)^\top \Pi_n y = y^\top \Pi_n y - y^\top \Pi_n y = 0. \end{aligned}$$

And $A_n = \frac{1}{2n}\big[\sum_{i\in I_+}\alpha_+ x_i x_i^\top + \sum_{i\in I_-}\alpha_- x_i x_i^\top\big]$ with $\alpha_+ = \big(\frac{1}{\pi y_+}-1\big)$ and $\alpha_- = \big(\frac{1}{-\pi y_-}-1\big)$. Unfortunately $\alpha_+\alpha_- \leq 0$, and A_n is not necessary negative. Even worse we will show that $\mathbb{E}A$ is not semi-definite negative which will conclude the proof since by the law of large number $\lim_{n\to\infty}\frac{1}{n}A_n = \mathbb{E}A$. Assume that the proportions of the two clusters stay constant with $n_\pm = \rho_\pm n$, then

$$\mathbb{E}A = \rho_+\alpha_+\begin{pmatrix}(\pi y_+)^2 & 0\\ 0 & I\end{pmatrix} + \rho_-\alpha_-\begin{pmatrix}(\pi y_-)^2 & 0\\ 0 & I\end{pmatrix}.$$

And $\rho_+\alpha_+(\pi y_+)^2 + \rho_-\alpha_-(\pi y_-)^2 = 0$ since $w^\top A_n w = 0$. Then

$$\begin{aligned} \rho_+\alpha_+ + \rho_-\alpha_- &= \frac{\rho_+\pi y_- - \rho_-\pi y_+ - \pi y_+\pi y_-}{\pi y_+\pi y_-} \\ &= \frac{-(\rho_+ + \rho_-) - \frac{1_n^\top y}{n}(\rho_+ - \rho_-) + (1 - (1_n^\top y)^2)}{-(1-(\frac{1_n^\top y}{n})^2)} \\ &= \frac{\frac{1_n^\top y}{n}(\rho_+ - \rho_-) + (\frac{1_n^\top y}{n})^2)}{(1-(\frac{1_n^\top y}{n})^2)} = \frac{2(\frac{1_n^\top y}{n})^2}{(1-(\frac{1_n^\top y}{n})^2)} \geq 0. \end{aligned}$$

Thus $A = \frac{2(1_n^\top y)^2}{(n^2-(1_n^\top y)^2)}\begin{pmatrix}0 & 0\\ 0 & I\end{pmatrix}$ is not semi-definite negative and V_* is not solution of the relaxation Eq. (5.12).

5.E Auxiliary Results for Sparse Extension

5.E.1 There Is a Rank-One Solution of the Relaxation Eq. (5.18)

Lemma 39. *The rank-one solution $V_* = v^* v^{*\top}$ is solution of the relaxation Eq. (5.18) if the design matrix X is such that $\frac{1}{n}X^\top X$ has all its diagonal entries less than one.*

Proof. The KKT conditions are

$$\frac{1}{n}\sum_{i=1}^{n}\frac{x_i x_i^\top}{\sqrt{x_i^\top W x_i}} - \lambda U - \frac{1}{n}X^\top X = A \preccurlyeq 0 \text{ and } AW = 0,$$

with U such that $U_{ij} = \text{sign}(W_{ij})$ if $W_{ij} \neq 0$ and $U_{ij} \in [-1, 1]$ otherwise. For $V_* = v^* v^{*\top}$ this gives

$$A = \frac{(1+\lambda)}{n}X^\top X - \lambda U - \frac{1}{n}X^\top X = \lambda\Big[\frac{X^\top X}{n} - U\Big],$$

with $U_{1,1} = 1$, and $U_{i,j} \in [-11]$ otherwise. We check that $AV_* = 0$. If the design matrix X satsifies assumption (A1), we can choose a sub-gradient U such that the dual variable $A = 0$ and thus V_* is solution. Otherwise by property of semi-definite matrices, there is a diagonal entry of $\frac{1}{n}X^\top X$ which is bigger than 1 which prevents A to be semi-definite negative since the corresponding diagonal entry of $\frac{X^\top X}{n} - U$ will be positive. This shows that V_* does not solve the problem. □

5.E.2 Proof of Proposition 20

Lemma 40. *For $\delta \in [0, 1)$, with probability $1 - 5d^2 \exp\big(-\frac{\delta^2 n(\beta-1)}{2dR^4(1/m^2+\beta+d)}\big)$, for any direction Δ such that $V_* + \Delta \succcurlyeq 0$, we have:*

$$\begin{aligned} g(V_*) - g(V_* + \Delta) \;>\; & (1-\delta)\Big[\lambda\|\Delta - \text{Diag}(\Delta)\|_1 + \frac{\beta-1}{\beta+d+1/m^2}\frac{(1+\lambda)^3}{4}\|\text{Diag}(\Delta)\|_2^2\Big] \\ & + o(\|\Delta\|^2) \\ \geq\; & 0. \end{aligned}$$

Moreover we also have with probability at least $1 - 5d^2 \exp\big(-\frac{\delta^2 n m^2(\beta-1)}{2dR^4}\big)$, for any symmetric matrix Δ such that $V_* + \Delta \succcurlyeq 0$ and $\text{Diag}(\Delta) \in (e_{\min})^\perp$:

$$g(V_*) - g(V_* + \Delta) > (1-\delta)\Big[\lambda\|\Delta - \text{Diag}(\Delta)\|_1 + m^2(\beta-1)\frac{(1+\lambda)^3}{4}\|\text{Diag}(\Delta)\|_2^2\Big] + o(\|\Delta\|^2) \geq 0.$$

where $e_{\min} = [1, c_{\min} 1_{d-1}]$ is defined in the proof and satisfies

$$|c_{\min}| \leq \frac{m}{|(d+\beta-2)m^2 - 1|}.$$

Proof Outline

We will investigate under which conditions on X the solution is unique, first for a deterministic design matrix. We make the following deterministic assumptions on X for $\delta, \zeta \geq 0$ and $\mathcal{S} \subset \mathbb{R}^d$:

(A1) $\|\frac{X^\top X}{n}\|_\infty \leq 1$ **(A3)** $\|\frac{Z^\top Z}{n} - \mathrm{Diag}(\mathrm{diag}(\frac{1}{n}Z^\top Z))\|_\infty \leq \delta$
(A2) $\|\frac{Z^\top y}{n}\|_\infty \leq \delta$ **(A4)** $\lambda^{\mathcal{S}}_{\min}\big(\frac{X^{\odot 2}(X^{\odot 2})^\top}{n}\big) \geq \zeta > 0.$

Where we denoted by $\odot$ the Hadamard (i.e., pointwise) product between matrices and $\lambda^{\mathcal{S}}_{\min}$ the minimum eigenvalue of a linear operator restricted to a subspace $\mathcal{S}$. Then with $g(V) = \frac{2}{n}\sum_{i=1}^n \sqrt{x_i^\top V x_i} - \lambda\|V\|_1 - \frac{1}{n}\operatorname{tr} X^\top XV$, we can certify that g will decrease around the solution V_*.

Lemma 41. *Let us assume that the noise matrix verifies assumption (A1,A2,A3,A4), then for all direction Δ such that $V_* + \Delta \succcurlyeq 0$ and $\mathrm{diag}(\Delta) \in \mathcal{S}$ we have:*

$$g(V_*) - g(V_* + \Delta) \geq \lambda(1-\delta)\|\Delta - \mathrm{Diag}(\mathrm{diag}(\Delta))\|_1 + \zeta\frac{(1+\lambda)^3}{4}\|\mathrm{Diag}(\Delta)\|_2^2 + o(\|\Delta\|^2) > 0.$$

Let us assume now that $(z^i)_{i=1,.,d}$ are i.i.d of law z symmetric with $\mathbb{E}z = \mathbb{E}z^3 = 0$, $\mathbb{E}z^2 = m = 1$, $\mathbb{E}z^4/(\mathbb{E}z^2)^2 = \beta$ and such that $\|z\|_\infty$ is a.s. bounded by $0 \leq R \leq 1$. Then the matrix X satisfies a.s. assumption (A1). Using multiple Hoeffding's inequalities we have

Lemma 42. *If z does not follow a Rademacher law, the design matrix X satsifies assumptions (A1,A2,A3,A4) with probability greater than $1 - 8d^2\exp\big(-\frac{\delta^2 n(\beta-1)}{2d(\beta+d)R^4}\big)$ for $\mathcal{S} = \mathbb{R}^d$, and with probability greater than $1 - 8d^2\exp\big(-\frac{\delta^2 n\min\{\beta-1,2\}}{2dR^4}\big)$ for $\mathcal{S} = [1, c_{\min}1_{d-1}]^\perp$ where c_{min} is defined in the proof and satisfies*

$$|e_{\min}| \leq \frac{1}{d+\beta-3}.$$

This lemma concludes the proof of proposition 20. We will now prove these two lemmas.

Proof of Lemma 41

Proof. Since the dual variable A for the PSD constraint is 0 (see the proof of lemma 39), this constraint $W \succcurlyeq 0$ is not active and we will show that the function decreases in a set of directions Δ which include the one for which $V_* + \Delta \succcurlyeq 0$.

Therefore we consider a direction $\Delta = \begin{pmatrix} a & b^\top \\ b & C \end{pmatrix}$, with $C \succcurlyeq 0$, which is slightly more general than $V_* + \Delta \succcurlyeq 0$. We denote by $f(W) = \frac{2}{n}\sum_{i=1}^n \sqrt{x_i^\top W x_i} - \frac{1}{n}\operatorname{tr} X^\top XW$ the smooth part of g. By Taylor-Young, we have for all W:

$$f(W) - f(W+\Delta) = -\langle f'(W), \Delta\rangle - \frac{1}{2}\langle \Delta, f''(W)\Delta\rangle + o(\|\Delta\|^2).$$

Thus:

$$g(W) - g(W+\Delta) = -\langle f'(W), \Delta\rangle - \frac{1}{2}\langle \Delta, f''(W)\Delta\rangle + \lambda(\|W+\Delta\|_1 - \|W\|_1) + o(\|\Delta\|^2).$$

In $W = V_*$ this gives with $X^\top X = \begin{pmatrix} n & y^\top Z \\ Z^\top y & Z^\top Z \end{pmatrix}$,

$$\begin{aligned} g(W) - g(W+\Delta) &= -\lambda\langle \frac{X^\top X}{n}, \Delta\rangle - \frac{1}{2}\langle \Delta, f''(V_*)\Delta\rangle \\ &\quad +\lambda(a + 2\|b\|_1 + \|C\|_1) + o(\|\Delta\|^2) \\ &= \lambda\big[2(\|b\|_1 - \frac{1}{n} b^\top Z^\top y) + \|C\|_1 - \frac{1}{n}\operatorname{tr}(Z^\top Z C)\big] \\ &\quad -\frac{1}{2}\langle \Delta, f''(V_*)\Delta\rangle + o(\|\Delta\|^2). \end{aligned}$$

And with Hölder's inequality and assumption (A2)

$$\|b\|_1 - \frac{1}{n} b^\top Z^\top y \geq \|b\|_1 (1 - \|\frac{1}{n} Z^\top y\|_\infty) \geq (1-\delta)\|b\|_1.$$

Nevertheless we will show in lemma 43 that $\|C\|_1 - \frac{1}{n}\operatorname{tr}(Z^\top Z C) \geq (1-\delta)\|C - \operatorname{diag}(C)\|_1$, thus

$$g(W) - g(W+\Delta) \geq \lambda(1-\delta)(2\|b\|_1 + \|C - \operatorname{diag}(C)\|_1) + o(\|\Delta\|^2). \tag{5.36}$$

However in Eq. (5.36), $g(W) - g(W+\Delta) = 0$ for $b = 0$ and C diagonal, therefore we have to investigate second order conditions, i.e. to show for $\Delta = \operatorname{diag}(e)$ with $e \in \mathbb{R}^d$ that $-\langle \Delta, f''(V_*)\Delta\rangle > 0$.

And with assumption (A4)

$$\begin{aligned} -\frac{4}{(1+\lambda)^3}\langle \operatorname{diag}(e), f''(V_*)\operatorname{diag}(e)\rangle &= \frac{1}{n}\sum_{i=1}^n (x_i^\top \operatorname{diag}(e) x_i)^2 \\ &= \frac{1}{n}\sum_{i=1}^n (\sum_{j=1}^d e_j (x_i^j)^2)^2 \\ &= \frac{1}{n}\sum_{i=1}^n e^\top [x_i^{\odot 2}(x_i^{\odot 2})^\top] e \\ &\geq \lambda_{\min}\big(\frac{X^{\odot 2}(X^{\odot 2})^\top}{n}\big)\|e\|^2 \geq \zeta\|e\|_2^2. \end{aligned}$$

Thus we can conclude:

$$g(W) - g(W+\Delta) \geq \lambda(1-\delta)(2\|b\|_1 + \|C - \operatorname{diag}(C)\|_1) + \zeta\frac{(1+\lambda)^3}{4}\|e\|_2^2 + o(\|\Delta\|^2).$$

□

Auxiliary Lemma

Lemma 43. *For all matrix C symmetric semi-definite positive we have under assumptions (A1) and (A3):*

$$\operatorname{tr}\Big(S - \frac{Z^\top Z}{n}\Big)C \geq (1-\delta)\|C - \operatorname{diag}(C)\|_1 > 0.$$

Proof. We denote by $\Sigma^n = \frac{Z^\top Z}{n}$. We always have $\|C\|_1 - \operatorname{tr}(\Sigma^n C) = \operatorname{tr}(S - \Sigma^n)C$ where $S_{i,j} = \operatorname{sign}(C_{i,j})$, thus if $\operatorname{diag}(C) > 0$ then $\operatorname{diag}(S) = 1$ and $\operatorname{diag}(S - \Sigma^n) \geq 0$ from assumption (A1). Moreover since ${\Sigma^n}_{i,j} \in [-1, 1]$ then $\operatorname{sign}(S - \Sigma^n) = \operatorname{sign}(S)$.

Thus $\operatorname{tr}(S - \Sigma^n)C = \sum_i C_{i,i}(S - \Sigma^n)_{i,i} + \sum_{i \neq j} C_{i,j}(S - \Sigma^n)_{i,j} \geq \sum_{i \neq j} C_{i,j}(S - \Sigma^n)_{i,j} \geq 0$. Furthermore from assumption (A3) $|(\Sigma^n)_{i,j}| \leq \delta$ for $i \neq j$. Therefore

$$\operatorname{tr}(S - \Sigma^n)C \geq \sum_{i \neq j} C_{i,j}(S - \Sigma^n)_{i,j} \geq \sum_{i \neq j} |C_{i,j}|(1-\delta) \geq (1-\delta)\|C - \operatorname{diag}(C)\|_1 > 0.$$

If there is a diagonal element of C which is 0, then all the corresponding line and column in C will also be 0 and we can look at the same problem as before by erasing of C and Σ^n the corresponding column and line. □

Proof of Lemma 42

Proof. We will first show that the noise matrix Z satisfies assumptions (A2,A3). By Hoeffding's inequality we have with probability $1 - 2\exp(-\delta^2 n/(2R^2))$

$$\frac{1}{n}|\sum_{i=1}^n z_i^j| \leq \delta.$$

Then, since the law of z is symmetric $y_i z_i$ will have the same law as z_i and with probability $1 - 2\exp(-\delta^2 n/(2R^2))$, the design matrix Z satisfies assumption (A2):

$$\|\frac{Z^\top y}{n}\|_\infty \leq \delta.$$

Likewise we have with probability $1 - 2\exp(-\delta^2 n/(2R^4))$ that for $j \neq j'$

$$|\frac{1}{n}\sum_{i=1}^n z_i^j z_i^{j'}| \leq \delta.$$

Thus we also have with probability $1 - 2d^2\exp(-\delta^2 n/(2R^4))$ that Z satisfies assumption (A3):

$$\|\frac{1}{n}Z^\top Z - \operatorname{diag}(\frac{1}{n}Z^\top Z)\|_\infty \leq \delta.$$

Thus with probability $1 - 4d^2\exp(-\delta^2 n/(2R^4))$, the noise matrix Z satisfies assumptions (A1, A2, A3).

We proceed as in the proof of proposition 19 to show that X satisfies assumption (A4). We first derive a condition to have the result in expectation, then we use an inequality concentration on matrix to bound the empirical expectation. This will be very similar, but we will get a better scaling since Δ is diagonal.

Using the same arguments as in the proof of proposition 19 we have for the diagonal matrix $\Delta = \mathrm{diag}(e)$ with $e = (a, c) \in \mathbb{R}^d$:

$$e^\top \mathbb{E}(x^{\odot 2}(x^{\odot 2})^\top)e = \mathbb{E}(x^\top \Delta x)^2 = (a + mc^\top 1_{n-1})^2 + m^2(\beta - 1)\|c\|_2^2 > 0 \;\text{ if }\; \beta > 1.$$

We can show that $m^2(\beta-1)$ is an eigenvalue of multiplicity $d-2$ and $\mu_\pm$ are eigenvalues of multiplicity one of the operator $\Delta \mapsto \mathbb{E}(x^\top \Delta x)^2$ with eigenvectors $e_\pm$. Thus we have

$$\begin{aligned}\lambda_{\min}(\mathbb{E}x^{\odot 2}(x^{\odot 2})^\top) &= \frac{1 + (d+\beta-2)m^2 - \sqrt{(1+(d+\beta-2)m^2)^2 - 4m^2(\beta-1)}}{2} \\ &\geq \frac{m^2(\beta-1)}{1+(d+\beta-2)m^2}, \end{aligned} \tag{5.37}$$

and

$$\lambda_{\min}^{e_-^\perp}(\mathbb{E}x^{\odot 2}(x^{\odot 2})^\top) = m^2(\beta - 2).$$

Moreover

$$\lambda_{\max}\Big(x^{\odot 2}(x^{\odot 2})^\top\Big) = (x^{\odot 2})^\top x^{\odot 2} = \sum_{j=1}^{d}(x_i)^4 \leq dR^4.$$

Thus we can apply the Matrix Chernoff inequality from [Tropp, 2012] for $\mu_{\mathcal{S}} = \lambda_{\min}^{\mathcal{S}}(\mathbb{E}x^{\odot 2}(x^{\odot 2})^\top)$:

$$\mathbb{P}\left(\lambda_{\min}^{\mathcal{S}}\Big(\frac{X^{\odot 2}(X^{\odot 2})^\top}{n}\Big) \leq (1-\delta)\mu_{\mathcal{S}}\right) \leq de^{-\delta^2 n\mu_{\mathcal{S}}/(2dR^4)}.$$

Thus with probability $1 - 5d^2\exp(-\delta^2 n\mu_-/(2dR^4))$ the design matrix X satisfies assumption (A1,A2,A3,A4) with $\zeta = (1-\delta)\mu_-$ and $\mathcal{S} = \mathbb{R}^d$. And with probability $1 - 5d^2\exp(-\delta^2 n\min\{\beta-1, 2\}/(2dR^4))$ the design matrix X satisfies assumption (A1,A2,A3,A4) with $\zeta = (1-\delta)\min\{\beta-1, 2\}$ and $\mathcal{S} = e_-^\perp$. □

5.F Proof of Multi-Label Results

We first prove the lemma 31:

Proof. Let $A \in \mathbb{R}^{k\times k}$ symmetric semi-definite positive such that $\mathrm{diag}(\tilde{y}A\tilde{y}^\top) = 1_n$, then

$$\mathrm{diag}(\tilde{y}A\tilde{y}^\top) = \sum_{i=0}^{k} a_{i,i}1_n + 2\sum_{i=1}^{k} a_{0,i}y_i + 2\sum_{1\leq i<j\leq k} a_{i,j}y_i \odot y_j$$

thus

$$2\sum_{i=1}^{k} a_{0,i}y_i + 2\sum_{1\leq i<j\leq k} a_{i,j}y_i \odot y_j = (1-\sum_{i=0}^{k} a_{i,i})1_n$$

And this system admits as unique solution 0_n if and only if the family $\{1_n, (y_i)_{1\leq i\leq k}, (y_iy_j)_{1\leq i<j\leq k}\}$ is *linearly independent.* □

Then we prove the lemma 32:

Proof. Since $a_0 + \sum_{i=1}^{k} a_i^2\alpha_i \geq \alpha_{\min}\sum_{i=0}^{k} a_i^2 = \alpha_{\min}$ we should have $\alpha \geq \alpha_{\min}$. We have already seen that such Y satisfies the constraint. The KKT conditions are: $B = \operatorname{diag}(\mu) - H - \nu 11^\top \succcurlyeq 0$ and $BY = 0$. Since $y_i = \Pi_n y_i + \frac{(y_i^\top 1_n)}{n}1_n$.

$$\begin{aligned} Hy_i &= H\Pi_n y_i + (y_i^\top 1_n)H1_n \\ &= \Pi_n y = (y_i - \frac{1_n^\top y_i}{n}1_n). \end{aligned}$$

Thus

$$\begin{aligned} HY &= \sum_{i=1}^{k} a_i^2 Hy_iy_i^\top \\ &= \sum_{i=1}^{k} a_i^2(y_i - \frac{1_n^\top y_i}{n}1_n)y_i^\top \\ &= \sum_{i=1}^{k} a_i^2(y_iy_i^\top - \frac{1_n^\top y_i}{n}1_n y_i^\top) \end{aligned}$$

and $\operatorname{tr}(HY) = \sum_{i=1}^{k} a_i^2(n - n\alpha_i) = n(1 - a_0^2 + a_0^2 - \alpha) = n(1-\alpha)$.

Furthermore since $1_n^\top \operatorname{diag}(Y) = n$ and $1_n^\top M 1_n = n^2\alpha$, for $\mu = 1_n$ and $\nu = 1/n$, $B.Y = n - n(1-\alpha) - n\alpha = 0$. And since $B = I_n - \frac{1}{n}1_n1_n^\top - H$, $B^2 = B$ and $B^\top = B$, thus B is a symmetric projection and consequently symmetric semi-definit positive.

Hence the primal variable Y and the dual variables $\mu = 1_n$ and $\nu = 1/n$ satisfy the KKT conditions, thus Y is solution of this problem. □

5.G Efficient Optimization Problem

5.G.1 Dual Computation

We consider the following strongly convex approximation of Eq. (5.24), augmented with the von-Neumann entropy:

$$\begin{aligned} \max_{V\succcurlyeq 0} \quad & \frac{1}{n}\sum_{i=1}^{n}\sqrt{(XVX^\top)_{ii}} - \|\operatorname{Diag}(c)V\operatorname{Diag}(c)\|_1 - \varepsilon\operatorname{tr}[(A^{\frac{1}{2}}VA^{\frac{1}{2}})\log(A^{\frac{1}{2}}VA^{\frac{1}{2}})] \\ \text{s.t.} \quad & \operatorname{tr}(A^{\frac{1}{2}}VA^{\frac{1}{2}}) = 1. \end{aligned}$$

Introducing dual variables, we have

$$\begin{aligned}\min_{u\in\mathbb{R}^n_+, C:|C_{ij}|\leqslant c_i c_j} \max_{V\succcurlyeq 0} \quad & \frac{1}{2n}\sum_{i=1}^n \Big(u_i((XVX^\top)_{ii}) + \frac{1}{u_i}\Big) - \operatorname{tr} CV \\ & -\varepsilon \operatorname{tr}[(A^{\frac{1}{2}}VA^{\frac{1}{2}})\log(A^{\frac{1}{2}}VA^{\frac{1}{2}})] \\ \text{s.t.} \quad & \operatorname{tr}(A^{\frac{1}{2}}VA^{\frac{1}{2}}) = 1.\end{aligned}$$

By fixing u and C, and letting $Q = A^{\frac{1}{2}}VA^{\frac{1}{2}}$, we can write the max problem as

$$\begin{aligned}\max_{Q\succcurlyeq 0} \quad & \operatorname{tr} A^{-\frac{1}{2}}(\frac{1}{2n}X^\top \operatorname{Diag}(u)X - C)A^{-\frac{1}{2}}Q - \varepsilon \operatorname{tr}[Q\log(Q)] \\ \text{s.t.} \quad & \operatorname{tr} Q = 1.\end{aligned}$$

This problem is of the form

$$\max_{Q\succcurlyeq 0} \operatorname{tr} DQ - \varepsilon \sum_{i=1}^n \sigma_i(Q)\log\sigma_i(Q) \quad \text{s.t. } \operatorname{tr} Q = 1$$

where $D = A^{-\frac{1}{2}}(\frac{1}{2n}X^\top \operatorname{Diag}(u)X - C)A^{-\frac{1}{2}}$ and $\sigma_i(Q)$ denotes the i-th largest eigen value of the matrix Q. If we consider the matrix D to be of the form $D = U\operatorname{Diag}(\theta)U^\top$ with θ denoting the vector of ordered eigen values of D, then it turns out that at optimality Q has the form $Q = U\operatorname{Diag}(\sigma)U^\top$, with σ denoting the ordered vector of eigen values of Q.

Therefore the above optimization problem can be cast in terms of σ as:

$$\max_{\sigma\in\mathbb{R}^n} \theta^\top\sigma - \varepsilon\sum_{i=1}^n \sigma_i\log\sigma_i \quad \text{s.t. } \sum_{i=1}^n \sigma_i = 1.$$

The solution of this problem is $\sigma_i = \frac{e^{\theta_i/\varepsilon}}{\sum_{j=1}^n e^{\theta_j/\varepsilon}}$, which leads to

$$\min_{\theta\in\mathbb{R}^n} \phi^\varepsilon(\theta) = \varepsilon\log\sum_{i=1}^n \Big(e^{\frac{\theta_i}{\varepsilon}}\Big).$$

In terms of the original matrix variables, we have $\min\phi^\varepsilon(D) = \varepsilon\log\operatorname{tr} e^{\frac{D}{\varepsilon}}$. Using the appropriate expansion of D, we have the overall optimization problem as

$$\min_{u\in\mathbb{R}^n_+, C:|C_{ij}|\leqslant c_i c_j} \frac{1}{2n}\sum_{i=1}^n \frac{1}{u_i} + \phi^\varepsilon(A^{-\frac{1}{2}}(\frac{1}{2n}X^\top \operatorname{Diag}(u)X - C)A^{-\frac{1}{2}}). \tag{5.38}$$

At optimality, we have

$$A^{\frac{1}{2}}VA^{\frac{1}{2}} = \Big(e^{\frac{(A^{-\frac{1}{2}}(\frac{1}{2n}X^\top \operatorname{Diag}(u)X - C)A^{-\frac{1}{2}})}{\varepsilon}}\Big) / \operatorname{tr}\Big(e^{\frac{(A^{-\frac{1}{2}}(\frac{1}{2n}X^\top \operatorname{Diag}(u)X - C)A^{-\frac{1}{2}})}{\varepsilon}}\Big).$$

The error of approximation is at most $\varepsilon \log d$ and the Lipschitz constant associated with the function $\dot{\phi^{\varepsilon}}()$ is $\frac{1}{\varepsilon}$.

5.G.2 Algorithm Details

We write the optimization problem Eq. (5.38) as:

$$\min_{u \in \mathbb{R}_+^n} F(u, C) + H(u, C)$$

where

$$H(u, C) = \phi^{\varepsilon}(A^{-\frac{1}{2}}(\frac{1}{2n} X^{\top} \operatorname{Diag}(u) X - C) A^{-\frac{1}{2}})$$

is the smooth part and

$$F(u, C) = \mathbb{I}_{C:|C_{ij}| \leqslant c_i c_j} + \frac{1}{2n} \sum_{i=1}^{n} \frac{1}{u_i}$$

is the non-smooth part.

The gradient ∇_u of $H(u, C)$ with respect to u is

$$\nabla_u = \operatorname{diag}(B^{\top} U \operatorname{Diag}(\sigma) U^{\top} B).$$

where $B = \frac{1}{\sqrt{2n}} A^{-\frac{1}{2}} X^{\top}$ and the gradient of $H(u, C)$ with respect to C is

$$\nabla_C = (A^{-\frac{1}{2}} U \operatorname{Diag}(\sigma) U^{\top} A^{-\frac{1}{2}}).$$

The Lipschitz constant L associated with the gradient $\nabla H(u, C)$ is

$$L = \frac{2}{\varepsilon} \max \Big(\lambda_{max}(B^{\top} B \odot B^{\top} B), \lambda_{max}^2(A^{-1}) \Big), \tag{5.39}$$

where $\lambda_{max}(M)$ denotes the maximum eigen value of matrix M. Computing L takes $O(\max(n, d)^3)$ time and L needs to be computed once at the beginning of the algorithm.

The resultant FISTA procedure is described in Algorithm 1. Note that the FISTA procedure first computes intermediate iterates $(\bar{u}^{k-\frac{1}{2}}, \bar{C}^{k-\frac{1}{2}})$ (Step 7, Algorithm 1) by taking descent steps along the respective gradient directions. Then two distinct problems in u and C (respectively Steps 8 and 9 in Algorithm 1) are solved. The sub-problem in u (Step 8) can be efficiently solved using a Newton procedure followed by a thresholding step, as illustrated in Algorithm 2. The sub-problem in C (Step 9) can also be solved using a simple thresholding step.

Algorithm 1 *FISTA Algorithm to solve Eq. (5.38)*

Input X.
Compute Lipschitz constant L.
Let (u^0, C^0) be an arbitrary starting point.
Let $(\bar{u}^0, \bar{C}^0) = (u^0, C^0)$, $t_0 = 1$.
Set the maximum iterations to be K.
for $k = 1, 2, \ldots, K$ **do** ▷ The loop can also be terminated based on duality gap.
    $(\bar{u}^{k-\frac{1}{2}}, \bar{C}^{k-\frac{1}{2}}) = \Big(\bar{u}^k - \frac{1}{L}\nabla_{\bar{u}^k}, \bar{C}^k - \frac{1}{L}\nabla_{\bar{C}^k}\Big)$.
    Obtain $u^k = \arg\min_{u\in\mathbb{R}^n_+} \Big\{\frac{L}{2}\|u - \bar{u}^{k-\frac{1}{2}}\|^2 + \frac{1}{2n}\sum_{i=1}^n \frac{1}{u_i}\}$ by Algorithm 2.
    Obtain $C^k = \arg\min_C \Big\{\mathbb{I}_{C:|C_{ij}|\leqslant c_i c_j} + \frac{L}{2}\|C - \bar{C}^{k-\frac{1}{2}}\|_F^2\Big\}$ by thresholding.
    $t_k = \frac{1+\sqrt{1+4t_{k-1}^2}}{2}$.
    $(\bar{u}^k, \bar{C}^k) = (u^k, C^k) + \frac{(t_{k-1}-1)}{t_k}\Big((u^k, C^k) - (u^{k-1}, C^{k-1})\Big)$.
end for
Output (u^K, C^K).

Algorithm 2 *Newton method to solve u sub-problem*

Input $u^{k-\frac{1}{2}}$, n, L.
$u_i^0 = \max(u_i^{k-\frac{1}{2}}, \frac{1}{(2nL)^{\frac{1}{3}}})$, $i = 1, 2, \ldots, n$.
Set $\mathcal{M}$ to be the max number of Newton steps.
for $t = 1, 2, \ldots, \mathcal{M}$ **do**
    for $i = 1, 2, \ldots, n$ **do**
        $u_i^t = \frac{2nL(u_i^{t-1})^3 u_i^{k-\frac{1}{2}} + 3u_i^t}{2(nL(u_i^{t-1})^3+1)}$.
    end for
end for
Output $\max(u^{\mathcal{M}}, 0)$.

Chapter 6

Application to Isotonic Regression and Seriation Problems

Abstract

Given a matrix, the seriation problem consists in permuting its rows in such way that all its columns have the same shape, for example, they are monotone increasing. We propose a statistical approach to this problem where the matrix of interest is observed with noise and study the corresponding minimax rate of estimation of the matrices. Specifically, when the columns are either unimodal or monotone, we show that the least-squares estimator is optimal up to logarithmic factors and adapts to matrices with a certain natural structure. Finally, we propose a computationally efficient estimator in the monotonic case and study its performance both theoretically and experimentally. Our work is at the intersection of shape constrained estimation and recent work that involves permutation learning, such as graph denoising and ranking.

This chapter is extracted from the paper: Optimal rates of Statistical Seriation, in collaboration with C. Mao and P. Rigollet, submitted to *Bernoulli*.

6.1 Introduction

The *consecutive 1's problem* (C1P) [Fulkerson and Gross, 1964] is defined as follows. Given a binary matrix A the goal is to permute its rows in such a way that the resulting matrix enjoys the *consecutive 1's property*: each of its columns is a vector $v = (v_1, \dots, v_n)^\top$ where $v_j = 1$ if and only if $a \leq j \leq b$ for two integers a, b between 1 and n.

This problem has its roots in archeology and especially *sequence dating* where the goal is to recover the chronological order of sepultures based on artifacts found in these sepultures where the entry $A_{i,j}$ of matrix A indicates the presence of artifact j in sepulture i. In his seminal work, egyptologist Flinders Petrie [1899] formulated the hypothesis that two sepultures should be close in the time domain if they present similar sets of artifacts. Already in the noiseless case, this problem presents an in-

teresting algorithmic challenge and is reducible to the famous Travelling Salesman Problem [Gertzen and Grötschel, 2012] as observed by statistician David Kendall [1963, 1969, 1970, 1971] who employed early tools from multidimensional scaling as a heuristic to solve it. C1P belongs to a more general class of so-called *seriation* problems that consist in optimizing various criteria over the discrete set of permutations. While such problems are hard in general, it can be shown that a subset of the these problems, including C1P, can be solved efficiently using spectral method [Atkins et al., 1998] or convex optimization [Fogel et al., 2013, Lim and Wright, 2014]. However, little is known about the robustness to noise of such methods.

In order to set the benchmark for the noisy case, we propose a statistical *seriation model* and study optimal rates of estimation in this model. Assume that we observe an $n \times m$ matrix $Y = \Pi A + Z$, where Π is an unknown $n \times n$ permutation matrix, Z is an $n \times m$ noise matrix and $A \in \mathbb{R}^{n \times m}$ is assumed to belong to a class of matrices that satisfy a certain shape constraint. Our goal is to give estimators $\hat{\Pi}$ and $\hat{A}$ so that $\hat{\Pi}\hat{A}$ is close to ΠA. The shape constraint can be the consecutive 1's property, but more generally, we consider the class of matrices that have unimodal columns, which also include monotonic columns as a special case. These terms will be formally defined at the end of this section.

The rest of the chapter is organized as follows. In Section 6.2 we formulate the model and discuss related work. Section 6.3 collects our main results, including uniform and adaptive upper bounds for the least-squares estimator together with corresponding minimax lower bounds in the general unimodal case. In Section 6.4, for the special case of monotone columns, we propose a computationally efficient alternative to the least-squares estimator and study its rates of convergence both theoretically and numerically. Appendix 6.A is devoted to the proofs of the upper bounds, which use the metric entropy bounds proved in Appendix 6.B. The proofs of the information-theoretic lower bounds are presented in Appendix 6.C. In Appendix 6.D, we study the rate of estimation of the efficient estimator for the monotonic case. Appendix 6.E contains a delayed proof of a trivial upper bound. Appendix 6.F presents new bounds for unimodal regression implied by our analysis, which are minimax optimal up to logarithmic factors.

NOTATION. For a positive integer n, define $[n] = \{1, \dots, n\}$. For a matrix $A \in \mathbb{R}^{n \times m}$, let $\|A\|_F$ denote its Frobenius norm, and let $A_{i,\cdot}$ be its i-th row and $A_{\cdot,j}$ be its j-th column. Let $\mathcal{B}^n(a,t)$ denote the Euclidean ball of radius t centered at a in $\mathbb{R}^n$. We use C and c to denote positive constants that may change from line to line. For any two sequences $(u_n)_n$ and $(v_n)_n$, we write $u_n \lesssim v_n$ if there exists an absolute constant $C > 0$ such that $u_n \le C v_n$ for all n. We define $u_n \gtrsim v_n$ analogously. Given two real numbers a, b, define $a \wedge b = \min(a,b)$ and $a \vee b = \max(a,b)$.

Denote the closed convex cone of increasing [1] sequences in $\mathbb{R}^n$ by $\mathcal{S}_n = \{a \in \mathbb{R}^n : a_1 \le \cdots \le a_n\}$. We define $\mathcal{S}^m$ to be the Cartesian product of m copies of $\mathcal{S}_n$ and we identify $\mathcal{S}^m$ to the set of $n \times m$ matrices with increasing columns.

For any $l \in [n]$, define the closed convex cone $\mathcal{C}_l = \{a \in \mathbb{R}^n : a_1 \le \cdots \le a_l\} \cap \{a \in$

1. Throughout the chapter, we loosely use the terms "increasing" and "decreasing" to mean "monotonically non-decreasing" and "monotonically non-increasing" respectively.

$\mathbb{R}^n : a_l \geq \cdots \geq a_n\}$, which consists of vectors in $\mathbb{R}^n$ that increase up to the l-th entry and then decrease. Define the set $\mathcal{U}$ of unimodal sequences in $\mathbb{R}^n$ by $\mathcal{U} = \bigcup_{l=1}^n \mathcal{C}_l$. We define $\mathcal{U}^m$ to be the Cartesian product of m copies of $\mathcal{U}$ and we identify $\mathcal{U}^m$ to the set of $n \times m$ matrices with unimodal columns. It is also convenient to write $\mathcal{U}^m$ as a union of closed convex cones as follows. For $\mathbf{l} = (l_1, \ldots, l_m) \in [n]^m$, let $\mathcal{C}_{\mathbf{l}}^m = \mathcal{C}_{l_1} \times \cdots \times \mathcal{C}_{l_m}$. Then $\mathcal{U}^m$ is the union of the n^m closed convex cones $\mathcal{C}_{\mathbf{l}}^m, \mathbf{l} \in [n]^m$.

Finally, let $\mathfrak{S}_n$ be the set of $n \times n$ permutation matrices and define $\mathcal{M} = \bigcup_{\Pi \in \mathfrak{S}_n} \Pi\mathcal{U}^m$ where $\Pi\mathcal{U}^m = \{\Pi A : A \in \mathcal{U}^m\}$, so that $\mathcal{M}$ is the union of the $n!n^m$ closed convex cones $\Pi\mathcal{C}_{\mathbf{l}}^m, \Pi \in \mathfrak{S}_n, \mathbf{l} \in [n]^m$.

6.2 Problem Setup and Related Work

In this section, we formally state the problem of interest and discuss several lines of related work.

6.2.1 The Seriation Model

Suppose that we observe a matrix $Y \in \mathbb{R}^{n \times m}$, $n \geq 2$ such that

$$Y = \Pi^* A^* + Z\,, \tag{6.1}$$

where $A^* \in \mathcal{U}^m$, $\Pi \in \mathfrak{S}_n$ and Z is a centered sub-Gaussian noise matrix with variance proxy $\sigma^2 > 0$. More specifically, Z is a matrix such that $\mathbb{E}[Z] = 0$ and, for any $M \in \mathbb{R}^{n \times m}$,

$$\mathbb{E}\big[\exp\big(\mathsf{Tr}(Z^\top M)\big)\big] \leq \exp\Big(\frac{\sigma^2 \|M\|_F^2}{2}\Big)\,,$$

where $\mathsf{Tr}(\cdot)$ is the trace operator. We write $Z \sim \mathrm{subG}_{n,m}(\sigma^2)$ or simply $Z \sim \mathrm{subG}(\sigma^2)$ when dimensions are clear from the context.

Given the observation Y, our goal is to estimate the unknown pair (Π^*, A^*). The performance of an estimator $(\hat{\Pi}, \hat{A}) \in \mathfrak{S}_n \times \mathcal{U}^m$, is measured by the quadratic loss:

$$\frac{1}{nm}\|\hat{\Pi}\hat{A} - \Pi^* A^*\|_F^2\,.$$

In particular, its expectation is the mean squared error. Since we are interested in estimating $\Pi^* A^* \in \mathcal{M}$, we can also view $\mathcal{M}$ as the parameter space.

In the general unimodal case, upper bounds on the above quadratic loss do not imply individual upper bounds on estimation of the matrix Π^* or the matrix A^* due to lack of identifiability. Nevertheless, if we further assume that the columns of A^* are monotone increasing, that is $A^* \in \mathcal{S}^m$, then the following lemma holds.

Lemma 44. *If $A^*, \tilde{A} \in \mathcal{S}^m$, then for any $\Pi^*, \tilde{\Pi} \in \mathfrak{S}_n$, we have that*

$$\|\tilde{A} - A^*\|_F^2 \leq \|\tilde{\Pi}\tilde{A} - \Pi^* A^*\|_F^2\,,$$

and that

$$\|\tilde{\Pi}A^* - \Pi^* A^*\|_F^2 \leq 4\|\tilde{\Pi}\tilde{A} - \Pi^* A^*\|_F^2 .$$

Proof. Let $a, b \in \mathcal{S}_n$ and $b_\pi = (b_{\pi(1)}, \dots, b_{\pi(n)})$ where $\pi : [n] \to [n]$ is a permutation. It is easy to check that $\sum_{i=1}^n a_i b_i \geq \sum_{i=1}^n a_i b_{\pi(i)}$, so $\|a - b\|_2^2 \leq \|a - b_\pi\|_2^2$. Applying this inequality to columns of matrices, we see that

$$\|\tilde{A} - A^*\|_F^2 \leq \|\tilde{A} - \tilde{\Pi}^{-1}\Pi^* A^*\|_F^2 = \|\tilde{\Pi}\tilde{A} - \Pi^* A^*\|_F^2,$$

since $A^*, \tilde{A} \in \mathcal{S}^m$. Moreover, $\|\tilde{\Pi}A^* - \tilde{\Pi}\tilde{A}\|_F = \|A^* - \tilde{A}\|_F$, so

$$\|\tilde{\Pi}A^* - \Pi^* A^*\|_F \leq \|A^* - \tilde{A}\|_F + \|\tilde{\Pi}\tilde{A} - \Pi^* A^*\|_F \leq 2\|\tilde{\Pi}\tilde{A} - \Pi^* A^*\|_F,$$

by the triangle inequality and the previous display. □

Lemma 44 guarantees that $\|\tilde{\Pi}A^* - \Pi^* A^*\|_F$ is a pertinent measure of the performance of $\tilde{\Pi}$. Note further that $\|\tilde{\Pi}A^* - \Pi^* A^*\|_F$ is large if $\tilde{\Pi}$ misplaces rows of A^* that have large differences, and is small if $\tilde{\Pi}$ only misplaces rows of A^* that are close to each other. We argue that, in the seriation context, this measure of distance between permutations is more natural than ad hoc choices such as the trivial 0/1 distance or popular choices such as Kendall's τ or Spearman's ρ.

Apart from Section 6.4 (and Appendix 6.D), the rest of this chapter focuses on the least-squares (LS) estimator defined by

$$(\hat{\Pi}, \hat{A}) \in \operatorname*{argmin}_{(\Pi, A) \in \mathfrak{S}_n \times \mathcal{U}^m} \|Y - \Pi A\|_F^2 . \tag{6.2}$$

Taking $\hat{M} = \hat{\Pi}\hat{A}$, we see that it is equivalent to define the LS estimator by

$$\hat{M} \in \operatorname*{argmin}_{M \in \mathcal{M}} \|Y - M\|_F^2 . \tag{6.3}$$

Note that in our case, the set of parameters $\mathcal{M}$ is not convex, but is a union of $n!n^m$ closed convex cones and it is not clear how to compute the LS estimator efficiently. We discuss this aspect in further details in the context of monotone columns in Section 6.4. Nevertheless, the main focus of this chapter is the least-squares estimator which, as we shall see, is near-optimal in a minimax sense and therefore serves as a benchmark for the statistical seriation model.

6.2.2 Related Work

Our work falls broadly in the scope of statistical inference under shape constraints but presents a major twist: the unknown latent permutation Π^*.

Shape Constrained Regression

To set our goals, we first consider the case where the permutation is known and assume without loss of generality that $\Pi^* = I_n$. In this case, we can estimate individ-

ually each column $A^*_{\cdot,j}$ by an estimator $\hat{A}_{\cdot,j}$ and then get an estimator $\hat{A}$ for the whole matrix by concatenating the columns $\hat{A}_{\cdot,j}$. Thus the task is reduced to estimation of a vector θ^* which satisfies a certain shape constraint from an observation $y = \theta^* + z$ where $z \sim \mathrm{subG}_{n,1}(\sigma^2)$.

When θ^* is assumed to be increasing we speak of isotonic regression [Barlow et al., 1972]. The LS estimator defined by $\hat{\theta} = \mathrm{argmin}_{\theta \in \mathcal{S}_n} \|\theta - y\|_2^2$ can be computed in closed form in $O(n)$ using the Pool-Adjacent-Violators algorithm (PAVA) [Ayer et al., 1955, Barlow et al., 1972, Robertson et al., 1988] and its statistical performance has been studied by Zhang [2002] (see also works by Nemirovski et al. [1985], Donoho [1990], van de Geer [1990], Mammen [1991], van de Geer [1993] for similar bounds using empirical process theory) who showed in the Gaussian case $z \sim N(0, \sigma^2 I_n)$ that the mean squared error behaves like

$$\frac{1}{n}\mathbb{E}\|\hat{\theta} - \theta^*\|_2^2 \asymp \Big(\frac{\sigma^2 V(\theta^*)}{n}\Big)^{2/3}, \tag{6.4}$$

where $V(\theta) = \max_{i \in [n]} \theta_i - \min_{i \in [n]} \theta_i$ is the variation of $\theta \in \mathbb{R}^n$. Note that $2/3 = 2\beta/(2\beta + 1)$ for $\beta = 1$ so that this is the minimax rate of estimation of Lipschitz functions [see, e.g., Tsybakov, 2009].

The rate in Eq. (6.4) is said to be *global* since it holds uniformly over the set of monotone vectors with variation $V(\theta^*)$. Recently, Chatterjee et al. [2015] have initiated the study of *adaptive* bounds that may be better if θ^* has a simpler structure in some sense. To define this structure, let $k(\theta) = \mathrm{Card}(\{\theta_1, \cdots, \theta_n\})$ denote the cardinality of entries of $\theta \in \mathbb{R}^n$. In this context, Chatterjee et al. [2015] showed that the LS estimator satisfies the adaptive bound

$$\frac{1}{n}\mathbb{E}\|\hat{\theta} - \theta^*\|_2^2 \le C \inf_{\theta \in \mathcal{S}_n} \Big(\frac{\|\theta - \theta^*\|^2}{n} + \frac{\sigma^2 k(\theta)}{n} \log \frac{en}{k(\theta)} \Big). \tag{6.5}$$

This result was extended by Bellec [2015] to a sharp oracle inequality where $C = 1$. This bound was also shown to be optimal in a minimax sense by Chatterjee et al. [2015], Bellec and Tsybakov [2015].

Unlike its monotone counterpart, unimodal regression where $\theta^* \in \mathcal{U}$ has received sporadic attention [Shoung and Zhang, 2001, Köllmann et al., 2014, Chatterjee and Lafferty, 2015]. This state of affairs is all the more surprising given that unimodal density estimation has been the subject of much more research [Bickel and Fan, 1996, Birge, 1997, Eggermont and LaRiccia, 2000, Daskalakis et al., 2012, 2013, Turnbull and Ghosh, 2014]. It was recently shown by Chatterjee and Lafferty [2015] that the LS estimator also adapts to $V(\theta^*)$ and $k(\theta^*)$ for unimodal regression:

$$\frac{1}{n}\|\hat{\theta} - \theta^*\|_2^2 \lesssim \min\Big(\sigma^{4/3} \Big(\frac{V(\theta^*) + \sigma}{n} \Big)^{2/3}, \frac{\sigma^2}{n} k(\theta^*)^{3/2} (\log n)^{3/2} \Big) \tag{6.6}$$

with probability at least $1 - n^{-\alpha}$ for some $\alpha > 0$. The exponent $3/2$ in the second term was improved to 1 in the new version by Chatterjee and Lafferty [2015] after the first version of our current paper was posted. Note that the exponents in Eq. (6.6)

are different from the isotonic case. Our results will imply that they are not optimal and in fact the LS estimator achieves the same rate as in isotonic regression. See Corollary 10 for more details. The algorithmic aspect of unimodal regression has received more attention [Frisen, 1986, Geng and Shi, 1990, Bro and Sidiropoulos, 1998, Boyarshinov and Magdon-Ismail, 2006] and Stout [2008] showed that the LS estimator can be computed with time complexity $O(n)$ using a modified version of PAVA. Hence there is little difference between isotonic and unimodal regressions from both computational and statistical points of views.

Latent Permutation Learning

When the permutation Π^* is unknown the estimation problem is more involved. Noisy permutation learning was explicitly addressed by Collier and Dalalyan [2016] where the problem of matching two sets of noisy vectors was studied from a statistical point of view. Given $n \times m$ matrices $Y = A^* + Z$ and $\tilde{Y} = \Pi^* A^* + \tilde{Z}$, where $A^* \in \mathbb{R}^{n \times m}$ is an unknown matrix and $\Pi^* \in \mathbb{R}^{n \times n}$ is an unknown permutation matrix, the goal is to recover Π^*. It was shown by [Collier and Dalalyan, 2016] that if $\min_{i \neq j} \|A_{i,\cdot} - A_{j,\cdot}\|_2 \geq c\sigma\big((\log n)^{1/2} \vee (m \log n)^{1/4}\big)$, then the LS estimator defined by $\hat{\Pi} = \operatorname{argmin}_{\Pi \in \mathfrak{S}_n} \|\Pi Y - \tilde{Y}\|_F^2$ recovers the true permutation with high probability. However they did not directly study the behavior of $\|\hat{\Pi} A^* - \Pi^* A^*\|_F^2$.

In his celebrated paper on matrix estimation, Sourav Chatterjee [2015] describes several noisy matrix models involving unknown latent permutations. One is the *nonparametric Bradley-Terry-Luce* (NP-BTL) model where we observe a matrix $Y \in \mathbb{R}^{n \times n}$ with independent entries $Y_{i,j} \sim \mathrm{Ber}(P_{i,j})$ for some unknown parameters $P = \{P_{i,j}\}_{1 \leq i,j \leq n}$ where $P_{i,j} \in [0,1]$ is equal to the probability that item i is preferred over item j and $P_{j,i} = 1 - P_{i,j}$. Crucially, the NP-BTL model assumes the so-called *strong stochastic transitivity (SST)* [Davidson and Marschak, 1959, Fishburn, 1973] assumption: there exists an unknown permutation matrix $\Pi \in \mathbb{R}^{n \times n}$ such that the ordered matrix $A = \Pi^\top P \Pi$ satisfies $A_{1,k} \leq \cdots \leq A_{n,k}$ for all $k \in [n]$. Note that the NP-BTL model is a special case of our model in Eq. (6.1) where $m = n$ and $Z \sim \mathrm{subG}(1/4)$ is taken to be Bernoulli. Chatterjee [2015] proposed an estimator $\hat{P}$ that leverages the fact that any matrix P in the NP-BTL model can be approximated by a low rank matrix and proved [Chatterjee, 2015, Theorem 2.11] that $n^{-2}\|\hat{P} - P\|_F^2 \lesssim n^{-1/4}$, which was improved to $n^{-1/2}$ by Shah et al. [2017] for a variation of the estimator. This method does not yield individual estimators of Π or A, and Chatterjee and Mukherjee [2016] proposed estimators $\hat{\Pi}$ and $\hat{A}$ so that $\hat{\Pi}\hat{A}\hat{\Pi}^\top$ estimates P with the same rate $n^{-1/2}$ up to a logarithmic factor. The non-optimality of this rate has been observed by Shah et al. [2017] who showed that the correct rate should be of order n^{-1} up to a possible $\log n$ factor. However, it is not known whether a computationally efficient estimator could achieve the fast rate. A recent work by Shah et al. [2016] explored a new notion of adaptivity for which the authors proved a computational lower bound, and also proposed an efficient estimator whose rate of estimation matches that lower bound.

Also mentioned by Chatterjee [2015] is the so-called *stochastic block model* that has since received such extensive attention in various communities that it is futile to

attempt to establish a comprehensive list of references. Instead, we refer the reader to the article by Gao et al. [2015] and references therein. This paper establishes the minimax rates for this problem and its continuous limit, the graphon estimation problem and, as such, constitutes the state-of-the-art in the statistical literature. In the stochastic block model with $k \geq 2$ blocks, we assume that we observe a matrix $Y = P + Z$ where $P = \Pi A \Pi^\top, \Pi \in \mathbb{R}^{\times n}$ is an unknown permutation matrix and A has a block structure, namely, there exist positive integers $n_1 < \ldots < n_k < n_{k+1} := n$, and k^2 real numbers $a_{s,t}, (s,t) \in [k]^2$ such that A has entries

$$A_{i,j} = \sum_{(s,t)\in[k]^2} a_{s,t} \mathbb{1}\{n_s \leq i \leq n_{s+1}, n_t \leq j \leq n_{t+1}\}, \qquad i,j \in [n].$$

While traditionally, the stochastic block model is a network model and therefore pertains only to Bernoulli observations, the more general case of sub-Gaussian additive error is also explicitly handled by Gao et al. [2015]. For this problem, Gao, Liu and Zhou have established that the least-squares estimator $\hat{P}$ satisfies $n^{-2}\|\hat{P} - P\|_F^2 \lesssim k^2/n^2 + (\log k)/n$ together with a matching lower bound. Using piecewise constant approximation to bivariate Hölder functions, they also establish that this estimator with a correct choice of k leads to minimax optimal estimation of smooth graphons. Both results exploit extensively the fact that the matrix P is equal to or can be well approximated by a piecewise constant matrix and our results below take a similar route by observing that monotone and unimodal vectors are also well approximated by piecewise constant ones. Moreover, we allow for rectangular matrices.

In fact, our result can be also formulated as a network estimation problem but on a bipartite graph, thus falling at the intersection of the above two examples. Assume that n left nodes represent items and that m right nodes represent users. Assume further that we observe the $n \times m$ adjacency matrix Y of a random graph where the presence of edge (i,j) indicates that user j has purchased or liked item i. Define $P = \mathbb{E}[Y]$ and assume SST across items in the sense that there exists an unknown $n \times n$ permutation matrix Π^* such that $P = \Pi^* A^*$ and A^* is such that $A^*_{1,j} \leq \cdots \leq A^*_{n,j}$ for all users $j \in [m]$. This model falls into the scope of the statistical seriation model in Eq. (6.1).

6.3 Main Results

6.3.1 Adaptive Oracle Inequalities

For a matrix $A \in \mathcal{U}^m$, let $k(A_{\cdot,j}) = \mathrm{Card}(\{A_{1,j}, \ldots, A_{n,j}\})$ be the number of values taken by the j-th column of A and define $K(A) = \sum_{j=1}^m k(A_{\cdot,j})$. Observe that $K(A) \geq m$. The first theorem shows that the LS estimator adapts to the complexity K.

Theorem 10. *For $A^* \in \mathbb{R}^{n\times m}$ and $Y = \Pi^* A^* + Z$, let $(\hat{\Pi}, \hat{A})$ be the LS estimator*

defined in Eq. (6.2). Then the following oracle inequality holds

$$\frac{1}{nm}\|\hat{\Pi}\hat{A} - \Pi^* A^*\|_F^2 \lesssim \min_{A \in \mathcal{U}^m} \Big(\frac{1}{nm}\|A - A^*\|_F^2 + \sigma^2 \frac{K(A)}{nm} \log \frac{enm}{K(A)} \Big) + \sigma^2 \frac{\log n}{m} \quad (6.7)$$

with probability at least $1 - e^{-c(n+m)}, c > 0$. *Moreover,*

$$\frac{1}{nm}\mathbb{E}\|\hat{\Pi}\hat{A} - \Pi^* A^*\|_F^2 \lesssim \min_{A \in \mathcal{U}^m} \Big(\frac{1}{nm}\|A - A^*\|_F^2 + \sigma^2 \frac{K(A)}{nm} \log \frac{enm}{K(A)} \Big) + \sigma^2 \frac{\log n}{m} \,. \quad (6.8)$$

Note that while we assume that $A^* \in \mathcal{U}^m$ in Eq. (6.1), the above oracle inequalities hold in fact for any $A^* \in \mathbb{R}^{n \times m}$ even if its columns are *not* assumed to be unimodal.

The above oracle inequalities indicate that the LS estimator automatically trades off the approximation error $\|A - A^*\|_F^2$ for the stochastic error $\sigma^2 K(A) \log(enm/K(A))$.

If A^* is assumed to have unimodal columns, then we can take $A = A^*$ in Eq. (6.7) and Eq. (6.8) to get the following corollary.

Corollary 8. *For* $A^* \in \mathcal{U}^m$ *and* $Y = \Pi^* A^* + Z$, *the LS estimator* $(\hat{\Pi}, \hat{A})$ *satisfies*

$$\frac{1}{nm}\|\hat{\Pi}\hat{A} - \Pi^* A^*\|_F^2 \lesssim \sigma^2 \Big(\frac{K(A^*)}{nm} \log \frac{enm}{K(A^*)} + \frac{\log n}{m} \Big)$$

with probability at least $1 - e^{-c(n+m)}, c > 0$. *Moreover, the corresponding bound with the same rate holds in expectation.*

The two terms in the adaptive bound can be understood as follows. The first term corresponds to the estimation of the matrix A^* with unimodal columuns if the permutation Π^* is known. It can be viewed as a matrix version of the adaptive bound in Eq. (6.5) in the vector case. The LS estimator adapts to the cardinality of entries of A^* as it achieves a provably better rate if $K(A^*)$ is smaller while not requiring knowledge of $K(A^*)$. The second term corresponds to the error due to the unknown permutation Π^*. As m grows to infinity this second term vanishes, because we have more samples to estimate Π^* better. If $m \geq n$, it is easy to check that the permutation term is dominated by the first term, so the rate of estimation is the same as if the permutation is known.

6.3.2 Global Oracle Inequalities

The bounds in Theorem 10 adapt to the cardinality of the oracle. In this subsection, we state another type of upper bounds for the LS estimator $(\hat{\Pi}, \hat{A})$. They are called global bounds because they hold uniformly over the class of matrices whose columns are unimodal and that have bounded variation. Recall that we call *variation* of a vector $a \in \mathbb{R}^n$ the scalar $V(a) \geq 0$ defined by

$$V(a) = \max_{1 \leq i \leq n} a_i - \min_{1 \leq i \leq n} a_i \,.$$

We extend this notion to a matrix $A \in \mathbb{R}^{n\times m}$ by defining

$$V(A) = \Big(\frac{1}{m}\sum_{j=1}^{m} V(A_{\cdot,j})^{2/3}\Big)^{3/2}.$$

While this 2/3-norm may seem odd at first sight, it turns out to be the correct extrapolation from vectors to matrices, at least in the context under consideration here. Indeed, the following upper bound, in which this quantity naturally appears, is matched by the lower bound of Theorem 13 up to logarithmic terms.

Theorem 11. *For $A^* \in \mathbb{R}^{n\times m}$ and $Y = \Pi^* A^* + Z$, let $(\hat{\Pi}, \hat{A})$ be the LS estimator defined in Eq. (6.2). Then it holds that*

$$\frac{1}{nm}\|\hat{\Pi}\hat{A} - \Pi^* A^*\|_F^2 \lesssim \min_{A\in\mathcal{U}^m}\Big[\frac{1}{nm}\|A - A^*\|_F^2 + \Big(\frac{\sigma^2 V(A)\log n}{n}\Big)^{2/3}\Big] + \sigma^2\frac{\log n}{n\wedge m}. \tag{6.9}$$

with probability at least $1 - e^{-c(n+m)}, c > 0$. Moreover, the corresponding bound with the same rate holds in expectation.

If $A^* \in \mathcal{U}^m$, then taking $A = A^*$ in Theorem 11 leads to the following corollary that indicates that the LS estimator is adaptive to the quantity $V(A^*)$.

Corollary 9. *For $A^* \in \mathcal{U}^m$ and $Y = \Pi^* A^* + Z$, the LS estimator $(\hat{\Pi}, \hat{A})$ satisfies*

$$\frac{1}{nm}\|\hat{\Pi}\hat{A} - \Pi^* A^*\|_F^2 \lesssim \Big(\frac{\sigma^2 V(A^*)\log n}{n}\Big)^{2/3} + \sigma^2\frac{\log n}{n\wedge m}$$

with probability at least $1 - e^{-c(n+m)}, c > 0$. Moreover, the corresponding bound with the same rate holds in expectation.

Akin to the adaptive bound, the above inequality can be viewed as a sum of a matrix version of Eq. (6.4) and an error due to estimation of the unknown permutation.

Having stated the main upper bounds, we digress a little to remark that the proofs of Theorem 10 and Theorem 11 also yield a minimax optimal rate of estimation (up to logarithmic factors) for unimodal regression, which improves the bound in Eq. (6.6). We discuss the details in Appendix 6.F.

6.3.3 Minimax Lower Bounds

Given the model $Y = \Pi^* A^* + Z$ where entries of Z are i.i.d. $N(0, \sigma^2)$ random variables, let $(\hat{\Pi}, \hat{A})$ denote any estimator of (Π^*, A^*), i.e., any pair in $\mathfrak{S}_n \times \mathbb{R}^{n\times m}$ that is measurable with respect to the observation Y. We will prove lower bounds that match the rates of estimation in Corollary 8 and Corollary 9 up to logarithmic factors. The combination of upper and lower bounds, implies simultaneous near optimality of the least-squares estimator over a large scale of matrix classes.

For $m \le K_0 \le nm$ and $V_0 > 0$, define $\mathcal{U}^m_{K_0} = \{A \in \mathcal{U}^m : K(A) \le K_0\}$ and $\mathcal{U}^m(V_0) = \{A \in \mathcal{U}^m : V(A) \le V_0\}$. We present below two lower bounds, one for the adaptive rate uniformly over $\mathcal{U}^m_{K_0}$ and one for the global rate uniformly over $\mathcal{U}^m(V_0)$. This splitting into two cases is solely justified by better readability but it is worth noting that a stronger lower bound that holds on the intersection $\mathcal{U}^m_{K_0} \cap \mathcal{U}^m(V_0)$ can also be proved and is presented as Proposition 21.

Theorem 12. *There exists a constant $c \in (0,1)$ such that for any $K_0 \ge m$, and any estimator $(\hat{\Pi}, \hat{A})$, it holds that*

$$\sup_{(\Pi,A)\in\mathfrak{S}_n\times\mathcal{U}^m_{K_0}} \mathbb{P}_{\Pi A}\Big[\frac{1}{nm}\|\hat{\Pi}\hat{A} - \Pi A\|_F^2 \gtrsim \sigma^2\Big(\frac{K_0}{nm} + \frac{\log l}{m}\Big)\Big] \ge c,$$

where $l = \min(K_0 - m, m) + 1$ and $\mathbb{P}_{\Pi A}$ is the probability distribution of $Y = \Pi A + Z$. It follows that the lower bound with the same rate holds in expectation.

In fact, the lower bound holds for any estimator of the matrix $\Pi^* A^*$, not only those of the form $\hat{\Pi}\hat{A}$ with $\hat{A} \in \mathcal{U}^m$. The above lower bound matches the upper bound in Corollary 8 up to logarithmic factors.

Note the presence of a $\log l$ factor in the second term. If $l = 1$ then $K_0 = m$ which means that each column of A is simply a constant block, so $\Pi A = A$ for any $\Pi \in \mathfrak{S}_n$. In this case, the second term vanishes because the permutation does not play a role. More generally, the number $l - 1$ can be understood as the maximal number of columns of A on which the permutation does have an effect. The larger l, the harder the estimation. It is easy to check that if $l \ge n$ the second term in the lower bound will be dominated by the first term in the upper bound.

A lower bound corresponding to Corollary 9 also holds:

Theorem 13. *There exists a constant $c \in (0,1)$ such that for any $V_0 \ge 0$, and any estimator $(\hat{\Pi}, \hat{A})$, it holds that*

$$\sup_{(\Pi,A)\in\mathfrak{S}_n\times\mathcal{U}^m(V_0)} \mathbb{P}_{\Pi A}\Big[\frac{1}{nm}\|\hat{\Pi}\hat{A} - \Pi A\|_F^2 \gtrsim \big(\frac{\sigma^2 V_0}{n}\big)^{2/3} + \frac{\sigma^2}{n} + \frac{\sigma^2}{m} \wedge m^2 V_0^2\Big] \ge c\,,$$

where $\mathbb{P}_{\Pi A}$ is the probability distribution of $Y = \Pi A + Z$. The lower bound with the same rate also holds in expectation.

There is a slight mismatch between the upper bound of Corollary 9 and the lower bound of Theorem 13 above. Indeed the lower bound features a term $\frac{\sigma^2}{m} \wedge m^2 V_0^2$ instead of just $\frac{\sigma^2}{m}$. In the regime $m^2 V_0^2 < \frac{\sigma^2}{m}$, where A has very small variation, the LS estimator may not be optimal. Proposition 22 indicates that a matrix with constant columns obtained by averaging achieves optimality in this extreme regime.

6.4 Further Results in the Monotone Case

A particularly interesting subset of unimodal matrices is $\mathcal{S}^m$, the set of $n \times m$ matrices with monotonically increasing columns. While it does not amount to the

seriation problem in its full generality, this special case is of prime importance in the context of shape constrained estimation as illustrated by the discussion and references in Section 6.2.2. In fact, it covers the example of bipartite ranking discussed at the end of Section 6.2.2. In the rest of this section, we devote further investigation to this important case. To that end, consider the model in Eq. (6.1) where we further assume that $A^* \in \mathcal{S}^m$. We refer to this model as the *monotone seriation model*. In this context, define the LS estimator by

$$(\hat{\Pi}, \hat{A}) \in \operatorname*{argmin}_{(\Pi, A) \in \mathfrak{S}_n \times \mathcal{S}^m} \|Y - \Pi A\|_F^2 .$$

Since $\mathcal{S}^m$ is a convex subset of $\mathcal{U}^m$, it is easily seen that the upper bounds in Theorem 10 and 11 remain valid in this case. The lower bounds of Theorem 12 (with $\log l$ replaced by 1) and Theorem 13 also extend to this case; see Appendix 6.C.

Although for unimodal matrices the established error bounds do not imply any bounds on estimation of A^* or Π^* in general, for the monotonic case, however, Lemma 44 yields that

$$\|\hat{A} - A^*\|_F^2 \vee \frac{1}{4}\|(\hat{\Pi} - \Pi^*)A^*\|_F^2 \leq \|\hat{\Pi}\hat{A} - \Pi^* A^*\|_F^2 .$$

so that the LS estimator $(\hat{\Pi}, \hat{A})$ also leads to good individual estimators of Π^* and A^* respectively.

Because it requires optimizing over a union of $n!$ cones $\Pi\mathcal{S}^m$, no efficient way of computing the LS estimator is known since. As an alternative, we describe a simple and efficient algorithm to estimate (Π^*, A^*) and study its rate of estimation.

Let $K(A)$ and $V(A)$ be defined as before. Moreover, for a matrix $A \in \mathcal{S}^m$, let $\mathcal{J}$ denote the set of pairs of indices $(i, j) \in [n]^2$ such that $A_{i,\cdot}$ and $A_{j,\cdot}$ are not identical. Define the quantity $R(A)$ by

$$R(A) = \frac{1}{n} \max_{\substack{\mathcal{I} \subset [n]^2 \\ |\mathcal{I}| = n}} \sum_{(i,j) \in \mathcal{I} \cap \mathcal{J}} \Big(\frac{\|A_{i,\cdot} - A_{j,\cdot}\|_2^2}{\|A_{i,\cdot} - A_{j,\cdot}\|_\infty^2} \wedge \frac{m\|A_{i,\cdot} - A_{j,\cdot}\|_2^2}{\|A_{i,\cdot} - A_{j,\cdot}\|_1^2} \Big) . \tag{6.10}$$

It can be shown (see Appendix 6.D) that $1 \leq R(A) \leq \sqrt{m}$. Intuitively, the quantity $R(A)$ is small if the difference u of any two rows of A is either very sparse ($\|u\|_2/\|u\|_\infty$ is small) or very dense ($m\|u\|_2/\|u\|_1$ is small). Indeed, for any nonzero vector $u \in \mathbb{R}^m$, $\|u\|_2^2/\|u\|_\infty^2 \geq 1$ with equality achieved when $\|u\|_0 = 1$, and $m\|u\|_2^2/\|u\|_1^2 \geq 1$ with equality achieved when all entries of u are the same.

For matrices with small $R(\cdot)$ values, it is possible to aggregate the information across each row to learn the unknown permutation Π^* in a simple fashion. Recovering the permutation Π^*, is equivalent to ordering (or ranking reversely) the rows of $\Pi^* A^*$ from their noisy version Y.

One simple method to achieve this goal, which we call RankSum, is to permute the rows of Y so that they have increasing row sums. However, it is easy to observe

that this method fails if

$$A^* = \begin{bmatrix} \sqrt{m} & 0 & \dots & 0 \\ 2\sqrt{m} & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ n\sqrt{m} & 0 & \dots & 0 \end{bmatrix} \tag{6.11}$$

where $A^*_{i,1} = i\sqrt{m}$ and entries of Z are i.i.d. standard Gaussian variables, because the sum of noise in a row has order $\sqrt{m}$ which is no less than the gaps between row sums of A^*. In fact, $R(A^*) = 1$ and it should be easy to distinguish the two types of rows of A^*, for example, by looking at the first entry of a row. This motivates us to consider the following method called RankScore.

For $i, i' \in [n]$, define

$$\Delta_{A^*}(i, i') = \max_{j \in [m]} (A^*_{i',j} - A^*_{i,j}) \vee \frac{1}{\sqrt{m}} \sum_{j=1}^{m} (A^*_{i',j} - A^*_{i,j})$$

and define $\Delta_Y(i, i')$ analogously. The RankScore procedure is defined as follows:

1. For each $i \in [n]$, define the score s_i of the i-th row of Y by

$$s_i = \sum_{l=1}^{n} \mathbb{1}(\Delta_Y(l, i) \geq 2\tau)$$

   where $\tau := C\sigma\sqrt{\log(nm)}$ for some tuning constant C (see Appendix 6.D for more details).
2. Then order the rows of Y so that their scores are increasing, with ties broken arbitrarily.

The RankScore procedure recovers an order of the rows of Y, which leads to an estimator $\tilde{\Pi}$ of the permutation. Then we define $\tilde{A} \in \mathcal{S}^m$ so that $\tilde{\Pi}\tilde{A}$ is the projection of Y onto the convex cone $\tilde{\Pi}\mathcal{S}^m$. The estimator $(\tilde{\Pi}, \tilde{A})$ enjoys the following rate of estimation.

Theorem 14. *For $A^* \in \mathcal{S}^m$ and $Y = \Pi^* A^* + Z$, let $(\tilde{\Pi}, \tilde{A})$ be the estimator defined above using the RankScore procedure with threshold $\tau = 3\sigma\sqrt{(C+1)\log(nm)}$, $C > 0$. Then it holds that*

$$\frac{1}{nm}\|\tilde{\Pi}\tilde{A} - \Pi^* A^*\|_F^2 \lesssim \min_{A \in \mathcal{S}^m} \Big(\frac{1}{nm}\|A - A^*\|_F^2 + \sigma^2 \frac{K(A)}{nm} \log \frac{enm}{K(A)} \Big) + (C+1)\sigma^2 \frac{R(A^*)\log(nm)}{m},$$

with probability at least $1 - e^{-c(n+m)} - (nm)^{-C}$ for some constant $c > 0$.

The quantity $R(A^*)$ only depends on the matrix A^*. If $R(A^*)$ is bounded logarithmically, the estimator $(\tilde{\Pi}, \tilde{A})$ achieves the minimax rate up to logarithmic factors. In any case, $R(A^*) \leq \sqrt{m}$, so the estimator is still consistent with the permutation error (the last term) decaying at a rate no slower than $\tilde{O}(\frac{1}{\sqrt{m}})$. Furthermore, it is

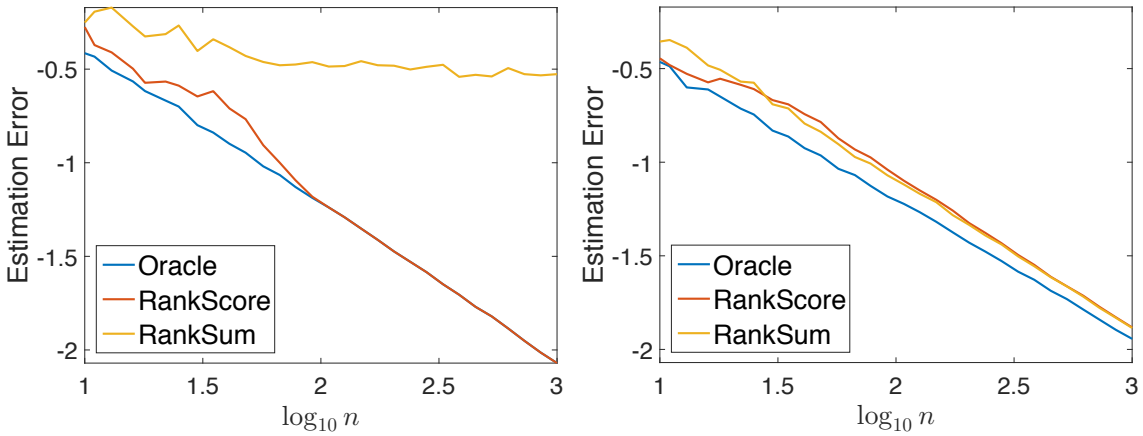


Figure 6-1 – Estimation errors of three estimators for two deterministic A^* of size $n \times n$. Left: rows of A^* are 1-sparse; Right: columns of A^* are identical.

worth noting that $R(A^*)$ is not needed to construct $(\tilde{\Pi}, \tilde{A})$, so the estimator adapts to $R(A^*)$ automatically.

Remark 4. *In the same way that Theorem 11 follows from Theorem 10, we can deduce from Theorem 14 a global bound for the estimator* $(\tilde{\Pi}, \tilde{A})$ *which has rate*

$$\Big(\frac{\sigma^2 V(A^*)\log n}{n}\Big)^{2/3} + \sigma^2\Big(\frac{\log n}{n} + R(A^*)\frac{\log(nm)}{m}\Big)\,.$$

We conclude this section with a numerical comparison between the RankSum and RankScore procedures.

Consider the model in Eq. (6.1) with $A^* \in \mathcal{S}^m$ and assume without loss of generality that $\Pi^* = I_n$. For various $n \times m$ matrices A^*, we generate observations $Y = A^* + Z$ where entries of Z are i.i.d. standard Gaussian variables. The performance of the estimators given by RankScore and RankSum defined above is compared to the performance of the oracle $\hat{A}^{\text{oracle}}$ defined by the projection of Y onto the cone $\mathcal{S}^m$. For the RankScore estimator we take $\tau = 6$. The curves are generated based on 30 equally spaced points on the base-10 logarithmic scale, and all results are averaged over 10 replications. The vertical axis represents the estimation error of an estimator $\hat{\Pi}\hat{A}$, measured by the sample mean of $\log_{10}\big(\frac{1}{nm}\|\hat{\Pi}\hat{A} - A^*\|_F^2\big)$ unless otherwise specified.

We begin with two simple examples for which we set $n = m$. In the left plot of Figure 6-1, A^* is defined as in Eq. (6.11). As expected, RankSum fails to estimate the true permutation and performs very poorly. On the other hand, RankScore succeeds in recovering the correct permutation and has roughly the same performance as the oracle. Because the difference of any two rows of A^* is 1-sparse, $R(A^*) = 1$ according to Eq. (6.10) and the discussion thereafter. Hence, Theorem 14 predicts the fast rate, which is verified by the experiment. The right plot illustrates another extreme case; more precisely, we set A^* to be the matrix with all m columns equal to $\frac{1}{n}(1, \cdots, n)^\top$. The difference of any two rows of A^* is constant across all entries, so again we have $R(A^*) = 1$ by Eq. (6.10). Thus RankScore achieves the fast rate as expected. Note that RankSum also performs well in this case.

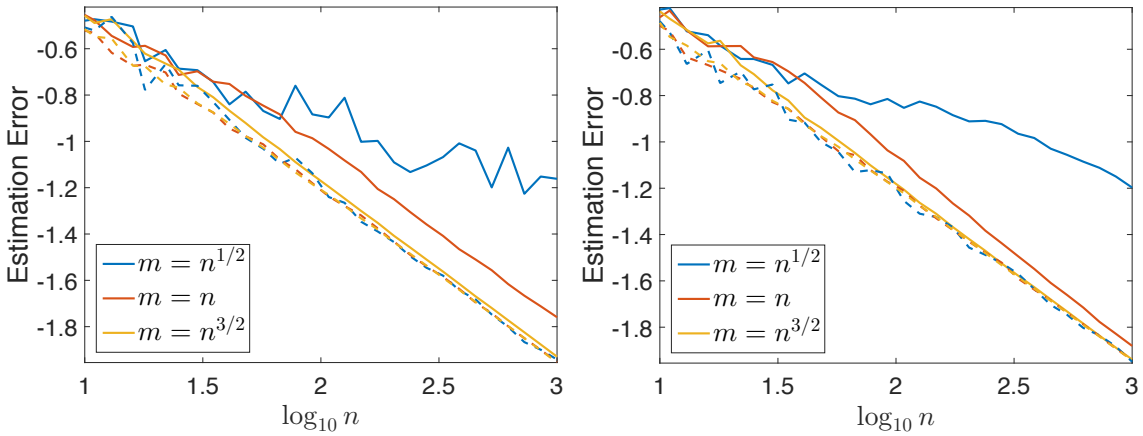


Figure 6-2 – Estimation errors of the oracle (dashed lines) and RankScore (solid lines) for different regimes of (n, m) and randomly generated A^* of size $n \times m$. Left: $K(A^*) = 5m$; Right: $V(A^*) \leq 1$.

In Figure 6-2, we compare the performance of RankScore to that of the oracle in three regimes of (n, m). The matrices A^* are randomly generated for different values of n and m as follows. For the right plot, A^* is generated so that $V(A^*) \leq 1$, by sorting the columns of a matrix with i.i.d. $U(0, 1)$ entries. For the left plot, we further require that $K(A^*) = 5m$ by uniformly partitioning each column of A^* into five blocks and assigning each block the corresponding value from a sorted sample of five i.i.d. $U(0, 1)$ variables.

Since the oracle knows the true permutation, its behavior is independent of m, and its rates of estimation are bounded by $\frac{\log n}{n}$ for $K(A^*) = 5m$ and $(\frac{\log n}{n})^{\frac{2}{3}}$ for $V(A^*) = 1$ respectively by Theorem 10 and 11. (The difference is minor in the plots as n is not sufficiently large). For RankScore, the permutation term dominates the estimation term when $m = n^{1/2}$ by Theorem 14. From the plots, the rates of estimation are better than $\tilde{O}(n^{-1/4})$ predicted by the worst-case analysis in both examples. For $m = n$, we also observe rates of estimation faster than the worst-case rate $\tilde{O}(n^{-1/2})$ and close to the oracle rates. We could explain this phenomenon by $R(A^*) < \sqrt{m}$, but such an interpretation may not be optimal since our analysis is based on worst-case deterministic A^*. Potential study of random designs of A^* is left open. Finally, for $m = n^{3/2}$, the permutation term is of order $\tilde{O}(n^{-3/4})$ theoretically, in between of the oracle rates for the two cases. Indeed RankScore has almost the same performance as the oracle experimentally. Overall Figure 6-2 illustrates the good behavior of RankScore in this random scenario.

To conclude our numerical experiments, we consider the $n \times n$ lower triangular matrix A^* defined by $A^*_{i,j} = \mathbb{1}(i \geq j)$. For this matrix, it is easy to check that $K(A^*) = 2n - 1$ and $R(A^*) \approx \sqrt{n}$. We plot in Figure 6-3 the estimation errors of $\tilde{\Pi}\tilde{A}$, $\tilde{\Pi}A^*$ and $\tilde{A}$ given by RankScore, in addition to the oracle. By Theorem 14, the rate of estimation achieved by $\tilde{\Pi}\tilde{A}$ is of order $\tilde{O}(n^{-1/2})$, while that achieved by the oracle is of order $\tilde{O}(n^{-1})$ since there is no permutation term. The plot confirms this discrepancy. Moreover, $\frac{1}{n^2}\|\tilde{\Pi}A^* - A^*\|_F^2$ is an appropriate measure of the performance

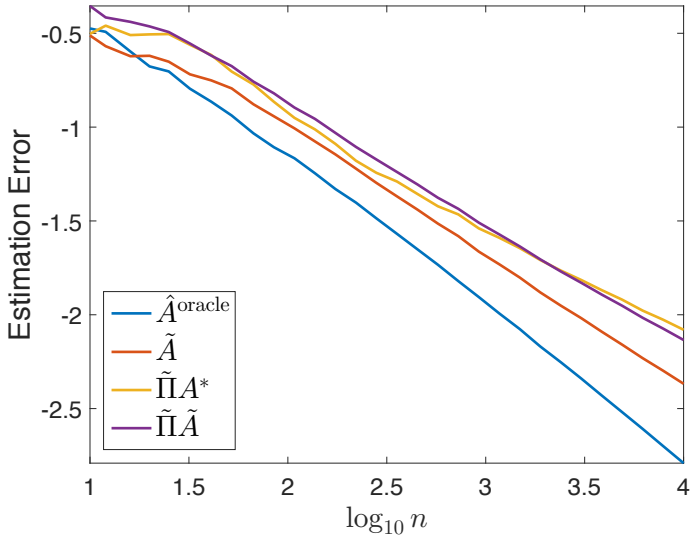


Figure 6-3 – Various estimation errors of the oracle and RankScore for the triangular matrix.

of $\tilde{\Pi}$ by Lemma 61 and 44, and the plot suggests that the rates of estimation achieved by $\tilde{\Pi}A^*$ and $\tilde{\Pi}\tilde{A}$ are about the same order. Finally $\tilde{A}$ seems to have a slightly faster rate of estimation than $\tilde{\Pi}\tilde{A}$, so in practice $\tilde{A}$ could be used to estimate A. However we refrain from making an explicit conjecture about the rate.

6.5 Discussion

While computational aspects of the seriation problem have received significant attention, the robustness of this problem to noise was still unknown to date. To overcome this limitation, we have introduced in this chapter the statistical seriation model and studied optimal rates of estimation by showing, in particular, that the least-squares estimator enjoys several desirable statistical properties such as adaptivity and minimax optimality (up to logarithmic terms).

While this work paints a fairly complete statistical picture of the statistical seriation model, it also leaves many unanswered questions. There are several logarithmic gaps in the bounds. In the case of adaptive bounds, some logarithmic terms are unavoidable as illustrated by Theorem 12 (for the permutation term) and also by statistical dimension consideration explained by Bellec [2015] (for the estimation term). However, a more refined argument for the uniform bound, namely one that uses covering in ℓ_2-norm rather than ℓ_∞-norm, would allow us to remove the $\log n$ factor from the estimation term in the upper bound of Corollary 9. Such an argument is provided by Birman and Solomjak [1967], Anuchina et al. [1979], van de Geer [1991] for the larger class of vectors with bounded total variation [see, e.g., Mammen and van de Geer, 1997] but we do not pursue sharp logarithmic terms in this work. For the permutation term, $\log n$ in the upper bound of Corollary 8 and $\log l$ in the lower bound of Theorem 12 do not match if $l < n$. We do not seek answers to these questions in

this chapter but note that their answers may be different for the unimodal and the monotone case.

Perhaps the most pressing question is that of computationally efficient estimators. Indeed, while statistically optimal, the least-squares estimator requires searching through $n!$ permutations, which is not realistic even for problems of moderate size, let alone genomics applications. We gave a partial answer to this question in the specific context of monotone columns by proposing and studying the performance of a simple and efficient estimator called RankScore. This study reveals the existence of a potentially intrinsic gap between the statistical performance achievable by efficient estimators and that achievable by estimators with access to unbounded computation. A similar gap is also observed in the SST model for pairwise comparisons by Shah et al. [2017]. We conjecture that achieving optimal rates of estimation in the seriation model is computationally hard in general but argue that the planted clique assumption that has been successfully used to establish statistical vs. computational gaps by Berthet and Rigollet [2013], Ma and Wu [2015], Shah et al. [2016] for example, is not the correct primitive. Instead, one has to seek for a primitive where hardness comes from searching through permutations rather than subsets.

Appendix

6.A Proof of the Upper Bounds

Before proving the main theorems, we discuss two methods adopted in recent works to bound the error of the LS estimator in shape constrained regression, in a general setting. Consider the least-squares estimator $\hat{\theta}$ of the model $y = \theta^* + z$, where θ^* lies in a parameter space Θ and z is Gaussian noise. One way to study $\mathbb{E}\|\hat{\theta} - \theta^*\|_2^2$ is to use the *statistical dimension* [Amelunxen et al., 2014] of a convex cone Θ defined by

$$\mathbb{E}\Big[\Big(\sup_{\theta\in\Theta,\,\|\theta\|_2\leq 1}\langle\theta, z\rangle\Big)^2\Big].$$

This has been successfully applied to isotonic and more general shape constrained regression by Chatterjee et al. [2015], Bellec [2015].

Another prominent approach is to express the error of the LS estimator via what is known as *Chatterjee's variational formula*, proved by Chatterjee [2014] and given by

$$\|\hat{\theta} - \theta^*\|_2 = \operatorname*{argmax}_{t\geq 0}\Big(\sup_{\theta\in\Theta,\|\theta-\theta^*\|_2\leq t}\langle\theta - \theta^*, z\rangle - \frac{t^2}{2}\Big). \tag{6.12}$$

Note that the first term is related to the *Gaussian width* [see, e.g., Chandrasekaran et al., 2012] of Θ defined by $\mathbb{E}[\sup_{\theta\in\Theta}\langle\theta, z\rangle]$, whose connection to the statistical dimension was studied by Amelunxen et al. [2014]. The variational formula was first proposed for convex regression by Chatterjee [2014], and later exploited in several different settings, including matrix estimation with shape constraints by Chatterjee et al. [2017] and unimodal regression by Chatterjee and Lafferty [2015]. Similar ideas have appeared in other works, for example, analysis of empirical risk minimization [Mendelson, 2015], ranking from pairwise comparison [Shah et al., 2017] and isotonic regression [Bellec, 2015]. Bellec [2015] has used the statistical dimension approach to prove spectacularly sharp oracle inequalities that seem to be currently out of reach for methods based on Chatterjee's variational formula in Eq. (6.12). On the other hand, Chatterjee's variational formula seems more flexible as computations of the statistical dimension based on tools developed by Amelunxen et al. [2014] are currently limited to convex sets Θ with a polyhedral structure. In this chapter, we use exclusively Chatterjee's variational formula.

6.A.1 A Variational Formula for the Error of the LS Estimator

We begin the proof by stating an extension of Chatterjee's variational formula. While we only need this lemma to hold for a union of closed convex sets we present a version that holds for all closed sets. The latter extension was suggested to us by Pierre C. Bellec [2016] in a private communication.

Lemma 45. *Let $\mathcal{C}$ be a closed subset of $\mathbb{R}^d$. Suppose that $y = a^* + z$ where $a^* \in \mathcal{C}$ and $z \in \mathbb{R}^d$. Let $\hat{a} \in \operatorname{argmin}_{a\in\mathcal{C}} \|y - a\|_2^2$ be a projection of y onto $\mathcal{C}$. Define the function $f_{a^*} : \mathbb{R}_+ \to \mathbb{R}$ by*

$$f_{a^*}(t) = \sup_{a\in\mathcal{C}\cap\mathcal{B}^d(a^*,t)} \langle a - a^*, z\rangle - \frac{t^2}{2}\,.$$

Then we have

$$\|\hat{a} - a^*\|_2 \in \operatorname*{argmax}_{t\geq 0} f_{a^*}(t). \tag{6.13}$$

Moreover, if there exists $t^* > 0$ such that $f_{a^*}(t) < 0$ for all $t \geq t^*$, then $\|\hat{a} - a^*\|_2 \leq t^*$.

Proof. By definition,

$$\begin{aligned}\hat{a} &\in \operatorname*{argmin}_{a\in\mathcal{C}} \left(\|a - a^*\|_2^2 - 2\langle a - a^*, z\rangle + \|z\|_2^2\right)\\ &= \operatorname*{argmax}_{a\in\mathcal{C}} \left(\langle a - a^*, z\rangle - \frac{1}{2}\|a - a^*\|_2^2\right).\end{aligned}$$

Together with the definition of f_{a^*}, this implies that

$$\begin{aligned}f_{a^*}(\|\hat{a} - a^*\|_2) &\geq \langle \hat{a} - a^*, z\rangle - \frac{1}{2}\|\hat{a} - a^*\|_2^2\\ &\geq \sup_{a\in\mathcal{C}\cap\mathcal{B}^d(a^*,t)} \left(\langle a - a^*, z\rangle - \frac{1}{2}\|a - a^*\|_2^2\right)\\ &\geq \sup_{a\in\mathcal{C}\cap\mathcal{B}^d(a^*,t)} \langle a - a^*, z\rangle - \frac{t^2}{2} = f_{a^*}(t)\,.\end{aligned}$$

Therefore Eq. (6.13) follows.

Furthermore, suppose that there is $t^* > 0$ such that $f_{a^*}(t) < 0$ for all $t \geq t^*$. Since $f_{a^*}(\|\hat{a} - a^*\|_2) \geq f_{a^*}(0) = 0$, we have $\|\hat{a} - a^*\|_2 \leq t^*$. □

Note that this structural result holds for any error vector $z \in \mathbb{R}^d$ and any closed set $\mathcal{C}$ which is not necessarily convex. In particular, this extends the results by Chatterjee [2014] and Chatterjee and Lafferty [2015] which hold for convex sets and finite unions of convex sets respectively.

6.A.2 Proof of Theorem 10

For our purpose, we need a standard chaining bound on the supremum of a sub-Gaussian process that holds in high probability. The interested readers can see the

proof, for example, by van Handel [2014, Theorem 5.29], and refer to the monograph of Ledoux and Talagrand [1991] for a more detailed account of the technique.

Lemma 46 (Chaining tail inequality)**.** *Let $\Theta \subset \mathbb{R}^d$ and $z \sim \mathrm{subG}(\sigma^2)$ in $\mathbb{R}^d$. For any $\theta_0 \in \Theta$, it holds that*

$$\sup_{\theta\in\Theta}\langle \theta - \theta_0, z\rangle \le C\sigma \int_0^{\mathrm{diam}(\Theta)} \sqrt{\log N(\Theta, \|\cdot\|_2, \varepsilon)}\, d\varepsilon + s$$

with probability at least $1 - C\exp(-\frac{cs^2}{\sigma^2 \mathrm{diam}(\Theta)^2})$ where C and c are positive constants.

Let $\tilde{A} \in \mathcal{U}^m$. To ligthen the notation, we define two rates of estimation:

$$R_1 = R_1(\tilde{A}, n) = \sigma\Big(\sqrt{K(\tilde{A}) \log \frac{enm}{K(\tilde{A})}} + \sqrt{n \log n}\Big) \tag{6.14}$$

and

$$R_2 = R_2(\tilde{A}, n) = \sigma^2\Big(K(\tilde{A}) \log \frac{enm}{K(\tilde{A})} + n \log n\Big). \tag{6.15}$$

Note that $R_2 \le R_1^2 \le 2R_2$.

Lemma 47. *Suppose $Y = A^* + Z$ where $A^* \in \mathbb{R}^{n\times m}$ and $Z \sim \mathrm{subG}(\sigma^2)$. For $\tilde{A} \in \mathcal{U}^m$ and all $t > 0$, define*

$$f_{\tilde{A}}(t) = \sup_{A\in\mathcal{M}\cap\mathcal{B}^{nm}(\tilde{A},t)} \langle A - \tilde{A}, Y - \tilde{A}\rangle - \frac{t^2}{2}.$$

Then for any $s > 0$, it holds simultaneously for all $t > 0$ that

$$f_{\tilde{A}}(t) \le CR_1 t + t\|A^* - \tilde{A}\|_F - \frac{t^2}{2} + st \tag{6.16}$$

with probability at least $1 - C\exp(-\frac{cs^2}{\sigma^2})$, where C and c are positive constants.

Proof. Define $\Theta = \Theta_{\mathcal{M}}(\tilde{A}, 1) = \bigcup_{\lambda\ge 0}\{B - \lambda\tilde{A} : B \in \mathcal{M} \cap \mathcal{B}^{nm}(\lambda\tilde{A}, 1)\}$ (see also Definition in Eq. (6.22)). In particular, $\Theta \subset \mathcal{B}^{nm}(0, 1)$ and $0 \in \Theta$. Since $\mathcal{M}$ is a finite union of convex cones and thus is star-shaped, by scaling invariance,

$$\sup_{A\in\mathcal{M}\cap\mathcal{B}^{nm}(\tilde{A},t)} \langle A - \tilde{A}, Z\rangle = t \sup_{B\in\mathcal{M}\cap\mathcal{B}^{nm}(t^{-1}\tilde{A},1)} \langle B - t^{-1}\tilde{A}, Z\rangle \le t \sup_{M\in\Theta} \langle M, Z\rangle.$$

By Lemma 46, with probability at least $1 - C\exp(-\frac{cs^2}{\sigma^2})$,

$$\sup_{M\in\Theta} \langle M, Z\rangle \le C\sigma \int_0^2 \sqrt{\log N(\Theta, \|\cdot\|_F, \varepsilon)}\, d\varepsilon + s\,.$$

Moreover, it follows from Lemma 54 that

$$\log N(\Theta, \|\cdot\|_F, \varepsilon) \le C\varepsilon^{-1} K(\tilde{A}) \log \frac{enm}{K(\tilde{A})} + n \log n\,.$$

Combining the previous three displays, we see that

$$\begin{aligned}\sup_{A\in\mathcal{M}\cap\mathcal{B}^{nm}(\tilde{A},t)}\langle A-\tilde{A},Z\rangle &\le C\sigma t\int_0^2\sqrt{C\varepsilon^{-1}K(\tilde{A})\log\frac{enm}{K(\tilde{A})}+n\log n}\,d\varepsilon+st\\ &\le C\sigma t\sqrt{K(\tilde{A})\log\frac{enm}{K(\tilde{A})}}+C\sigma t\sqrt{n\log n}+st\\ &= CR_1t+st\end{aligned}$$

with probability at least $1-C\exp(-\frac{cs^2}{\sigma^2})$. Therefore

$$\begin{aligned}f_{\tilde{A}}(t) &= \sup_{A\in\mathcal{M}\cap\mathcal{B}^{nm}(\tilde{A},t)}\langle A-\tilde{A},Y-\tilde{A}\rangle-\frac{t^2}{2}\\ &\le \sup_{A\in\mathcal{M}\cap\mathcal{B}^{nm}(\tilde{A},t)}\langle A-\tilde{A},Z\rangle+\sup_{A\in\mathcal{M}\cap\mathcal{B}^{nm}(\tilde{A},t)}\langle A-\tilde{A},A^*-\tilde{A}\rangle-\frac{t^2}{2}\\ &\le CR_1t+st+t\|A^*-\tilde{A}\|_F-\frac{t^2}{2}\end{aligned}$$

with probability at least $1-C\exp(-\frac{cs^2}{\sigma^2})$ simultaneously for all $t>0$. □

We are now in a position to prove the adaptive oracle inequalities in Theorem 10. Recall that $(\hat{\Pi},\hat{A})$ denotes the LS estimator defined in Eq. (6.2). Without loss of generality, assume that $\Pi^*=I_n$ and $Y=A^*+Z$.

Fix $\tilde{A}\in\mathcal{U}^m$ and define $f_{\tilde{A}}$ as in Lemma 47. We can apply Lemma 45 with $a^*=\tilde{A}$, $z=Y-\tilde{A}$, $y=Y$ and $\hat{a}=\hat{\Pi}\hat{A}$ to achieve an error bound on $\|\hat{\Pi}\hat{A}-\tilde{A}\|_F$, since $\hat{\Pi}\hat{A}\in\operatorname{argmin}_{M\in\mathcal{M}}\|Y-M\|_F^2$. To be more precise, for any $s>0$ we define $t^*=3C_1R_1+2\|A^*-\tilde{A}\|_F+2s$ where C_1 is the constant in Eq. (6.16). Then it follows from Lemma 47 that with probability at least $1-C\exp(-\frac{cs^2}{\sigma^2})$, it holds for all $t\ge t^*$ that

$$f_{\tilde{A}}(t)\le C_1R_1t+t\|A^*-\tilde{A}\|_F-\frac{t^2}{2}+st<0\,.$$

Therefore by Lemma 45,

$$\|\hat{\Pi}\hat{A}-\tilde{A}\|_F\le t^*=3C_1R_1+2\|A^*-\tilde{A}\|_F+2s\,,$$

and thus

$$\|\hat{\Pi}\hat{A}-A^*\|_F\le C(R_1+\|A^*-\tilde{A}\|_F)+2s \tag{6.17}$$

with probability at least $1-C\exp(-\frac{cs^2}{\sigma^2})$.

In particular, if $s=R_1$, then $s\ge\sigma\sqrt{n+m}$ as $K(\tilde{A})\ge m$. We see that with probability at least $1-C\exp(-\frac{cs^2}{\sigma^2})\ge 1-e^{-c(n+m)}$,

$$\|\hat{\Pi}\hat{A}-A^*\|_F\lesssim R_1+\|A^*-\tilde{A}\|_F$$

and thus

$$\|\hat{\Pi}\hat{A} - A^*\|_F^2 \lesssim \|A^* - \tilde{A}\|_F^2 + \sigma^2 K(\tilde{A}) \log \frac{enm}{K(\tilde{A})} + \sigma^2 n \log n\,.$$

Finally, Eq. (6.7) follows by taking the infimum over $\tilde{A} \in \mathcal{U}^m$ on the right-hand side and dividing both sides by nm.

Next, to prove the bound in expectation, observe that Eq. (6.17) yields

$$\mathbb{P}\Big[\|\hat{\Pi}\hat{A} - A^*\|_F^2 - C(R_2 + \|A^* - \tilde{A}\|_F^2) \geq s\Big] \leq C \exp(-\frac{cs}{\sigma^2}),$$

where R_2 is defined in Eq. (6.15). Integrating the tail probability, we get that

$$\mathbb{E}\|\hat{\Pi}\hat{A} - A^*\|_F^2 - C(R_2 + \|A^* - \tilde{A}\|_F^2) \lesssim \int_0^\infty \exp(-\frac{cs}{\sigma^2})\, ds = \frac{\sigma^2}{c}$$

and therefore

$$\mathbb{E}\|\hat{\Pi}\hat{A} - A^*\|_F^2 \lesssim R_2 + \|A^* - \tilde{A}\|_F^2\,.$$

Dividing both sides by nm and minimizing over $\tilde{A} \in \mathcal{U}^m$ yields Eq. (6.8).

6.A.3 Proof of Theorem 11

In the setting of isotonic regression, Bellec and Tsybakov [2015] derived global bounds from adaptive bounds by a block approximation method, which also applies to our setting. For $k \in [n]$, let

$$\mathcal{U}_k = \big\{a \in \mathcal{U} : \mathrm{Card}(\{a_1, \dots, a_n\}) \leq k\big\}.$$

Define $k^* = \big\lceil \big(\frac{V(a)^2 n}{\sigma^2 \log(en)}\big)^{1/3} \big\rceil$. The lemma below is very similar to Lemma 2 of Bellec and Tsybakov [2015] and their proof also extends to the unimodal case with minor modifications. We present the result with proof for completeness.

Lemma 48. *For $a \in \mathcal{U}$ and $k \in [n]$, there exists $\tilde{a} \in \mathcal{U}_k$ such that*

$$\frac{1}{\sqrt{n}}\|\tilde{a} - a\|_2 \leq \frac{V(a)}{2k}. \tag{6.18}$$

In particular, there exists $\tilde{a} \in \mathcal{U}_{k^*}$ such that

$$\frac{1}{n}\|\tilde{a} - a\|_2^2 \leq \frac{1}{4} \max\Big(\Big(\frac{\sigma^2 V(a) \log(en)}{n}\Big)^{2/3}, \frac{\sigma^2 \log(en)}{n}\Big).$$

Moreover,

$$\frac{\sigma^2 k^*}{n} \log(en) \leq 2 \max\Big(\Big(\frac{\sigma^2 V(a) \log(en)}{n}\Big)^{2/3}, \frac{\sigma^2 \log(en)}{n}\Big).$$

Proof. Let $\underline{a} = \min(a_1, a_n)$, $\bar{a} = \max_{i\in[n]} a_i$ and $i_0 \in \mathrm{argmax}_{i\in[n]} a_i$. For $j \in [k-1]$,

consider the intervals

$$I_j = \Big[\underline{a} + \frac{j-1}{k}V(a), \underline{a} + \frac{j}{k}V(a)\Big],$$

and $I_k = \Big[\underline{a} + \frac{k-1}{k}V(a), \bar{a}\Big]$. Also for $j \in [k]$, let $J_j = \{i \in [n] : a_i \in I_j\}$. We define the vector $\tilde{a} \in \mathbb{R}^n$ by $\tilde{a}_i = \underline{a} + \frac{j-1/2}{k}V(a)$ for $i \in [n]$, where j is uniquely determined by $i \in I_j$. Since a is increasing on $\{1, \dots, i_0\}$ and decreasing $\{i_0, \dots, n\}$, so is $\tilde{a}$. Thus $\tilde{a} \in \mathcal{U}_k$. Moreover, $|\tilde{a}_i - a_i| \le \frac{V(a)}{2k}$ for $i \in [n]$, which implies Eq. (6.18).

Next we prove the latter two assertions. Since $k^* = \lceil \big(\frac{V(a)^2 n}{\sigma^2 \log(en)}\big)^{1/3} \rceil$, if $\tilde{a} \in \mathcal{U}_{k^*}$ and $k^* = 1$ then

$$\frac{1}{n}\|\tilde{a} - a\|_2^2 \le \frac{V(a)^2}{4} \le \frac{\sigma^2}{4n}\log(en)$$

and

$$\frac{\sigma^2 k^*}{n}\log(en) = \frac{\sigma^2}{n}\log(en).$$

On the other hand, if $k^* > 1$, then

$$\frac{1}{n}\|\tilde{a} - a\|_2^2 \le \frac{V(a)}{4(k^*)^2} \le \frac{1}{4}\Big(\frac{\sigma^2 V(a)\log(en)}{n}\Big)^{2/3}$$

and

$$\frac{\sigma^2 k^*}{n}\log(en) \le 2\Big(\frac{\sigma^2 V(a)\log(en)}{n}\Big)^{2/3}.$$

□

It is straightforward to generalize the lemma to matrices. For $\mathbf{k} \in [n]^m$, we write $\mathbf{k} = (k_1, \dots, k_m)$ and let

$$\mathcal{U}_{\mathbf{k}}^m = \{A \in \mathcal{U}^m : \mathrm{Card}(\{A_{1,j}, \dots, A_{n,j}\}) = k_j \text{ for } 1 \le j \le m\}.$$

Then $K(A) = \sum_{j=1}^m k_j$ for $A \in \mathcal{U}_{\mathbf{k}}^m$. Define $\mathbf{k}^*$ by

$$k_j^* = \Big\lceil \Big(\frac{V(A_{\cdot,j})^2 n}{\sigma^2 \log(en)}\Big)^{1/3} \Big\rceil.$$

Lemma 49. *For $A \in \mathcal{U}^m$, there exists $\tilde{A} \in \mathcal{U}_{\mathbf{k}^*}^m$ such that*

$$\frac{1}{nm}\|\tilde{A} - A\|_F^2 \le \frac{1}{4}\Big(\frac{\sigma^2 V(A)\log(en)}{n}\Big)^{2/3} + \frac{\sigma^2}{4n}\log(en)$$

and

$$\frac{\sigma^2 K(\tilde{A})}{nm}\log(en) \le 2\Big(\frac{\sigma^2 V(A)\log(en)}{n}\Big)^{2/3} + \frac{2\sigma^2}{n}\log(en)\,.$$

Proof. Applying Lemma 48 to columns of A, we see that there exists $\tilde{A} \in \mathcal{U}_{\mathbf{k}^*}^m$ such

that

$$\frac{1}{n}\|\tilde{A}_{\cdot,j} - A_{\cdot,j}\|_2^2 \le \frac{1}{4}\max\Big(\big(\frac{\sigma^2 V(A_{\cdot,j})\log(en)}{n}\big)^{2/3}, \frac{\sigma^2}{n}\log(en)\Big)$$

and

$$\frac{\sigma^2 k_j^*}{n}\log(en) \le 2\max\Big(\big(\frac{\sigma^2 V(A_{\cdot,j})\log(en)}{n}\big)^{2/3}, \frac{\sigma^2}{n}\log(en)\Big).$$

Summing over $1 \le j \le m$, we get that

$$\begin{aligned}\frac{1}{nm}\|\tilde{A} - A\|_F^2 &\le \frac{1}{4m}\Big(\frac{\sigma^2\log(en)}{n}\Big)^{2/3}\sum_{j=1}^m V(A_{\cdot,j})^{2/3} + \frac{\sigma^2\log(en)}{4n} \\ &= \frac{1}{4}\Big(\frac{\sigma^2 V(A)\log(en)}{n}\Big)^{2/3} + \frac{\sigma^2}{4n}\log(en)\,,\end{aligned}$$

and similarly

$$\frac{\sigma^2 K(\tilde{A})}{nm}\log(en) \le 2\Big(\frac{\sigma^2 V(A)\log(en)}{n}\Big)^{2/3} + \frac{2\sigma^2}{n}\log(en)\,.$$

□

For $A \in \mathcal{U}^m$, choose $\tilde{A} \in \mathcal{U}^m_{\mathbf{k}^*}$ according to Lemma 49. Then

$$\begin{aligned}\frac{1}{nm}\|\tilde{A} - A^*\|_F^2 &\le \frac{2}{nm}\|A - A^*\|_F^2 + \frac{2}{nm}\|\tilde{A} - A\|_F^2 \\ &\le \frac{2}{nm}\|A - A^*\|_F^2 + \frac{5}{4}\Big(\frac{\sigma^2 V(A)\log n}{n}\Big)^{2/3} + \frac{5\sigma^2}{4n}\log n \qquad (6.19)\end{aligned}$$

by noting that $\log(en) \le 2.5\log n$ for $n \ge 2$, and similarly

$$\frac{\sigma^2 K(\tilde{A})}{nm}\log(en) \le 5\Big(\frac{\sigma^2 V(A)\log n}{n}\Big)^{2/3} + \frac{5\sigma^2}{n}\log n\,. \qquad (6.20)$$

Plugging Eq. (6.19) and Eq. (6.20) into the right-hand side of Eq. (6.7) and Eq. (6.8), and then minimizing over $A \in \mathcal{U}^m$, we complete the proof.

6.B Metric Entropy

In this section, we study various *covering numbers* or *metric entropy* related to the parameter space of the model in Eq. (6.1). First recall some standard definitions that date back at least to Kolmogorov and Tihomirov [1961]. An ε-net of a subset $G \subset \mathbb{R}^n$ with respect to a norm $\|\cdot\|$ is a set $\{w_1, \cdots, w_N\} \subset G$ such that for any $w \in G$, there exists $i \in [N]$ for which $\|w - w_i\| \le \varepsilon$. The covering number $N(G, \|\cdot\|, \varepsilon)$ is the cardinality of the smallest ε-net with respect to the norm $\|\cdot\|$. Metric entropy is defined as the logarithm of a covering number. In the following, we will consider the Euclidean norm unless otherwise specified.

6.B.1 Cartesian Product of Cones

Lemma 51 below bounds covering numbers of product spaces and is useful in later proofs. We start with a well-known result on the covering number of a Euclidean ball with respect to the ℓ_∞-norm [see, e.g., Massart, 2007, Lemma 7.14, for an analogous result].

Lemma 50. *For any $\varepsilon \in (0, 1]$,*

$$N\Big(\mathcal{B}^m(0,1), \|\cdot\|_\infty, \frac{\varepsilon}{\sqrt{m}}\Big) \le (C/\varepsilon)^m,$$

for some constant $C > 0$.

Proof. We aim at bounding the covering number of a Euclidean ball by cubes. Let $\{x^1, \dots, x^M\}$ be a maximal $\frac{\varepsilon}{\sqrt{m}}$-packing of $\mathcal{B}^m(0,1)$ with respect to the ℓ_∞-norm, where a δ-packing of a set G with respect to a norm $\|\cdot\|$ is a set $\{w_1, \cdots, w_N\} \subset G$ such that $\|w_i - w_j\| \ge \delta$ for all distinct $i, j \in [N]$. Then this set is necessarily an $\frac{\varepsilon}{\sqrt{m}}$-net of $\mathcal{B}^m(0,1)$ by maximality, so $N(\mathcal{B}^m(0,1), \|\cdot\|_\infty, \frac{\varepsilon}{\sqrt{m}}) \le M$. Consider the cubes with side length $\frac{\varepsilon}{\sqrt{m}}$ centered at x^i for $1 \le i \le M$. These cubes are disjoint and contained in the set $\mathcal{B}^m(0,1) + Q^m(\frac{\varepsilon}{\sqrt{m}})$, where $Q^m(\frac{\varepsilon}{\sqrt{m}})$ is the cube with side length $\frac{\varepsilon}{\sqrt{m}}$ centered at the origin in $\mathbb{R}^m$. Since $Q^m(\frac{\varepsilon}{\sqrt{m}}) \subset \mathcal{B}^m(0, \varepsilon)$,

$$\begin{aligned} M \operatorname{Vol}\Big(Q^m\big(\frac{\varepsilon}{\sqrt{m}}\big)\Big) &\le \operatorname{Vol}\Big(\mathcal{B}^m(0,1) + Q^m\big(\frac{\varepsilon}{\sqrt{m}}\big)\Big) \\ &\le \operatorname{Vol}(\mathcal{B}^m(0, 1+\varepsilon)) \\ &\le \operatorname{Vol}(\mathcal{B}^m(0,2)). \end{aligned}$$

This proves the following bound on the covering number in terms of a volume ratio:

$$N\Big(\mathcal{B}^m(0,1), \|\cdot\|_\infty, \frac{\varepsilon}{\sqrt{m}}\Big) \le \frac{\operatorname{Vol}(\mathcal{B}^m(0,2))}{\operatorname{Vol}(Q^m(\frac{\varepsilon}{\sqrt{m}}))} \le \frac{C^m m^{-m/2}}{\varepsilon^m m^{-m/2}} = (C/\varepsilon)^m.$$

□

Now we study the metric entropy of a Cartesian product of convex cones. Let $\{I_i\}_{i=1}^m$ be a partition of $[n]$ with $|I_i| = n_i$ and $\sum_{i=1}^m n_i = n$. For $a \in \mathbb{R}^n$, the restriction of a to the coordinates in I_i is denoted by $a_{I_i} \in \mathbb{R}^{n_i}$. Let $\mathcal{C}_i$ be a convex cone in $\mathbb{R}^{n_i}$ and $\mathcal{C} = \mathcal{C}_1 \times \cdots \times \mathcal{C}_m$.

Lemma 51. *With the notation above, suppose that $a_{I_i} \in \mathcal{C}_i \cap (-\mathcal{C}_i)$. Then for any $t > 0$ and $\varepsilon \in (0, t]$,*

$$\log N\big(\mathcal{C} \cap \mathcal{B}^n(a,t), \|\cdot\|_2, \varepsilon\big) \le m \log \frac{Ct}{\varepsilon} + \sum_{i=1}^m \log N\Big(\mathcal{C}_i \cap \mathcal{B}^{n_i}(a_{I_i}, t), \|\cdot\|_2, \frac{\varepsilon}{3}\Big)$$

for some constant $C > 0$.

Proof. Since a product of balls $\mathcal{B}^{n_1}(0, \frac{\varepsilon}{\sqrt{m}}) \times \cdots \times \mathcal{B}^{n_m}(0, \frac{\varepsilon}{\sqrt{m}})$ is contained in $\mathcal{B}^n(0, \varepsilon)$, one could try to cover $\mathcal{C} \cap \mathcal{B}^n(a, t)$ by such products of balls. It turns out that this yields an upper bound of order $m^{3/2}$, which is too loose for our purpose. Fortunately, the following argument corrects this dependency.

Without loss of generality, we assume that $t = 1$. We construct a 3ε-net of $\mathcal{C} \cap \mathcal{B}^n(a, 1)$ as follows. First, let $\mathcal{N}_\mathcal{B}$ be an $\frac{\varepsilon}{2\sqrt{m}}$-net of $\mathcal{B}^m(0, 1)$ with respect to the ℓ_∞-norm. Define

$$\mathcal{N}_\mathcal{D} = \Big\{\mu \in \mathcal{N}_\mathcal{B} : \min_{i\in[m]} \mu_i \geq -\frac{1}{2\sqrt{m}}\Big\}.$$

Note that $\mu_i + \frac{1}{\sqrt{m}} > 0$ for $\mu \in \mathcal{N}_\mathcal{D}$, and let $\mathcal{N}_{\mu_i}$ be a $(\mu_i + \frac{1}{\sqrt{m}})\varepsilon$-net of $\mathcal{C}_i \cap \mathcal{B}^{n_i}(a_{I_i}, \mu_i + \frac{1}{\sqrt{m}})$. Define $\mathcal{N}_\mu = \mathcal{N}_{\mu_1} \times \cdots \times \mathcal{N}_{\mu_m}$, i.e.,

$$\mathcal{N}_\mu = \{w \in \mathbb{R}^n : w = (w_{I_1}, \cdots, w_{I_m}),\ w_{I_i} \in \mathcal{N}_{\mu_i}\}.$$

We claim that $\bigcup_{\mu\in\mathcal{N}_\mathcal{D}} \mathcal{N}_\mu$ is an 3ε-net of $\mathcal{C} \cap \mathcal{B}^n(a, 1)$.

Fix $v \in \mathcal{C} \cap \mathcal{B}^n(a, 1)$. Let $v_{I_i} \in \mathbb{R}^{n_i}$ be the restriction of v to the component space $\mathbb{R}^{n_i}$. Then $v_{I_i} \in \mathcal{C}_i$. Let $\lambda \in \mathbb{R}^m$ be defined by $\lambda_i = \|v_{I_i} - a_{I_i}\|_2$, so $\|\lambda\|_2 = \|v - a\|_2 \leq 1$. Hence we can find $\mu \in \mathcal{N}_\mathcal{B}$ such that $\|\mu - \lambda\|_\infty \leq \frac{\varepsilon}{2\sqrt{m}}$. In particular, for all $i \in [m]$, $\mu_i \geq \lambda_i - \frac{\varepsilon}{2\sqrt{m}} \geq -\frac{1}{2\sqrt{m}}$, so $\mu \in \mathcal{N}_\mathcal{D}$. Moreover, $\|v_{I_i} - a_{I_i}\|_2 = \lambda_i < \mu_i + \frac{1}{\sqrt{m}}$ and $v_{I_i} \in \mathcal{C}_i$, so by definition of $\mathcal{N}_{\mu_i}$, there exists $w_{I_i} \in \mathcal{N}_{\mu_i}$ such that $\|w_{I_i} - v_{I_i}\|_2 \leq (\mu_i + \frac{1}{\sqrt{m}})\varepsilon$. Let $w = (w_{I_1}, \ldots, w_{I_m}) \in \mathcal{N}_\mu$. Since

$$\sum_{i=1}^m \mu_i^2 \leq \sum_{i=1}^m (\lambda_i + |\lambda_i - \mu_i|)^2 \leq \sum_{i=1}^m 2\lambda_i^2 + \frac{\varepsilon^2}{2} \leq \frac{5}{2},$$

we conclude that

$$\|w - v\|_2^2 \leq \sum_{i=1}^m \Big(\mu_i + \frac{1}{\sqrt{m}}\Big)^2 \varepsilon^2 \leq 7\varepsilon^2.$$

Therefore $\bigcup_{\mu\in\mathcal{N}_\mathcal{D}} \mathcal{N}_\mu$ is a 3ε-net of $\mathcal{C} \cap \mathcal{B}^n(a, 1)$.

It remains to bound the cardinality of this net. By Lemma 50, $|\mathcal{N}_\mathcal{D}| \leq |\mathcal{N}_\mathcal{B}| \leq (C/\varepsilon)^m$. Moreover, recall that $\mathcal{N}_{\mu_i}$ is a $(\mu_i + \frac{1}{\sqrt{m}})\varepsilon$-net of $\mathcal{C}_i \cap \mathcal{B}^{n_i}(a_{I_i}, \mu_i + \frac{1}{\sqrt{m}})$. Since $a_{I_i} \in \mathcal{C}_i \cap (-\mathcal{C}_i)$, for any $t > 0$, $\mathcal{C}_i \cap \mathcal{B}^{n_i}(a_{I_i}, t) = \{x + a_{I_i} : x \in \mathcal{C}_i \cap \mathcal{B}^{n_i}(0, t)\}$. Hence we can choose the net so that

$$\begin{aligned}
|\mathcal{N}_{\mu_i}| &= N\Big(\mathcal{C}_i \cap \mathcal{B}^{n_i}\big(0, \mu_i + \frac{1}{\sqrt{m}}\big), \|\cdot\|_2, \big(\mu_i + \frac{1}{\sqrt{m}}\big)\varepsilon\Big) \\
&= N(\mathcal{C}_i \cap \mathcal{B}^{n_i}(0, 1), \|\cdot\|_2, \varepsilon) \\
&= N(\mathcal{C}_i \cap \mathcal{B}^{n_i}(a_{I_i}, 1), \|\cdot\|_2, \varepsilon).
\end{aligned}$$

As $|\mathcal{N}_\mu| \leq \prod_{i=1}^m |\mathcal{N}_{\mu_i}|$, therefore

$$\Big|\bigcup_{\mu\in\mathcal{N}_\mathcal{D}} \mathcal{N}_\mu\Big| \leq \Big(\frac{C}{\varepsilon}\Big)^m \prod_{i=1}^m N(\mathcal{C}_i \cap \mathcal{B}^{n_i}(a_{I_i}, 1), \|\cdot\|_2, \varepsilon).$$

Taking the logarithm completes the proof. □

6.B.2 Unimodal Vectors and Matrices

Recall that $\mathcal{S}_n$ denotes the closed convex cone of increasing vectors in $\mathbb{R}^n$. First, we prove a result on the metric entropy of $\mathcal{S}_n$ intersecting with a ball using Lemma 51.

Lemma 52. *Let $b \in \mathbb{R}^n$ be such that $b_1 = \cdots = b_n$. Then for any $t > 0$ and $\varepsilon > 0$,*

$$\log N(\mathcal{S}_n \cap \mathcal{B}^n(b,t), \|\cdot\|_2, \varepsilon) \le C\varepsilon^{-1} t \log(en).$$

Proof. The majority of the proof is due to Lemma 5.1 in an old version of the article of Chatterjee and Lafferty [2015], but we improve their result by a factor $\sqrt{\log n}$ and provide the whole proof for completeness.

The bound holds trivially if $\varepsilon > t$, since the left-hand side is zero. It also clearly holds when $n = 1$. Hence we can assume without loss of generality that $\varepsilon \le t$ and $n = 2n' \ge 2$. Moreover, assume that $t = 1$ for simplicity and the proof will work for any $t > 0$. Let $I = \{1, \dots, n'\}$ and observe that

$$\log N(\mathcal{S}_n \cap \mathcal{B}^n(b,1), \|\cdot\|_2, \varepsilon) \le 2 \log N(\mathcal{S}_{n'} \cap \mathcal{B}^{n'}(b_I,1), \|\cdot\|_2, \varepsilon/\sqrt{2}\,).$$

Let k be the smallest integer for which $2^k > n'$. We partition I into k blocks $A_j = I \cap [2^j, 2^{j+1})$ for $j \in [k]$ and let $m_j = |A_j|$. Since $\mathcal{S}_{n'} \subset \mathcal{S}_{m_1} \times \cdots \times \mathcal{S}_{m_k}$, Lemma 51 yields that

$$\log N\big(\mathcal{S}_{n'} \cap \mathcal{B}^{n'}(b_I,1), \|\cdot\|_2, \varepsilon/\sqrt{2}\,\big) \le k \log \frac{C}{\varepsilon} + \sum_{j=1}^{k} \log N\Big(\mathcal{S}_{m_j} \cap \mathcal{B}^{m_j}(b_{A_j},1), \|\cdot\|_2, \frac{\varepsilon}{3\sqrt{2}}\Big). \quad (6.21)$$

We know from Chatterjee [2014, Lemma 4.20] that for any $c \le d$ and $n \ge 1$,

$$\log N\Big(\mathcal{S}_n \cap [c,d]^n \cap \mathcal{B}^n(b,1), \|\cdot\|_2, \varepsilon\Big) \le \frac{C\sqrt{n}(d-c)}{\varepsilon}.$$

For each $a \in \mathcal{S}_n \cap \mathcal{B}^n(0,1)$, it holds that $|a_i| \le \frac{1}{\sqrt{i}}$ for $i \in I$ (since either $|a_l| \ge |a_i|$ for all $l \le i$ or $|a_l| \ge |a_i|$ for all $i \le l \le n$; see e.g. Dai et al. [2014]), so $\max_{i \in A_j} |a_i| \le 2^{-j/2}$. Also $m_j \le 2^j$, so we get that

$$\log N\Big(\mathcal{S}_{m_j} \cap \mathcal{B}^{m_j}(b_{A_j},1), \|\cdot\|_2, \frac{\varepsilon}{3\sqrt{2}}\Big) \le \frac{C}{\varepsilon}$$

for all $j \in [k]$. Substituting this bound into Eq. (6.21) and noting that $k \le \log_2 n$, we reach the conclusion

$$\log N(\mathcal{S}_{n'} \cap \mathcal{B}^{n'}(b_I,1), \|\cdot\|_2, \varepsilon/\sqrt{2}) \le C\varepsilon^{-1} \log(en).$$

□

Next, we study the metric entropy of the set of matrices with unimodal columns.

Recall that $\mathcal{C}_l = \{a \in \mathbb{R}^n : a_1 \le \cdots \le a_l\} \cap \{a \in \mathbb{R}^n : a_l \ge \cdots \ge a_n\}$ for $l \in [n]$. For $\mathbf{l} = (l_1, \ldots, l_m) \in [n]^m$, define $\mathcal{C}_{\mathbf{l}}^m = \mathcal{C}_{l_1} \times \cdots \times \mathcal{C}_{l_m}$. Moreover, for $A \in \mathbb{R}^{n\times m}$, $t > 0$ and $\mathcal{C} \subset \mathbb{R}^{n\times m}$, define

$$\begin{aligned}\Theta_{\mathcal{C}}(A,t) &= \bigcup_{\lambda\ge 0}\{B - \lambda A : B \in \mathcal{C} \cap \mathcal{B}^{nm}(\lambda A, t)\} \\ &= \bigcup_{\lambda\ge 0}\Big(\mathcal{C} \cap \mathcal{B}^{nm}(\lambda A, t) - \lambda A\Big).\end{aligned} \tag{6.22}$$

Note that in particular $\Theta_{\mathcal{C}}(A,t) \subset \mathcal{B}^{nm}(0,t)$.

Lemma 53. *Given $A \in \mathbb{R}^{n\times m}$ and $\mathbf{l} = (l_1, \ldots, l_m) \in [n]^m$, define $K(A) = \sum_{j=1}^m k(A_{\cdot,j})$ and $k(A_{\cdot,j}) = \mathrm{Card}(\{A_{1,j}, \ldots, A_{n,j}\})$. Then for any $t > 0$ and $\varepsilon > 0$,*

$$\log N\big(\Theta_{\mathcal{C}_{\mathbf{l}}^m}(A,t), \|\cdot\|_F, \varepsilon\big) \le C\varepsilon^{-1} t\, K(A) \log \frac{enm}{K(A)}.$$

Proof. Assume that $\varepsilon \le t$ since otherwise the left-hand side is zero and the bound holds trivially. For $j \in [m]$, define $I^{j,1} = [l_j]$ and $I^{j,2} = [n]\backslash[l_j]$. Define $k_{j,1} = k(A_{I^{j,1},j})$ and $k_{j,2} = k(A_{I^{j,2},j})$. Let $\varkappa = \sum_{j=1}^m (k_{j,1} + k_{j,2})$ and observe that $K(A) \le \varkappa \le 2K(A)$. Moreover, let $\{I_1^{j,1}, \ldots, I_{k_{j,1}}^{j,1}\}$ be the partition of $I^{j,1}$ such that $A_{I_i^{j,1},j}$ is a constant vector for $i \in [k_{j,1}]$. Note that elements of $I_i^{j,1}$ need not to be consecutive. Define the partition for $I^{j,2}$ analogously.

For $j \in [m]$ and $i \in [k_{j,1}]$ (resp. $[k_{j,2}]$), let $\mathcal{S}_{I_i^{j,1},j}$ (resp. $\mathcal{S}_{I_i^{j,2},j}$) denote the set of increasing (resp. decreasing) vectors in the component space $\mathbb{R}^{|I_i^{j,1}|}$ (resp. $\mathbb{R}^{|I_i^{j,2}|}$). Lemma 52 implies that

$$\log N(\mathcal{S}_{I_i^{j,r},j} \cap \mathcal{B}^{|I_i^{j,r}|}(A_{I_i^{j,r},j}, t), \|\cdot\|_F, \varepsilon) \le C\varepsilon^{-1} t \log(e|I_i^{j,r}|).$$

As a matrix in $\mathbb{R}^{n\times m}$ can be viewed as a concatenation of $\varkappa = \sum_{j=1}^m (k_{j,1} + k_{j,2})$ vectors of length $|I_i^{j,r}|, r \in [2], j \in [m]$, we define the cone $\mathcal{S}^*$ in $\mathbb{R}^{n\times m}$ by $\mathcal{S}^* = \prod_{j=1}^m \prod_{r=1}^2 \prod_{i=1}^{k_{j,r}} \mathcal{S}_{I_i^{j,r},j}$, which is clearly a superset of $\mathcal{C}_{\mathbf{l}}^m$. It also follows that $A \in \mathcal{S}^* \cap (-\mathcal{S}^*)$, and thus by Lemma 51 and the previous display,

$$\begin{aligned}\log N(\mathcal{S}^* \cap \mathcal{B}^{nm}(A,t), \|\cdot\|_F, \varepsilon) &\le \varkappa \log \frac{Ct}{\varepsilon} + \sum_{j=1}^m \sum_{r=1}^2 \sum_{i=1}^{k_{j,r}} C\varepsilon^{-1} t \log(e|I_i^{j,r}|) \\ &\le C\varepsilon^{-1} t\, \varkappa + C\varepsilon^{-1} t\, \varkappa \log \frac{e \sum_{j,r,i} |I_i^{j,r}|}{\varkappa} \\ &\le C\varepsilon^{-1} t\, K(A) \log \frac{enm}{K(A)},\end{aligned}$$

where we used the concavity of the logarithm and Jensen's inequality in the second step, and that $K(A) \le \varkappa \le 2K(A)$ in the last step.

Since $A \in \mathcal{S}^* \cap (-\mathcal{S}^*)$ (the cone $\mathcal{S}^*$ is pointed at A) we have that $\mathcal{S}^* \cap \mathcal{B}^{nm}(\lambda A, t) -$

$\lambda A = \mathcal{S}^* \cap \mathcal{B}^{nm}(0,t)$ for any $\lambda \geq 0$. In view of Definition in Eq. (6.22), it holds

$$\Theta_{\mathcal{S}^*}(A,t) = \bigcup_{\lambda \geq 0} \mathcal{S}^* \cap \mathcal{B}^{nm}(\lambda A, t) - \lambda A = \mathcal{S}^* \cap \mathcal{B}^{nm}(\lambda A, t) - \lambda A\,, \quad \forall \lambda \geq 0\,.$$

In particular, taking $\lambda = 1$, we get $\Theta_{\mathcal{S}^*}(A,t) = \mathcal{S}^* \cap \mathcal{B}^{nm}(A,t) - A$. Moreover, $\mathcal{C}_{\mathrm{I}}^m \subset \mathcal{S}^*$, so that $\Theta_{\mathcal{C}_{\mathrm{I}}^m}(A,t) \subset \Theta_{\mathcal{S}^*}(A,t) = \mathcal{S}^* \cap \mathcal{B}^{nm}(A,t) - A$. Thus the metric entropy of $\Theta_{\mathcal{C}_{\mathrm{I}}^m}(A,t)$ is subject to the above bound as well. □

Finally, we consider the metric entropy of $\Theta_{\mathcal{M}}(A,t)$ for $A \in \mathbb{R}^{n\times m}$, $t > 0$ and $\mathcal{M} = \bigcup_{\Pi \in \mathfrak{S}_n} \Pi \mathcal{U}^m$. The above analysis culminates in the following lemma which we use to prove the main upper bounds.

Lemma 54. *Let $A \in \mathbb{R}^{n\times m}$ and $K(A)$ be defined as in the previous lemma. Then for any $\varepsilon > 0$ and $t > 0$,*

$$\log N\big(\Theta_{\mathcal{M}}(A,t), \|\cdot\|_F, \varepsilon\big) \leq C\varepsilon^{-1} t\, K(A) \log \frac{enm}{K(A)} + n \log n.$$

Proof. Assume that $\varepsilon \leq t$ since otherwise the left-hand side is zero and the bound holds trivially. Note that $\mathcal{U}^m = \bigcup_{\mathrm{I} \in [n]^m} \mathcal{C}_{\mathrm{I}}^m$, and that $\mathcal{M} = \bigcup_{\Pi \in \mathfrak{S}_n} \Pi \mathcal{U}^m$. Thus $\mathcal{M}$ is the union of $n^m n!$ cones of the form $\Pi \mathcal{C}_{\mathrm{I}}^m$. By Definition in Eq. (6.22), $\Theta_{\mathcal{M}}(A,t)$ is also the union of $n^m n!$ sets $\Theta_{\Pi \mathcal{C}_{\mathrm{I}}^m}(A,t)$, each having metric entropy subject to the bound in Lemma 53. Therefore, a union bound implies that

$$\begin{aligned}
\log N\big(\Theta_{\mathcal{M}}(A,t), \|\cdot\|_F, \varepsilon\big) &\leq \log N\big(\Theta_{\mathcal{C}_{\mathrm{I}}^m}(A,t), \|\cdot\|_F, \varepsilon\big) + \log(n^m n!) \\
&\leq C\varepsilon^{-1} t\, K(A) \log \frac{enm}{K(A)} + m \log n + n \log n \\
&\leq C\varepsilon^{-1} t\, K(A) \log \frac{enm}{K(A)} + n \log n,
\end{aligned}$$

where the last step follows from that $K \log(enm/K) \geq m \log n$ for $m \leq K \leq nm$ and that $\varepsilon \leq t$. □

6.C Proof of the Lower Bounds

For minimax lower bounds, we consider the model $Y = \Pi^* A^* + Z$ where entries of Z are i.i.d. $N(0, \sigma^2)$. The Varshamov-Gilbert lemma [Massart, 2007, Lemma 4.7] is a standard tool for proving lower bounds.

Lemma 55 (Varshamov-Gilbert)**.** *Let δ denote the Hamming distance on $\{0,1\}^d$ where $d \geq 2$. Then there exists a subset $\Omega \subset \{0,1\}^d$ such that $\log|\Omega| \geq d/8$ and $\delta(\omega, \omega') \geq d/4$ for distinct $\omega, \omega' \in \Omega$.*

We also need the following useful lemma.

Lemma 56. *Consider the model $y = \theta + z$ where $\theta \in \Theta \subset \mathbb{R}^d$ and $z \sim N(0, \sigma^2 I_d)$. Suppose that $|\Theta| \geq 3$ and for distinct $\theta, \theta' \in \Theta$, $4\phi \leq \|\theta - \theta'\|_2^2 \leq \frac{\sigma^2}{8} \log |\Theta|$ where $\phi > 0$. Then there exists $c > 0$ such that*

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{P}_\theta \big[\|\hat{\theta} - \theta\|_F^2 \geq \phi \big] \geq c.$$

Proof. Let $\mathbb{P}_\theta$ denote the probability with respect to $\theta + z$. Then the Kullback-Leibler divergence between $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta'}$ satisfies that

$$\mathrm{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \frac{\|\theta - \theta'\|_F^2}{2\sigma^2} \leq \frac{\log |\Theta|}{16} \leq \frac{\log(|\Theta| - 1)}{10},$$

since $|\Theta| \geq 3$. Applying [Tsybakov, 2009, Theorem 2.5] with $\alpha = \frac{1}{10}$ gives the conclusion. □

6.C.1 Proof of Theorem 12

We define $\mathcal{U}_{K_0}^m(V_0) = \mathcal{U}_{K_0}^m \cap \mathcal{U}^m(V_0)$ and $\mathcal{M}_{K_0}(V_0) = \bigcup_{\Pi \in \mathfrak{S}_n} \Pi \mathcal{U}_{K_0}^m(V_0)$. Define the subset of $\mathcal{M}_{K_0}(V_0)$ containing permutations of monotonic matrices by $\mathcal{M}_{K_0}^{\mathcal{S}}(V_0) = \{\Pi A \in \mathcal{M}_{K_0}(V_0) : \Pi \in \mathfrak{S}_n, A \in \mathcal{S}^m\}$. Since each estimator pair $(\hat{\Pi}, \hat{A})$ gives an estimator $\hat{M} = \hat{\Pi}\hat{A}$ of $M = \Pi A$, it suffices to prove a lower bound on $\|\hat{M} - M\|_F^2$. In fact, we prove a stronger lower bound than the one in Theorem 12.

Proposition 21. *Suppose that $K_0 \leq m(\frac{16n}{\sigma^2})^{1/3} V_0^{2/3} - m$. Then*

$$\inf_{\hat{M}} \sup_{M \in \mathcal{M}_{K_0}(V_0)} \mathbb{P}_M \Big[\frac{1}{nm} \|\hat{M} - M\|_F^2 \geq c\sigma^2 \frac{K_0}{nm} + c \max_{1 \leq l \leq \min(K_0 - m, m) + 1} \min \Big(\frac{\sigma^2}{m} \log l, m^2 l^{-3} V_0^2 \Big) \Big] \geq c' \quad (6.23)$$

for some $c, c' > 0$, where $\mathbb{P}_M$ is the probability with respect to $Y = M + Z$. This bound remains valid for the parameter subset $\mathcal{M}_{K_0}^{\mathcal{S}}(V_0)$ if $l = 1$ or 2.

Note that the bound clearly holds for the larger parameter space $\mathcal{M}_{K_0} = \bigcup_{\Pi \in \mathfrak{S}_n} \Pi \mathcal{U}_{K_0}^m$. By taking $l = \min(K_0 - m, m) + 1$ and V_0 large enough, we see that the assumption in Proposition 21 is satisfied and the second term becomes simply $\frac{\sigma^2}{m} \log l$, so Theorem 12 follows. In the monotonic case, by the last statement of the proposition, if $K_0 \geq m + 1$ then taking $l = 2$ and V_0 large enough yields a lower bound of rate $\sigma^2(\frac{K_0}{nm} + \frac{1}{m})$ for the set of matrices A with increasing columns and $K(A) \leq K_0$.

The proof of Proposition 21 has two parts which correspond to the two terms respectively. First, the term $\sigma^2 \frac{K_0}{nm}$ is derived from the proof of lower bounds for isotonic regression by Bellec and Tsybakov [2015]. Then we derive the other term $\frac{\sigma^2}{m} \log l$ for any $1 \leq l \leq \min(K_0 - m, m) + 1$, which is due to the unknown permutation.

Lemma 57. *Suppose that* $K_0 \le m(\frac{16n}{\sigma^2})^{1/3}V_0^{2/3} - m$. *For some* $c, c' > 0$,

$$\inf_{\hat{M}} \sup_{M \in \mathcal{M}_{K_0}^{\mathcal{S}}(V_0)} \mathbb{P}_M\big[\|\hat{M} - M\|_F^2 \ge c\sigma^2 K_0\big] \ge c\,,$$

where $\mathbb{P}_M$ *is the probability with respect to* $Y = M + Z$.

Proof. We adapt the proof of Bellec and Tsybakov [2015, Theorem 4] to the case of matrices. Let $V_j = V_0$ for all $j \in [m]$. Since

$$K_0 \le m\big(\frac{16n}{\sigma^2}\big)^{1/3}V_0^{2/3} - m = \sum_{j=1}^{m}\Big[\big(\frac{16n}{\sigma^2}\big)^{1/3}V_j^{2/3} - 1\Big],$$

we can choose $k_j \in [n]$ so that $k_j \le (\frac{16n}{\sigma^2})^{1/3}V_j^{2/3}$ and $K_0 = \sum_{j=1}^m k_j$. According to Lemma 55, there exists $\Omega \subset \{0,1\}^{K_0}$ such that $\log|\Omega| \ge K_0/8$ and $\delta(\omega, \omega') \ge K_0/4$ for distinct $\omega, \omega' \in \Omega$. Consider the partition $[K_0] = \cup_{m=1}^j I_j$ with $|I_j| = k_j$. For each $\omega \in \Omega$, let $\omega^j \in \{0,1\}^{k_j}$ be the restriction of ω to coordinates in I_j. Define $M^\omega \in \mathbb{R}^{n\times m}$ by

$$M_{i,j}^\omega = \frac{\lfloor (i-1)k_j/n \rfloor V_j}{2k_j} + \gamma_j \omega_{\lfloor (i-1)k_j/n\rfloor + 1},$$

where $\gamma_j = \frac{\sigma}{8}\sqrt{k_j/2n}$. It is straightforward to check that $k(M_{\cdot,j}) \le k_j$, $V(M_{\cdot,j}) \le V_j$ and $M_{\cdot,j}$ is increasing, so M is in the parameter space. Moreover, for distinct $\omega, \omega' \in \Omega$,

$$\|M^\omega - M^{\omega'}\|_F^2 \ge c\sum_{j=1}^m \frac{n}{k_j}\gamma_j^2\delta(\omega^j, (\omega')^j) \ge c\sigma^2\sum_{j=1}^m \delta(\omega^j, (\omega')^j) = c\sigma^2 K_0.$$

On the other hand,

$$\|M^\omega - M^{\omega'}\|_F^2 \le 2\sum_{j=1}^m \frac{n}{k_j}\gamma_j^2\delta(\omega^j, (\omega')^j) \le \frac{\sigma^2}{64}\delta(\omega, \omega') \le \frac{\sigma^2 K_0}{64} \le \frac{\sigma^2}{8}\log|\Omega|.$$

Applying Lemma 56 completes the proof. □

For the second term in Eq. (6.23), we first note that the bound is trivial for $l = 1$ since $\log l = 0$. The next lemma deals with the case $l = 2$.

Lemma 58. *There exist constants* $c, c' > 0$ *such that for any* $K_0 \ge m+1$ *and* $V_0 \ge 0$,

$$\inf_{\hat{M}} \sup_{M \in \mathcal{M}_{K_0}^{\mathcal{S}}(V_0)} \mathbb{P}_M\Big[\|\hat{M} - M\|_F^2 \ge cn\min\big(\sigma^2, m^3V_0^2\big)\Big] \ge c'\,,$$

where $\mathbb{P}_M$ *is the probability with respect to* $Y = M + Z$.

Proof. By Lemma 55, there exists $\Omega \subset \{0,1\}^n$ such that $\log|\Omega| \ge n/8$ and $\delta(\omega, \omega') \ge n/4$ for distinct $\omega, \omega' \in \Omega$. For each $\omega \in \Omega$, define $M^\omega \in \mathbb{R}^{n\times m}$ by setting the first column of M^ω to be $\alpha\omega$ and all other entries to be zero, where $\alpha = \min\big(\frac{\sigma}{8}, m^{3/2}V_0\big)$. Then

1. $M^\omega \in \mathcal{M}^{\mathcal{S}}_{K_0}(V_0)$ since $K(M) = m+1 \le K_0$, $V(M) \le V_0$ and we can permutate the rows of M^ω so that its first column is increasing;
2. $\|M^\omega - M^{\omega'}\|_F^2 \ge \min(\frac{\sigma^2}{64}, m^3V_0^2)\,\delta(\omega,\omega') \ge \min(\frac{n\sigma^2}{256}, \frac{n}{4}m^3V_0^2)$ for distinct $\omega, \omega' \in \Omega$;
3. $\|M^\omega - M^{\omega'}\|_F^2 \le \frac{\sigma^2}{64}\delta(\omega,\omega') \le \frac{\sigma^2}{64}n \le \frac{\sigma^2}{8}\log|\Omega|$ for $\omega, \omega' \in \Omega$.

Applying Lemma 56 completes the proof. □

For the previous two lemmas, we have only used matrices with increasing columns. However, to achieve the second term in Eq. (6.23) for $l \ge 3$, we need matrices with unimodal columns. The following packing lemma is the key.

Lemma 59. *For $l \in [m]$, consider the set $\mathfrak{M}$ of $n \times m$ matrices of the form*

$$M = \begin{cases} 1 & \text{for exactly one } j_i \in [l] \text{ for each } i \in [n], \\ 0 & \text{otherwise.} \end{cases}$$

For $\varepsilon > 0$, define $k = \lfloor \frac{\varepsilon^2 n}{2} \rfloor$. Then there exists an $\varepsilon\sqrt{n}$-packing $\mathcal{P}$ of $\mathfrak{M}$ such that $|\mathcal{P}| \ge l^{n-k}(\frac{k}{en})^k$ if $k \ge 1$ and $|\mathcal{P}| = l^n$ if $k = 0$.

Proof. There are l choices of entries to put the one in each row of M, so $|\mathfrak{M}| = l^n$. Fix $M_0 \in \mathfrak{M}$. If $\|M - M_0\|_F \le \varepsilon\sqrt{n}$ where $M \in \mathfrak{M}$, then M differs from M_0 in at most k rows. If $k = 0$, taking $\mathcal{P} = \mathfrak{M}$ gives the result. If $k \ge 1$ then

$$\left|\mathfrak{M} \cap B^{nm}(M_0, \varepsilon\sqrt{n})\right| \le \binom{n}{k} l^k \le \left(\frac{en}{k}\right)^k l^k .$$

Moreover, let $\mathcal{P}$ be a maximal $\varepsilon\sqrt{n}$-packing of $\mathfrak{M}$. Then $\mathcal{P}$ is also an $\varepsilon\sqrt{n}$-net, so $\mathfrak{M} \subset \bigcup_{M_0 \in \mathcal{P}} B^{nm}(M_0, \varepsilon\sqrt{n})$. It follows that

$$l^n = |\mathfrak{M}| \le \sum_{M_0 \in \mathcal{P}} \left|\mathfrak{M} \cap B^{nm}(M_0, \varepsilon\sqrt{n})\right| \le |\mathcal{P}| \cdot \left(\frac{en}{k}\right)^k l^k .$$

We conclude that $|\mathcal{P}| \ge l^{n-k}(\frac{k}{en})^k$. □

For notational simplicity, we now consider $2 \le l \le \min(K_0 - m, m)$ instead of $3 \le l \le \min(K_0 - m, m) + 1$.

Lemma 60. *There exist constants $c, c' > 0$ such that for any $K_0 \ge m$, $V_0 \ge 0$ and $2 \le l \le \min(K_0 - m, m)$,*

$$\inf_{\hat{M}} \sup_{M \in \mathcal{M}_{K_0}(V_0)} \mathbb{P}_M\left[\|\hat{M} - M\|_F^2 \ge cn \min\left(\sigma^2 \log(l+1), m^3(l+1)^{-3}V_0^2\right)\right] \ge c' ,$$

where $\mathbb{P}_M$ is the probability with respect to $Y = M + Z$.

Proof. Set $\varepsilon = 1/2$ and let $\mathcal{P}$ be the $\sqrt{n}/2$-packing given by Lemma 59. If $k = \lfloor \frac{n}{8} \rfloor = 0$, then $\log|\mathcal{P}| = n \log l$. Now assume that $k \ge 1$. Since $(\frac{x}{en})^x$ is decreasing on $[1, n]$,

we have that $|\mathcal{P}| \geq l^{7n/8}(\frac{1}{8e})^{n/8}$. Hence for $l \geq 2$,

$$\log|\mathcal{P}| \geq \frac{7n}{8}\log l - \frac{n}{8}\log(8e) \geq \frac{n}{4}\log l\,. \tag{6.24}$$

Moreover, for each $M_0 \in \mathcal{P}$, consider the rescaled matrix

$$M = \min\Big(\frac{\sigma}{8}\sqrt{\frac{\log l}{2}}, \big(\frac{m}{l}\big)^{3/2}V_0\Big)M_0\,.$$

1. We can permute the rows of M_0 so that each column has consecutive ones (or all zeros), so $M \in \mathcal{M}$. Moreover,

   $$K(M) = 2l + m - l \leq \min(m, K_0 - m) + m \leq K_0$$

   and

   $$V(M) \leq \Big(\frac{1}{m}\sum_{j=1}^{l}\big((m/l)^{3/2}V_0\big)^{2/3}\Big)^{3/2} = V_0\,,$$

   so $M \in \mathcal{M}_{K_0}(V_0)$ for $M_0 \in \mathcal{P}$.
2. For $M_0, M_0' \in \mathcal{P}$, $\|M_0 - M_0'\|_F^2 \geq n/4$, so

   $$\begin{aligned}\|M - M'\|_F^2 &= \min\Big(\frac{\sigma^2\log l}{128}, (m/l)^3V_0^2\Big)\|M_0 - M_0'\|_F^2\\ &\geq \min\Big(\frac{\sigma^2}{512}n\log l, \frac{n}{4}\big(\frac{m}{l}\big)^3V_0^2\Big)\,.\end{aligned}$$

3. For $M_0, M_0' \in \mathcal{P}$, $\|M_0 - M_0'\|_F^2 \leq 2\|M_0\|_F^2 + 2\|M_0'\|_F^2 \leq 4n$, so by Eq. (6.24),

   $$\|M - M'\|_F^2 \leq \frac{\sigma^2\log l}{128}\|M_0 - M_0'\|_F^2 \leq \frac{\sigma^2}{32}n\log l \leq \frac{\sigma^2}{8}\log|\mathcal{P}|\,.$$

Since $\log l \geq \frac{1}{2}\log(l+1)$ for $l \geq 2$, applying Lemma 56 completes the proof. □

Combining Lemma 57, 58 and 60, and dividing the bound by nm, we get Eq. (6.23) because the max of two terms is lower bounded by a half of their sum. The last statement in Proposition 21 holds since Lemma 57 and 58 are proved for matrices with increasing columns.

6.C.2 Proof of Theorem 13

The proof will only use Lemma 57 and 58, so the lower bound of rate $(\frac{\sigma^2V_0}{n})^{2/3} + \frac{\sigma^2}{n} + \min(\frac{\sigma^2}{m}, m^2V_0^2)$ holds even if the matrices are required to have increasing columns.

The last term $\min(\frac{\sigma^2}{m}, m^2V_0^2)$ is achieved by Lemma 58, so we focus on the trade-off between the first two terms. Suppose that $(\frac{16n}{\sigma^2})^{1/3}V_0^{2/3} \geq 3$, in which case the first

term $(\frac{\sigma^2 V_0}{n})^{2/3}$ dominates the second term. Then $m(\frac{16n}{\sigma^2})^{1/3}V_0^{2/3} - m \geq 2m$. Setting

$$K_0 = \lfloor m\big(\frac{16n}{\sigma^2}\big)^{1/3}V_0^{2/3} - m \rfloor,$$

we see that $K_0 \geq \lfloor \frac{m}{2}(\frac{16n}{\sigma^2})^{1/3}V_0^{2/3} \rfloor$. Lemma 57 can be applied with this choice of K_0. Then the term $c\sigma^2\frac{K_0}{nm}$ is lower bounded by $c(\frac{\sigma^2 V_0}{n})^{2/3}$.

On the other hand, if $(\frac{16n}{\sigma^2})^{1/3}V_0^{2/3} \leq 3$, then the second term $\frac{\sigma^2}{n}$ dominates the first up to a constant. To deduce a lower bound of this rate, we apply Lemma 55 to get $\Omega \subset \{0,1\}^m$ such that $\log|\Omega| \geq m/8$ and $\delta(\omega,\omega') \geq m/4$ for distinct $\omega, \omega' \in \Omega$. For each $\omega \in \Omega$, define $M^\omega \in \mathbb{R}^{n\times m}$ by setting every row of M^ω equal to $\frac{\sigma}{8\sqrt{n}}\omega^\top$. Then

1. $M^\omega \in \mathcal{U}^m(V_0)$ since $V(M^\omega) = 0$;
2. $\|M^\omega - M^{\omega'}\|_F^2 = \frac{\sigma^2}{64}\delta(\omega,\omega') \geq c\sigma^2 m$;
3. $\|M^\omega - M^{\omega'}\|_F^2 = \frac{\sigma^2}{64}\delta(\omega,\omega') \leq \frac{\sigma^2}{64}m \leq \frac{\sigma^2}{8}\log|\Omega|$.

Hence Lemma 56 implies a lower bound on $\frac{1}{nm}\|\hat{M} - M\|_F^2$ of rate $\frac{\sigma^2 m}{nm} = \frac{\sigma^2}{n}$.

6.D Matrices with Increasing Columns

For the model $Y = \Pi^* A^* + Z$ where $A^* \in \mathcal{S}^m$ and $Z \sim \mathrm{subG}(\sigma^2)$, a computationally efficient estimator $(\tilde{\Pi}, \tilde{A})$ has been constructed in Section 6.4 using the RankScore procedure. We will bound its rate of estimation in this section. Recall that the definition of $(\tilde{\Pi}, \tilde{A})$ consists of two steps. First, we recover an order (or a ranking) of the rows of Y, which leads to an estimator $\tilde{\Pi}$ of the permutation. Then define $\tilde{A} \in \mathcal{S}^m$ so that $\tilde{\Pi}\tilde{A}$ is the projection of Y onto the convex cone $\tilde{\Pi}\mathcal{S}^m$. For the analysis of the algorithm, we deal with the projection step first, and then turn to learning the permutation.

6.D.1 Projection

In fact, for *any* estimator $\tilde{\Pi}$, if $\tilde{A}$ is defined as above by the projection corresponding to $\tilde{\Pi}$, then the error $\|\tilde{\Pi}\tilde{A} - \Pi^* A^*\|_F^2$ can be split into two parts: the permutation error $\|(\tilde{\Pi} - \Pi^*)A^*\|_F^2$ and the estimation error of order $\tilde{O}(\sigma^2 K(A^*))$.

The proof of the following oracle inequality is very similar to that of Theorem 10, so we will sketch the proof without providing all the details.

Lemma 61. *Consider the model $Y = \Pi^* A^* + Z$ where $A^* \in \mathcal{S}^m$ and $Z \sim \mathrm{subG}(\sigma^2)$. For any $\tilde{\Pi} \in \mathfrak{S}_n$, define $\tilde{A} \in \mathcal{S}^m$ so that $\tilde{\Pi}\tilde{A}$ is the projection of Y onto $\tilde{\Pi}\mathcal{S}^m$. Then with probability at least $1 - e^{-c(n+m)}$,*

$$\|\tilde{\Pi}\tilde{A} - \Pi^* A^*\|_F^2 \lesssim \min_{A\in\mathcal{S}^m}\Big(\|A - A^*\|_F^2 + \sigma^2 K(A)\log\frac{enm}{K(A)}\Big) + \|(\tilde{\Pi} - \Pi^*)A^*\|_F^2.$$

Proof. Assume without loss of generality that $\Pi^* = I_n$. Let $A \in \mathcal{S}^m$ and define

$$f_{\tilde{\Pi}A}(t) = \sup_{M \in \tilde{\Pi}\mathcal{S}^m \cap \mathcal{B}^{nm}(\tilde{\Pi}A, t)} \langle M - \tilde{\Pi}A, Y - \tilde{\Pi}A \rangle - \frac{t^2}{2} \,.$$

Since $\mathcal{S}^m = \mathcal{C}^m_{\mathbf{l}}$ with $\mathbf{l} = (n, \dots, n)$, by Lemma 53,

$$\log N\big(\Theta_{\tilde{\Pi}\mathcal{S}^m}(A,t), \|\cdot\|_F, \varepsilon\big) \le C\varepsilon^{-1} t\, K(A) \log \frac{enm}{K(A)} \,.$$

Using the proof of Lemma 47, we see that

$$f_{\tilde{\Pi}A}(t) \le C\sigma t \sqrt{K(A) \log \frac{enm}{K(A)}} + t\|\tilde{\Pi}A - A^*\|_F - \frac{t^2}{2} + st$$

with probability at least $1 - C\exp(-\frac{cs^2}{\sigma^2})$. Then the proof of Theorem 10 implies that with probability at least $1 - e^{-c(n+m)}$,

$$\begin{aligned} \|\tilde{\Pi}\tilde{A} - A^*\|_F^2 &\lesssim \sigma^2 K(A) \log \frac{enm}{K(A)} + \|\tilde{\Pi}A - A^*\|_F^2 \\ &\lesssim \sigma^2 K(A) \log \frac{enm}{K(A)} + \|A - A^*\|_F^2 + \|\tilde{\Pi}A^* - A^*\|_F^2 \,. \end{aligned}$$

Minimizing over $A \in \mathcal{S}^m$ yields the desired result. □

The idea of splitting the error into two terms as in Lemma 61 has appeared in previous works by Shah et al. [2017], Chatterjee and Mukherjee [2016].

6.D.2 Permutation

By virtue of Lemma 61, it remains to control the permutation error $\|\tilde{\Pi}A^* - \Pi^* A^*\|_F^2$ where $\tilde{\Pi}$ is given by the RankScore procedure defined in Section 6.4. Recall that for $i, i' \in [n]$,

$$\Delta_{A^*}(i,i') = \max_{j \in [m]} (A^*_{i',j} - A^*_{i,j}) \vee \frac{1}{\sqrt{m}} \sum_{j=1}^m (A^*_{i',j} - A^*_{i,j})$$

and $\Delta_Y(i,i')$ is defined analogously. Since columns of A^* are increasing,

$$\big|\Delta_{A^*}(i,i')\big| = \|A^*_{i',\cdot} - A^*_{i,\cdot}\|_\infty \vee \frac{1}{\sqrt{m}} \|A^*_{i',\cdot} - A^*_{i,\cdot}\|_1 \,. \tag{6.25}$$

Recall that the RankScore procedure is defined as follows. First, for $i \in [n]$, we associate with the i-th row of Y a score s_i defined by

$$s_i = \sum_{l=1}^n \mathbb{1}(\Delta_Y(l,i) \ge 2\tau) \tag{6.26}$$

for the threshold $\tau := 3\sigma\sqrt{\log(nm\delta^{-1})}$ where δ is the probability of failure. Then we order the rows of Y so that the scores are increasing with ties broken arbitrarily.

This is equivalent to requiring that the corresponding permutation $\tilde{\pi} : [n] \to [n]$ satisfies that if $s_i < s_{i'}$ then $\tilde{\pi}^{-1}(i) < \tilde{\pi}^{-1}(i')$. Define $\tilde{\Pi}$ to be the $n \times n$ permutation matrix corresponding to $\tilde{\pi}$ so that $\tilde{\Pi}_{\tilde{\pi}(i),i} = 1$ for $i \in [n]$ and all other entries of $\tilde{\Pi}$ are zero. Moreover, let $\pi^* : [n] \to [n]$ be the permutation corresponding to Π^*.

To control the permutation error, we first state a lemma which asserts that if the gap between two rows of A^* is sufficiently large, then the permutation defined above will recover their relative order with high probability.

Lemma 62. *There is an event $\mathcal{E}$ of probability at least $1 - \delta$ on which the following holds. For any $i, i' \in [n]$, if $\Delta_{A^*}(i, i') \geq 4\tau$, then $\tilde{\pi}^{-1} \circ \pi^*(i) < \tilde{\pi}^{-1} \circ \pi^*(i')$.*

Proof. Since $Z \sim \mathrm{subG}(\sigma^2)$, $Z_{i,j}$ and $\frac{1}{\sqrt{m}} \sum_{j=1}^m Z_{i,j}$ are sub-Gaussian random variables with variance proxy σ^2. A standard union bound yields that

$$\max\Big(\max_{i\in[n], j\in[m]} |Z_{i,j}|, \max_{i\in[n]} \frac{1}{\sqrt{m}} \Big| \sum_{j=1}^m Z_{i,j} \Big| \Big) \leq \tau = 3\sigma\sqrt{\log(nm\delta^{-1})}$$

on an event $\mathcal{E}$ of probability at least $1 - 2(nm + n)\exp(-\frac{\tau^2}{2\sigma^2}) \geq 1 - \delta$.

In the sequel, we make statements that are valid on the event $\mathcal{E}$. Since $Y_{\pi^*(i),j} = A^*_{i,j} + Z_{i,j}$, by the triangle inequality,

$$|\Delta_Y(\pi^*(i), \pi^*(i')) - \Delta_{A^*}(i, i')| \leq 2\tau. \tag{6.27}$$

Suppose that $\Delta_{A^*}(i, i') \geq 4\tau$. We claim that $s_{\pi^*(i)} < s_{\pi^*(i')}$. If $\Delta_Y(\pi^*(l), \pi^*(i)) \geq 2\tau$, for $l \in [n]$, then $\Delta_{A^*}(l, i) \geq 0$ by Eq. (6.27). Since A^* has increasing columns, $\Delta_{A^*}(l, i') \geq 4\tau$. Again by Eq. (6.27), $\Delta_Y(\pi^*(l), \pi^*(i')) \geq 2\tau$. By definition in Eq. (6.26), we see that $s_{\pi^*(i)} \leq s_{\pi^*(i')}$. Moreover, $\Delta_{A^*}(i, i') \geq 4\tau$ so $\Delta_Y(\pi^*(i), \pi^*(i')) \geq 2\tau$. Therefore $s_{\pi^*(i)} < s_{\pi^*(i')}$. According to the construction of $\tilde{\pi}$, $\tilde{\pi}^{-1} \circ \pi^*(i) < \tilde{\pi}^{-1} \circ \pi^*(i')$. □

Next, recall that for a matrix $A \in \mathcal{S}^m$, $\mathcal{J}$ denotes the set of pairs of indices $(i, j) \in [n]^2$ such that $A_{i,\cdot}$ and $A_{j,\cdot}$ are not identical. The quantity $R(A)$ is defined by

$$R(A) = \frac{1}{n} \max_{\substack{\mathcal{I}\subset[n]^2 \\ |\mathcal{I}|=n}} \sum_{(i,j)\in\mathcal{I}\cap\mathcal{J}} \Big(\frac{\|A_{i,\cdot} - A_{j,\cdot}\|_2^2}{\|A_{i,\cdot} - A_{j,\cdot}\|_\infty^2} \wedge \frac{m\|A_{i,\cdot} - A_{j,\cdot}\|_2^2}{\|A_{i,\cdot} - A_{j,\cdot}\|_1^2} \Big).$$

For any nonzero vector $u \in \mathbb{R}^m$, $\|u\|_2^2/\|u\|_\infty^2 \geq 1$ with equality achieved when $\|u\|_0 = 1$, and $\|u\|_2^2/\|u\|_1^2 \geq m^{-1}$ with equality achieved when all entries of u are the same. Hence $R(A) \geq 1$. Moreover, $\|u\|_2^2 \leq \|u\|_1\|u\|_\infty$ by Hölder's inequality, so $\frac{\|u\|_2^2}{\|u\|_\infty^2} \wedge \frac{m\|u\|_2^2}{\|u\|_1^2} \leq \sqrt{m}$ as the product of the two terms is no larger than m. The equality is achieved by $u = (1, \ldots, 1, 0, \ldots, 0)$ where the first $\sqrt{m}$ entries are equal to one. Therefore,

$$R(A) \in \big[1, \sqrt{m}\,\big].$$

Intuitively, the quantity $R(A)$ is small if the difference of any two rows of A is either very sparse or very dense.

Lemma 63. *There is an event $\mathcal{E}$ of probability at least $1-\delta$ on which*

$$\|\tilde{\Pi}A^* - \Pi^* A^*\|_F^2 \lesssim \sigma^2 R(A^*)\, n \log(nm\delta^{-1})\,.$$

Proof. Throughout the proof, we restrict ourselves to the event $\mathcal{E}$ defined in Lemma 62. To simplify the notation, we define $\alpha_i = A^*_{\tilde{\pi}^{-1}\circ\pi^*(i),\cdot} - A^*_{i,\cdot}$. Then

$$\|\tilde{\Pi}A^* - \Pi^* A^*\|_F^2 = \sum_{i=1}^n \|A^*_{\tilde{\pi}(i),\cdot} - A^*_{\pi^*(i),\cdot}\|_2^2 = \sum_{i\in I} \|\alpha_i\|_2^2\,, \tag{6.28}$$

where I is the set of indices i for which α_i is nonzero. For each $i \in I$,

$$\begin{aligned}\|\alpha_i\|_2^2 &= \min\Big(\frac{\|\alpha_i\|_2^2}{\|\alpha_i\|_\infty^2}, \frac{m\|\alpha_i\|_2^2}{\|\alpha_i\|_1^2}\Big)\cdot \max\Big(\|\alpha_i\|_\infty^2, \frac{\|\alpha_i\|_1^2}{m}\Big)\\ &= \min\Big(\frac{\|\alpha_i\|_2^2}{\|\alpha_i\|_\infty^2}, \frac{m\|\alpha_i\|_2^2}{\|\alpha_i\|_1^2}\Big)\cdot \Delta_{A^*}\big(i, \tilde{\pi}^{-1}\circ\pi^*(i)\big)^2 \end{aligned}\tag{6.29}$$

by Eq. (6.25).

Next, we proceed to showing that $|\Delta_{A^*}(i,\nu(i))| \le 4\tau$ for any $i \in [n]$, where $\nu = \tilde{\pi}^{-1}\circ\pi^*$. To that end, note that if $\Delta_{A^*}(i,\nu(i)) > 4\tau$, in which case $\Delta_{A^*}(i,i') > 4\tau$ for all $i' \in I' := \{i' \in [n] \,:\, i' \ge \nu(i)\}$, then it follows from Lemma 62 that on $\mathcal{E}$, $\nu(i) < \nu(i')$, $\forall\, i \in I'$. Note that $|\nu(I')| = |I'| = n - \nu(i) + 1$. Hence $\nu(i) < \nu(i')$, $\forall\, i \in I'$ implies that $\nu(i) \le n - |\nu(I')| = \nu(i) - 1$, which is a contradiction. Therefore, there does not exist such $i \in [n]$ on $\mathcal{E}$. The case where $\Delta_{A^*}(i,\nu(i)) < -4\tau$ is treated in a symmetric manner.

Combining this bound with Eq. (6.28) and Eq. (6.29), we conclude that

$$\begin{aligned}\|\tilde{\Pi}A^* - \Pi^* A^*\|_F^2 &\lesssim \sum_{i\in I} \min\Big(\frac{\|\alpha_i\|_2^2}{\|\alpha_i\|_\infty^2}, \frac{m\|\alpha_i\|_2^2}{\|\alpha_i\|_1^2}\Big)\cdot \tau^2\\ &\lesssim \sigma^2 R(A^*)\, n\log(nm\delta^{-1})\,.\end{aligned}$$

by the definitions of $R(A^*)$ and τ. □

6.D.3 Proof of Theorem 14

The bound is an immediate consequence of Lemma 61 and Lemma 63 with $\delta = (nm)^{-C}$ for $C > 0$.

6.E Upper bounds in a Trivial Case

In Theorem 13, we have observed the term $\frac{\sigma^2}{m} \wedge m^2 V(A)^2$, whereas the LS estimator only has $\frac{\sigma^2}{m}\log n$ in the upper bounds. The next proposition shows that in the case

$m^2V(A)^2 \le \frac{\sigma^2}{m}$, we can simply use an averaging estimator that achieves the term $m^2V(A)^2$.

Proposition 22. *For $Y = \Pi^* A^* + Z$ where $Z \sim \mathrm{subG}(\sigma^2)$, let $\hat{\Pi} = I_n$ and $\hat{A}$ be defined by $\hat{A}_{i,j} = \frac{1}{n}\sum_{k=1}^n Y_{k,j}$ for all $(i,j) \in [n] \times [m]$. Then,*

$$\frac{1}{nm}\|\hat{\Pi}\hat{A} - \Pi^* A^*\|_F^2 \lesssim \frac{\sigma^2}{n} + m^2 V(A)^2$$

with probability at least $1 - \exp(-m)$ and

$$\frac{1}{nm}\mathbb{E}\|\hat{\Pi}\hat{A} - \Pi^* A^*\|_F^2 \lesssim \frac{\sigma^2}{n} + m^2 V(A)^2 \,.$$

Proof. Recall that $V(A) = (\frac{1}{m}\sum_{j=1}^m V_j(A)^{2/3})^{3/2}$. Since the ℓ_2-norm of a vector is no larger than the $\ell_{\frac{2}{3}}$-norm,

$$\sum_{j=1}^m V_j(A)^2 \le \Big(\sum_{j=1}^m V_j(A)^{2/3}\Big)^3 = m^3 V(A)^2.$$

On the other hand,

$$\hat{A}_{i,j} = \frac{1}{n}\sum_{k=1}^n A^*_{k,j} + \frac{1}{n}\sum_{k=1}^n Z_{k,j}\,,$$

so we have that

$$\begin{aligned}
&\|\hat{\Pi}\hat{A} - \Pi^* A^*\|_F^2 \\
&= \sum_{i\in[n], j\in[m]} \Big(\frac{1}{n}\sum_{k=1}^n A^*_{k,j} + \frac{1}{n}\sum_{k=1}^n Z_{k,j} - A^*_{i,j}\Big)^2 \\
&\le 2\sum_{i\in[n], j\in[m]} \Big(\frac{1}{n}\sum_{k=1}^n A^*_{k,j} - A^*_{i,j}\Big)^2 + \frac{2}{n^2}\sum_{i\in[n], j\in[m]} \Big(\sum_{k=1}^n Z_{k,j}\Big)^2 \\
&\le 2n\sum_{j\in[m]} V_j(A)^2 + \frac{2}{n}\sum_{j\in[m]} \Big(\sum_{k=1}^n Z_{k,j}\Big)^2 \\
&\le 2nm^3 V(A)^2 + 2\sum_{j\in[m]} g_j^2\,,
\end{aligned}$$

where $g_j = \frac{1}{\sqrt{n}}\sum_{k=1}^n Z_{k,j}$ for $j \in [m]$ so that $g_1, \dots, g_m$ are centered sub-Gaussian variables with variance proxy σ^2. It is well-known that $\mathbb{E}g_j^2 \lesssim \sigma^2$, so

$$\mathbb{E}\|\hat{\Pi}\hat{A} - \Pi^* A^*\|_F^2 \lesssim nm^3 V(A)^2 + m\sigma^2.$$

Moreover, since $(g_1, \dots, g_m)$ is a sub-Gaussian vector with variance proxy σ^2, it follows from Hsu et al. [2012, Theorem 2.1] that $\sum_{j=1}^m g_j^2 \lesssim \sigma^2 m$ with probability at least

$1 - \exp(-m)$. On this event,

$$\|\hat{\Pi}\hat{A} - \Pi^* A^*\|_F^2 \lesssim nm^3 V(A)^2 + m\sigma^2.$$

Dividing the previous two displays by nm completes the proof. □

6.F Unimodal Regression

If the permutation in the main model in Eq. (6.1) is known, then the estimation problem simply becomes a concatenation of m unimodal regressions. In fact, our proofs imply new oracle inequalities for unimodal regression. Recall that $\mathcal{U}$ denotes the cone of unimodal vectors in $\mathbb{R}^n$. Suppose that we observe

$$y = \theta^* + z\,,$$

where $\theta^* \in \mathbb{R}^n$ and z is a sub-Gaussian vector with variance proxy σ^2. Define the LS estimator $\hat{\theta}$ by

$$\hat{\theta} \in \operatorname*{argmin}_{\theta \in \mathcal{U}} \|\theta - y\|_2^2\,.$$

Moreover let $k(\theta) = \mathrm{Card}(\{\theta_1, \dots, \theta_n\})$ and $V(\theta) = \max_{i\in[n]} \theta_i - \min_{i\in[n]} \theta_i$.

Corollary 10. *There exists a constant $c > 0$ such that with probability at least $1 - n^{-\alpha}$, $\alpha \geq 1$,*

$$\frac{1}{n}\|\hat{\theta} - \theta^*\|_2^2 \lesssim \min_{\theta \in \mathcal{U}} \Big(\frac{1}{n}\|\theta - \theta^*\|_2^2 + \sigma^2 \frac{k(\theta)}{n} \log \frac{en}{k(\theta)} \Big) + \alpha\sigma^2 \frac{\log n}{n} \tag{6.30}$$

and

$$\frac{1}{n}\|\hat{\theta} - \theta^*\|_2^2 \lesssim \min_{\theta \in \mathcal{U}} \Big[\frac{1}{n}\|\theta - \theta^*\|_2^2 + \Big(\frac{\sigma^2 V(\theta) \log n}{n} \Big)^{2/3} \Big] + \alpha\sigma^2 \frac{\log n}{n}\,.$$

The corresponding bounds in expectation also hold.

Proof. The proof closely follows that of Theorem 10 and Theorem 11.

First note that the term $n \log n$ in the bound of Lemma 54 comes from a union bound applied to the set of permutations, so it is not present if we consider only the set of unimodal matrices $\mathcal{U}^m$ instead of $\mathcal{M}$. Hence taking $m = 1$ in the lemma yields that

$$\log N\big(\Theta_{\mathcal{U}}(\tilde{\theta}, t), \|\cdot\|_2, \varepsilon\big) \leq C\varepsilon^{-1} t\, k(\tilde{\theta}) \log \frac{en}{k(\tilde{\theta})}\,.$$

For $\tilde{\theta} \in \mathcal{U}$, define

$$f_{\tilde{\theta}}(t) = \sup_{\theta \in \mathcal{U} \cap \mathcal{B}^n(\tilde{\theta}, t)} \langle \theta - \tilde{\theta}, y - \tilde{\theta} \rangle - \frac{t^2}{2}\,.$$

Following the proof of Lemma 47 and using the above metric entropy bound, we see

that

$$f_{\tilde{\theta}}(t) \le C\sigma t\sqrt{k(\tilde{\theta})\log\frac{en}{k(\tilde{\theta})}} + t\|\tilde{\theta}-\theta^*\|_2 - \frac{t^2}{2} + st$$

with probability at least $1 - C\exp(-\frac{cs^2}{\sigma^2})$. Then the proof of Theorem 10 gives that with probability at least $1 - C\exp(-\frac{cs^2}{\sigma^2})$,

$$\|\hat{\theta}-\theta^*\|_2 \le C\Big(\sigma\sqrt{k(\tilde{\theta})\log\frac{en}{k(\tilde{\theta})}} + \|\tilde{\theta}-\theta^*\|_2\Big) + 2s\,.$$

Taking $s = C\sigma\sqrt{\alpha\log n}$ for $\alpha \ge 1$ and C sufficiently large, we get that with probability at least $1 - n^{-\alpha}$,

$$\|\hat{\theta}-\theta^*\|_2^2 \lesssim \sigma^2 k(\tilde{\theta})\log\frac{en}{k(\tilde{\theta})} + \|\tilde{\theta}-\theta^*\|_2^2 + \alpha\sigma^2\log n\,.$$

Minimizing over $\tilde{\theta} \in \mathcal{U}$ yields Eq. (6.30). The corresponding bound in expectation follows from integrating the tail probability as in the proof of Theorem 10.

Finally, we can apply the proof of Theorem 11 with $m = 1$ to achieve the global bound. □

Note that the bounds in Corollary 10 match the minimax lower bounds for isotonic regression of Bellec and Tsybakov [2015] up to logarithmic factors. Since every monotonic vector is unimodal, lower bounds for isotonic regression automatically hold for unimodal regression. Therefore, we have proved that the LS estimator is minimax optimal up to logarithmic factors for unimodal regression.

A result similar to Eq. (6.30) was obtained by Bellec [2015] in its revision that was prepared independently and contemporaneously to this chapter. Chatterjee and Lafferty [2015] also improved their bounds to having optimal exponents after the first version of our current paper was posted. Interestingly Bellec [2015] employs bounds on the statistical dimension by leveraging results from Amelunxen et al. [2014], and Chatterjee and Lafferty [2015] use both the variational formula and the statistical dimension. Moreover, their results are presented in the well-specified case where $\theta^* \in \mathcal{U}$ and $\theta = \theta^*$.

Chapter 7

Conclusion and Future Work

7.1 Summary of the Thesis

In this thesis, we focused on some problems specific to optimization of quadratic functions and their applications to machine learning.

In our first contribution we provided a unified framework for minimizing non-strongly convex quadratic functions, both with noiseless and noisy gradients. It encompassed averaged gradient descent, accelerated gradient descent and the heavy ball method. They were jointly analyzed as constant parameter second-order difference equation algorithms, where stability of the system was equivalent to convergence at rate $O(1/n^2)$. This suggested a new class of algorithms that could profit at the same time from the known enhancements of averaging and acceleration: faster forgetting the initial conditions (for acceleration), and improved robustness to noise when the noise covariance was proportional to the Hessian (for averaging).

In our second main contribution, we continued this pursuit and showed that stochastic averaged accelerated gradient descent was robust to structured noise in the gradients present in least-squares regression. Consequently we proposed the first algorithm that achieved simultaneously the optimal prediction error rates for least-squares regression, both in terms of forgetting the initial conditions in $O(1/n^2)$, and in terms of dependence on the noise and dimension d of the problem, as $O(\sigma^2 d/n)$. Furthermore we also provided an analysis adapted to finer assumptions such as fast decays of the covariance matrices or optimal predictors with large norms leading to dimension-free quantities that may still be small in some distances while the "optimal" terms above were large.

In our third main contribution we extended the problem studied beforehand to composite settings where the objective function is composed of the expectation of quadratic functions and an arbitrary convex function. We showed that stochastic dual averaging algorithm with a constant step-size achieved a convergence rate of $O(1/n)$ for stochastic composite objectives, without strong convexity assumptions. Accordingly we widened the previous results on least-squares regression of Bach and Moulines [2013] to all convex regularizers and all geometries which may be represented by a Bregman divergence.

In our fourth main contribution we considered the problem of clustering high-dimensional data by finding a low-dimensional projection of the data which was well-clustered. We related this formulation to the discriminative clustering framework with the square loss for which we proposed a novel sparse extension and provided the first theoretical analysis. We also proposed a new efficient algorithm with an improved linear complexity in the number of observations and a natural extension to the multi-label scenario.

In our final contribution we considered the seriation problem which consists in permuting the rows of an observed matrix in such way that all its columns have the same shape. When the matrix was observed with additional random noise, we analyzed the minimax rates of estimation and we also designed a computationally efficient estimator in case the columns of the initial matrix were monotone increasing. A theoretical and experimental studies of this estimator were also provided.

7.2 Perspectives

Our work has triggered a few questions, which are still open.

1. Our current analysis of algorithms presented in Chapter 2 and Chapter 3 is only provided for additive noise, e.g., for least-squares regression, with knowledge of the population covariance matrix. This drawback raises two different issues: (a) common applications use the stochastic oracle with a multiplicative noise, (b) the stochastic oracle with additive noise has a computational complexity $O(d^2)$ (for multiplying the Hessian and the iterate). Thus the total running-time complexity is $O(\sigma^2 d^3/n)$ which is comparable to the one obtained by minimizing the empirical risk with the conjugate gradient. Nevertheless Jain et al. [2017] have recently provided a lower-bound which prevents from directly extending our results to least-squares regression with multiplicative noise without assuming extra assumptions. This will lead to future work.
2. Algorithms studied in Chapter 2, Chapter 3 and Chapter 4 only work with quadratic functions. In fact, these constant step-size stochastic algorithms are not converging for general smooth objectives. However each of these methods may be extended to non-quadratic functions through online Newton algorithm [Bach and Moulines, 2013] which iteratively solves quadratic approximations of the smooth problems with the algorithm to achieve the same rate $O(1/n)$.
3. Research on connections between discrete and continuous dynamics has recently gained interest in the optimization community. For the moment it enables to have a deeper intuition on accelerated gradient descent by drawing links between the proofs of convergence of the discrete and continuous dynamics. One of the strength of continuous dynamics is that their corresponding Lyapunov functions are significantly easier to design. This advantage can be leveraged to provide Lyapunov functions for discrete dynamics by discretizing the one corresponding to the continuous dynamic. Therefore further investigating these connections may yield to systematic proofs of convergence of optimization algorithms. This may be used to prove the convergence of online Newton algorithm

[Bach and Moulines, 2013] by considering its connection with continuous dynamics studied by Alvarez et al. [2002], Attouch et al. [2016b]. On the other hand an extension to stochastic approximation and corresponding stochastic differential equations would provide great insights by studying and discretizing the resulting diffusions.

4. While stochastic gradient descent is broadly used by practitioners and intensively studied by theoreticians, the choice of its step-size remains heuristic. Yet its performance depends highly on how the step-size is tuned and decreased over time. For quadratic functions, we studied the behavior of constant step-size which is easy to calibrate. However for general convex functions the stochastic gradient descent with constant step-size is not converging and the choice of its decrease remains challenging. Thus parameter-free methods are very popular [see, e.g., Kingma and Ba, 2015] and deserve further investigations.
5. Designing better stochastic algorithms for non-convex optimization is also very challenging. It could take several forms to avoid saddle-points and insure convergence to local minima: (a) by exploiting second order information (b) by adding Gaussian noise to the gradient estimate in stochastic gradient descent as in Langevin algorithms [Dalalyan, 2014, Durmus and Moulines, 2017, Durmus et al., 2016].
6. We have only provided a computationally efficient estimator for statistical seriation in the simple case of matrix with monotonic columns. We conjectured that achieving optimal rates of estimation in the seriation model is computationally hard in general. However we could also investigate approximate message passing [Donoho et al., 2009, Mézard and Montanari, 2009], and auto-encoder [Hazan and Ma, 2016] techniques which might efficiently solved this problem.

Bibliography

J. Abernethy, E. Hazan, and A. Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Proceedings of the International Conference on Learning Theory (COLT)*, 2008.

A. Agarwal and L. Bottou. A lower bound for the optimization of finite sums. In *Proceedings of the Conference on Machine Learning (ICML)*, 2015.

A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Trans. Inform. Theory*, 58(5):3235–3249, 2012.

J-B. Alayrac, P. Bojanowski, N. Agrawal, I. Laptev, J. Sivic, and S. Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

Z. Allen-Zhu and L. Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. In *Proceedings of the Innovations in Theoretical Computer Science (ITCS)*, 2017.

F. Alvarez, H. Attouch, J. Bolte, and P. Redont. A second-order gradient-like dissipative dynamical system with Hessian-driven damping. Application to optimization and mechanics. *J. Math. Pures Appl. (9)*, 81(8):747–779, 2002.

D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: phase transitions in convex programs with random data. *Information and Inference*, 2014.

N. N. Anuchina, K. I. Babenko, S. K. Godunov, N. A. Dmitriev, L. V. Dmitrieva, V. F. Dʹyachenko, A. V. Zabrodin, O. V. Lokutsievskiĭ, E. V. Malinovskaya, I. F. Podlivaev, G. P. Prokopov, I. D. Sofronov, and R. P. Fedorenko. *Teoreticheskie osnovy i konstruirovanie chislennykh algoritmov zadach matematicheskoĭ fiziki*. "Nauka", Moscow, 1979.

Y. Arjevani and O. Shamir. On the iteration complexity of oblivious first-order optimization algorithms. In *Proceedings of the Conference on Machine Learning (ICML)*, 2016.

L. Arnold. *Random Dynamical Systems*. Springer Monographs in Mathematics. Springer, 1998.

D. Arthur and S. Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2007.

J. E. Atkins, E. G. Boman, and B. Hendrickson. A spectral algorithm for seriation and the consecutive ones problem. *SIAM Journal on Computing*, 28(1):297–310, 1998.

H. Attouch and J. Peypouquet. The rate of convergence of Nesterov's accelerated forward-backward method is actually faster than $1/k^2$. *SIAM J. Optim.*, 26(3): 1824–1834, 2016.

H. Attouch, Z. Chbani, J. Peypouquet, and P. Redont. Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity. *Math. Program.*, pages 1–53, 2016a.

H. Attouch, J. Peypouquet, and P. Redont. Fast convex optimization via inertial dynamics with Hessian driven damping. *J. Differential Equations*, 261(10):5734–5783, 2016b.

J.-Y. Audibert and O. Catoni. Robust linear least squares regression. *Ann. Statist.*, 39(5):2766–2794, 2011.

M. Ayer, H. D. Brunk, G. M. Ewing, W. T. Reid, and E. Silverman. An empirical distribution function for sampling with incomplete information. *Ann. Math. Statist.*, 26(4):641–647, 12 1955.

K. S. Azoury and M. K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Mach. Learn.*, 43(3), 2001.

F. Bach. Self-concordant analysis for logistic regression. *Electron. J. Stat.*, 4:384–414, 2010.

F. Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *J. Mach. Learn. Res.*, 15(1):595–627, 2014.

F. Bach. Duality between subgradient and conditional gradient methods. *SIAM J. Optim.*, 25(1):115–129, 2015.

F. Bach and Z. Harchaoui. DIFFRAC : a discriminative and flexible framework for clustering. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

F. Bach and E. Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Found. Trends Mach. Learn.*, 4(1):1–106, 2012.

R. E. Barlow, D. J. Bartholomew, J. M. Bremner, and H. D. Brunk. *Statistical Inference under Order Restrictions. The Theory and Application of Isotonic Regression.* John Wiley & Sons, 1972.

H. H. Bauschke and J. M. Borwein. Legendre functions and the method of random Bregman projections. *J. Convex Anal.*, 4(1):27–67, 1997.

H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces.* CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer, 2011.

H. H. Bauschke, J. Bolte, and M. Teboulle. A descent Lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 2016.

A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Lett.*, 31(3):167–175, 2003.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

P. C. Bellec. Sharp oracle inequalities for least squares estimators in shape restricted regression. *arXiv preprint arXiv:1510.08029*, 2015.

P. C. Bellec. Private communication, July 2016.

P. C. Bellec and A. B. Tsybakov. Sharp oracle bounds for monotone and convex regression through aggregation. *J. Mach. Learn. Res.*, 16:1879–1892, 2015.

R. Bellman. A note on cluster analysis and dynamic programming. *Mathematical Biosciences*, 1973.

A. Ben-Tal and A. Nemirovski. *Lectures on Modern Convex Optimization.* MPS Series on Optimization. Society for Industrial and Applied Mathematics (SIAM), 2001.

Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, 2013.

A. Benveniste, P. Priouret, and M. Métivier. *Adaptive Algorithms and Stochastic Approximations.* Springer, 1990.

Q. Berthet and P. Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In *Proceedings of the International Conference on Learning Theory (COLT)*, 2013.

D. P. Bertsekas. *Nonlinear Programming.* Athena scientific, 1999.

P. J. Bickel and J. Fan. Some problems on the estimation of unimodal densities. *Statist. Sinica*, 6(1), 1996.

L. Birge. Estimation of unimodal densities without smoothness assumptions. *Ann. Statist.*, 25(3):pp. 970–981, 1997.

M. S. Birman and M. Z. Solomjak. Piecewise polynomial approximations of functions of classes $W_p{}^{\alpha}$. *Mat. Sb. (N.S.)*, 73 (115):331–355, 1967.

G. Blanchard, M. Kawanabe, M. Sugiyama, V. Spokoiny, and K.-R. Müller. In search of non-Gaussian components of a high-dimensional distribution. *J. Mach. Learn. Res.*, 7:247–282, 2006.

P. Bojanowski, F. Bach, I. Laptev, J. Ponce, C. Schmid, and J. Sivic. Finding actors and actions in movies. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2013.

J. Bolte and M. Teboulle. Smooth optimization with approximate gradient. *SIAM J. Optim.*, 43(3):1266–1292, 2003.

V. S. Borkar. Stochastic approximation with two time scales. *Systems & Control Letters*, 29(5):291–294, 1997.

V. S. Borkar. *Stochastic Approximation: a Dynamical Systems Viewpoint*. Cambridge University Press, 2008.

J. M. Borwein and A. S. Lewis. *Convex Analysis and Nonlinear Optimization*, volume 3 of *CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC*. Springer, 2000. Theory and examples.

L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

L. Bottou and Y. Le Cun. On-line learning for very large data sets. *Appl. Stoch. Models Bus. Ind.*, 21(2):137–151, 2005.

S. Boucheron and P. Massart. A high-dimensional Wilks phenomenon. *Probab. Theory Related Fields*, 150(3-4):405–433, 2011.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities*. Oxford University Press, 2013. A Nonasymptotic Theory of Independence.

N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre. Manopt, a Matlab toolbox for optimization on manifolds. *J. Mach. Learn. Res.*, 2014.

J. Bourgain, V. H. Vu, and P. M. Wood. On the singularity probability of discrete random matrices. *Journal of Functional Analysis*, 258(2):559–603, 2010.

V. Boyarshinov and M. Magdon-Ismail. Linear time isotonic and unimodal regression in the L_1 and L_∞ norms. *J. Discrete Algorithms*, 4(4), 2006.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan. *Linear Matrix Inequalities in System and Control Theory*, volume 15 of *SIAM Studies in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), 1994.

L. M. Bregman. A relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming. *Ž. Vyčisl. Mat. i Mat. Fiz.*, 7:620–631, 1967.

R. Bro and N. Sidiropoulos. Least squares algorithms under unimodality and non-negativity constraints. *J. Chemometrics*, 12:223–247, 1998.

S. Bubeck. Convex optimization: Algorithms and complexity. *Found. Trends Mach. Learn.*, 8(3-4), 2015.

S. Bubeck, Y.-T. Lee, and M. Singh. A geometric alternative to nesterov's accelerated gradient descent. *arXiv:1506.08187*, 2015.

G. Casella and R. L. Berger. *Statistical Inference*. Statistics/Probability Series. Wadsworth & Brooks/Cole, 1990.

A. Cauchy. Méthode générale pour la résolution des systemes d'équations simultanées. *Comp. Rend. Sci. Paris*, 25(1847):536–538, 1847.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *IEEE Trans. Inform. Theory*, 50(9):2050–2057, 2004.

V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *Found. Comput. Math.*, 12(6):805–849, 2012.

S. Chatterjee. A new perspective on least squares under convex constraint. *Ann. Statist.*, 42(6):2340–2381, 12 2014.

S. Chatterjee. Matrix estimation by universal singular value thresholding. *Ann. Statist.*, 43(1):177–214, 2015.

S. Chatterjee and J. Lafferty. Adaptive risk bounds in unimodal regression. *arXiv preprint arXiv:1512.02956*, 2015.

S. Chatterjee and S. Mukherjee. On estimation in tournaments and graphs under monotonicity constraints. *arXiv preprint arXiv:1603.04556*, 2016.

S. Chatterjee, A. Guntuboyina, and B. Sen. On risk bounds in isotonic and other shape restricted regression problems. *Ann. Statist.*, 43(4):1774–1800, 2015.

S. Chatterjee, A. Guntuboyina, and B. Sen. On matrix estimation under monotonicity constraints. *Bernoulli (to appear)*, 2017.

G. Chen and M. Teboule. Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM J. Optim.*, 3(3):538–543, 1993.

K. L. Chung. On a stochastic approximation method. *Ann. Math. Statistics*, 25: 463–483, 1954.

F. Clarke. *Functional Analysis, Calculus of Variations and Optimal Control*, volume 264 of *Graduate Texts in Mathematics*. Springer, 2013.

I. Colin, A. Bellet, J. Salmon, and S. Clémençon. Gossip dual averaging for decentralized optimization of pairwise functions. In *Proceedings of the Conference on Machine Learning (ICML)*, 2016.

O. Collier and A. S. Dalalyan. Minimax rates in permutation estimation for feature matching. *J. Mach. Learn. Res.*, 17(6):1–32, 2016.

P. L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, volume 49 of *Springer Optim. Appl.*, pages 185–212. Springer, 2011.

A. Cotter, O. Shamir, N. Srebro, and K. Sridharan. Better mini-batch algorithms via accelerated gradient methods. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Trans. Inform. Theory*, 13(1):21–27, 1967.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, second edition, 2006.

J. B. Crockett and H. Chernoff. Gradient methods of maximization. *Pacific J. Math.*, 5:33–50, 1955.

H. B. Curry. The method of steepest descent for non-linear minimization problems. *Quart. Appl. Math.*, 2:258–261, 1944.

D. Dai, P. Rigollet, L. Xia, and T. Zhang. Aggregation of affine estimators. *Electron. J. Statist.*, 8(1):302–327, 2014.

A. S. Dalalyan. Theoretical guarantees for approximate sampling from a smooth and log-concave density. to appear in jrss b, arXiv preprint arXiv:1412.7392, 2014.

C. Daskalakis, I. Diakonikolas, and R. A. Servedio. Learning k-modal distributions via testing. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2012.

C. Daskalakis, I. Diakonikolas, R. A. Servedio, G. Valiant, and P. Valiant. Testing k-modal distributions: Optimal algorithms via reductions. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2013.

A. d'Aspremont. Smooth optimization with approximate gradient. *SIAM J. Optim.*, 19(3):1171–1183, 2008.

D. Davidson and J. Marschak. Experimental tests of a stochastic decision theory. *Measurement: Definitions and theories*, 1959.

T. De Bie and N. Cristianini. Convex methods for transduction. In *Advances in Neural Information Processing Systems (NIPS)*, 2003.

F. De la Torre and T. Kanade. Discriminative cluster analysis. In *Proceedings of the Conference on Machine Learning (ICML)*, 2006.

A. Defazio, F. Bach, and S. Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

A. Défossez and F. Bach. Averaged least-mean-squares: bias-variance trade-offs and optimal sampling distributions. In *Proceedings of the International Conference on Artificial Intelligence and Statistics, (AISTATS)*, 2015.

O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao. Optimal distributed online prediction using mini-batches. *J. Mach. Learn. Res.*, 13:165–202, 2012.

O. Devolder, F. Glineur, and Y. Nesterov. First-order methods with inexact oracle: the strongly convex case. *CORE Discussion Papers*, 2013016, 2013.

O. Devolder, F. Glineur, and Y. Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Math. Program.*, 146(1-2, Ser. A):37–75, 2014.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31 of *Applications of Mathematics*. Springer, 1996.

E. Diederichs, A. Juditsky, A. Nemirovski, and V. Spokoiny. Sparse non-Gaussian component analysis by semidefinite programming. *Machine learning*, 91(2):211–238, 2013.

A. Dieuleveut and F. Bach. Non-parametric stochastic approximation with large step sizes. *Ann. Statist.*, 44(4):1363–1399, 2015.

A. Dieuleveut, N. Flammarion, and F. Bach. Harder, Better, Faster, Stronger Convergence Rates for Least-Squares Regression. *arXiv preprint arXiv:1602.05419v2*, 2016.

C. Ding and T. Li. Adaptive dimension reduction using discriminant analysis and K-means clustering. In *Proceedings of the Conference on Machine Learning (ICML)*, 2007.

D. L. Donoho. Gelfand n-widths and the method of least squares. Statistics Technical Report 282, University of California, Berkeley, December 1990.

D. L. Donoho, A. Maleki, and A. Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

R. D. Driver. Note on a paper of Halanay on stability for finite difference equations. *Arch. Rational Mech. Anal.*, 18:241–243, 1965.

J. Duchi and F. Ruan. Local asymptotics for some stochastic optimization problems: optimality, constraint identification, and dual averaging. *arXiv preprint arXiv:1612.05612*, 2016.

J. Duchi and Y. Singer. Efficient online and batch learning using forward backward splitting. *J. Mach. Learn. Res.*, 10:2899–2934, 2009.

J. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *Proceedings of the International Conference on Learning Theory (COLT)*, 2010.

J. Duchi, A. Agarwal, and M. Wainwright. Dual averaging for distributed optimization: convergence analysis and network scaling. *IEEE Trans. Automat. Control*, 57 (3):592–606, 2012.

M. Duflo. *Random Iterative Models*. Springer, 1997.

A. Durmus and E. Moulines. Non-asymptotic convergence analysis for the unadjusted langevin algorithm. *Ann. Appl. Prob.*, 2017.

A. Durmus, U. Simsekli, E. Moulines, R. Badeau, and G. Richard. Stochastic gradient Richardson-Romberg markov chain monte carlo. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

P. P. B. Eggermont and V. N. LaRiccia. Maximum likelihood estimation of smooth monotone and unimodal densities. *Ann. Statist.*, 28(3), 2000.

Y. M. Ermoliev. Methods of solution of nonlinear extremal problems. *Cybernetics*, 2 (4):1–14, 1966.

Y. M. Ermoliev. The method of generalized stochastic gradients and stochastic quasi-Fejér sequences. *Kibernetika*, (2):73–83, 1969.

V. Fabian. On asymptotic normality in stochastic approximation. *Ann. Math. Statist*, 39:1327–1332, 1968.

P. C. Fishburn. Binary choice probabilities: on the varieties of stochastic transitivity. *Journal of Mathematical psychology*, 10(4), 1973.

F. Fogel, R. Jenatton, F. Bach, and A. d'Aspremont. Convex relaxations for permutation problems. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

D. Freedman. *Statistical Models: Theory and Practice*. Cambridge University Press, 2009.

J. H. Friedman and W. Stuetzle. Projection pursuit regression. *Journal of the American Statistical Association*, 76(376):817–823, 1981.

A. Frieze and M. Jerrum. Improved approximation algorithms for MAX k-CUT and MAX BISECTION. In *Integer Programming and Combinatorial Optimization*. Springer, 1995.

M. Frisen. Unimodal regression. *Journal of the Royal Statistical Society. Series D*, 35(4):479–485, 1986.

R. Frostig, R. Ge, S. M. Kakade, and A. Sidford. Un-regularizing: Approximate proximal point and faster stochastic algorithms for empirical risk minimization. In *Proceedings of the Conference on Machine Learning (ICML)*, 2015.

D. R. Fulkerson and O. A. Gross. Incidence matrices with the consecutive 1's property. *Bull. Amer. Math. Soc.*, 70:681–684, 1964.

C. Gao, Y. Lu, and H. H. Zhou. Rate-optimal graphon estimation. *Ann. Statist.*, 43 (6):2624–2652, 12 2015.

M. R. Garey, D. S. Johnson, and L. Stockmeyer. Some simplified NP-complete graph problems. *Theoret. Comput. Sci.*, 1(3):237–267, 1976.

R. Ge, J. D. Lee, and T. Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

Z. Geng and N. Z. Shi. Algorithm as 257: Isotonic regression for umbrella orderings. *Journal of the Royal Statistical Society. Series C*, 39(3):397–402, 1990.

C. Gentile and N. Littlestone. The robustness of the p-norm algorithms. In *Proceedings of the International Conference on Learning Theory (COLT)*, 1999.

T. L. Gertzen and M. Grötschel. Flinders Petrie, the travelling salesman problem, and the beginning of mathematical modeling in archaeology. *Doc. Math.*, (Extra volume: Optimization stories):199–210, 2012.

M. X. Goemans and D. P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. ACM*, 42 (6):1115–1145, 1995.

A. A. Goldstein. Cauchy's method of minimization. *Numer. Math.*, 4:146–150, 1962.

G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, fourth edition, 2013.

G. J. Gordon. Regret bounds for prediction problems. In *Proceedings of the International Conference on Learning Theory (COLT)*, 1999.

J. C. Gower and G. J. S. Ross. Minimum spanning trees and single Linkage cluster analysis. *J. Roy. Statist. Soc. Ser. B*, 18(1), 1969.

M. Grant and S. Boyd. Graph implementations for nonsmooth convex programs. In *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer, 2008.

M. Grant and S. Boyd. CVX: Matlab Software for Disciplined Convex Programming, version 2.1, 2014.

E. Grave. A convex relaxation for weakly supervised relation extraction. In *EMNLP*, 2014.

C. Gu. *Smoothing Spline ANOVA Models*. Springer, 2013.

L. Györfi and H. Walk. On the averaged stochastic approximation for linear regression. *SIAM J. Optim.*, 34(1):31–61, 1996.

L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2006.

M. Gürbüzbalaban, A. Ozdaglar, and P. Parrilo. Why random reshuffling beats stochastic gradient descent. *arXiv preprint arXiv:1510.08560*, 2015.

W. Hahn. Über die anwendung der methode von Ljapunov auf differenzengleichungen. *Math. Ann.*, 136:430–441, 1958.

A. Halanay. Quelques questions de la théorie de la stabilité pour les systèmes aux différences finies. *Arch. Rational Mech. Anal.*, 12:150–154, 1963.

J. Hannan. Approximation to Bayes risk in repeated play. In *Contributions to the Theory of Games, vol. 3*, Annals of Mathematics Studies, no. 39, pages 97–139. Princeton University Press, 1957.

O. Hanner. On the uniform convexity of L^p and l^p. *Ark. Mat.*, 3:239–244, 1956.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, second edition, 2009.

E. Hazan. The convex optimization approach to regret minimization. *Optimization for Machine Learning*, pages 287–303, 2012.

E. Hazan and S. Kale. Extracting certainty from uncertainty: Regret bounded by variation in costs. *Mach. Learn.*, 80(2-3), 2010.

E. Hazan and T. Ma. A non-generative framework and convex relaxations for unsupervised learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Mach. Learn.*, 69(2-3), 2007.

M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *J. Research Nat. Bur. Standards*, 49:409–436, 1952.

J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis*. Grundlehren Text Editions. Springer, 2001.

A. E. Hoerl. Application of ridge analysis to regression problems. *Chemical Engineering Progress*, 58(3):54–59, 1962.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

D. Hsu, S. M. Kakade, and T. Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab.*, 17, 2012.

D. Hsu, S. M. Kakade, and T. Zhang. Random design analysis of ridge regression. *Found. Comput. Math.*, 14(3):569–600, 2014.

C. Hu, W. Pan, and J. T. Kwok. Accelerated gradient methods for stochastic optimization and online learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

G. Huang, J. Zhang, S. Song, and Z. Chen. Maximin separation probability clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015.

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*, volume 46. John Wiley & Sons, 2004.

P. Jain, S. M. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford. Parallelizing stochastic approximation through mini-batching and tail-averaging. *arXiv preprint arXiv:1610.03774*, 2016.

P. Jain, S. M. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford. Accelerating stochastic gradient descent. *arXiv preprint arXiv:1704.08227*, 2017.

C. Jin, S.M. Kakade, and P. Netrapalli. Provable efficient online matrix completion via non-convex stochastic gradient descent. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

A. Joulin and F. Bach. A convex relaxation for weakly supervised classifiers. In *Proceedings of the Conference on Machine Learning (ICML)*, 2012.

A. Joulin, F. Bach, and J. Ponce. Discriminative clustering for image co-segmentation. In *Computer Vision and Pattern Recognition (CVPR)*, 2010a.

A. Joulin, J. Ponce, and F. Bach. Efficient optimization for discriminative latent class models. In *Advances in Neural Information Processing Systems (NIPS)*, 2010b.

M. Journée, F. Bach, P-A Absil, and R. Sepulchre. Low-rank optimization on the cone of positive semidefinite matrices. *SIAM J. Optim.*, 2010.

A. Juditsky and A. S. Nemirovski. Functional aggregation for nonparametric regression. *Ann. Statist.*, 28(3):681–712, 2000.

A. Kalai and S. Vempala. Efficient algorithms for online decision problems. *J. Comput. System Sci.*, 71(3):291–307, 2005.

A. T. Kalai, A. Moitra, and G. Valiant. Efficiently learning mixtures of two Gaussians. In *International Symposium on Theory of Computing (STOC)*, 2010.

R. E. Kalman. Lyapunov functions for the problem of Lur'e in automatic control. *Proc. Nat. Acad. Sci. U.S.A.*, 49:201–205, 1963.

R. E. Kalman and J. E. Bertram. Control system analysis and design via the "second method" of Lyapunov. II. Discrete-time systems. *Trans. ASME Ser. D. J. Basic Engrg.*, 82:394–400, 1960.

L. V. Kantorovitch. On an effective method of solving extremal problems for quadratic functionals. *Dokl. Akad. Nauk SSSR*, 48:455–460, 1945.

S. B. Karmakar. An algorithm for finding a circuit of even length in a directed graph. *Internat. J. Systems Sci.*, 15(11):1197–1201, 1984.

R. M. Karp. Reducibility among combinatorial problems. In *Complexity of Computer Computations*, pages 85–103. Plenum, 1972.

D. G. Kendall. A statistical approach to Flinders Petrie's sequence-dating. *Bull. Inst. Internat. Statist.*, 40:657–681, 1963.

D. G. Kendall. Incidence matrices, interval graphs and seriation in archeology. *Pacific J. Math.*, 28:565–570, 1969.

D. G. Kendall. A mathematical approach to seriation. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 269 (1193):pp. 125–134, 1970.

D. G. Kendall. Abundance matrices and seriation in archaeology. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 17:104–112, 1971.

L. G. Khachiyan. A polynomial algorithm in linear programming. *Dokl. Akad. Nauk SSSR*, 244(5):1093–1096, 1979.

D.P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings of the international conference on learning representations (ICLR)*, 2015.

J. Kivinen and M. K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Inform. and Comput.*, 132(1):1–63, 1997.

K. C. Kiwiel. Proximal minimization methods with generalized Bregman functions. *SIAM J. Control Optim.*, 35(4):1142–1168, 1997.

A. J. Kleywegt, A. Shapiro, and T. Homem-de Mello. The sample average approximation method for stochastic discrete optimization. *SIAM J. on Optimization*, 12 (2), 2002.

C. Köllmann, B. Bornkamp, and K. Ickstadt. Unimodal regression using Bernstein-Schoenberg splines and penalties. *Biometrics*, 70(4), 2014.

A. N. Kolmogorov and V. M. Tihomirov. ε-entropy and ε-capacity of sets in functional space. *Amer. Math. Soc. Transl. (2)*, 17, 1961.

W. Krichene, A. Bayen, and P. L. Bartlett. Accelerated mirror descent in continuous and discrete time. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

H. W. Kuhn and A. W. Tucker. Nonlinear programming. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 481–492. University of California Press, 1951.

H. Kushner and G G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2003.

S. Lacoste-Julien, M. Schmidt, and F. Bach. A simpler approach to obtaining an $O(1/t)$ convergence rate for the projected stochastic subgradient method. *arXiv preprint arXiv:1212.2002*, 2012.

R. Lajugie, P. Bojanowski, P. Cuvillier, S. Arlot, and F. Bach. A weakly-supervised discriminative model for audio-to-score alignment. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016.

G. Lan. An optimal method for stochastic composite optimization. *Math. Program.*, 133(1-2, Ser. A):365–397, 2012.

G. Lan and Y. Zhou. An optimal randomized incremental gradient method. *arXiv preprint arXiv:1507.02000*, 2015.

J. LaSalle and S. Lefschetz. *Stability by Liapunov's Direct Method, with Applications*. Mathematics in Science and Engineering, Vol. 4. Academic Press, 1961.

N. Le Roux and F. Bach. Local component analysis. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2013.

G. Lecué. Optimal oracle inequality for aggregation of classifiers under low noise condition. In *Learning Theory*, volume 4005 of *Lecture Notes in Comput. Sci.*, pages 364–378. Springer, 2006.

G. Lecué. Optimal rates of aggregation in classification under low noise assumption. *Bernoulli*, 13(4):1000–1022, 2007.

G. Lecué and S. Mendelson. Learning subgaussian classes: Upper and minimax bounds. *arXiv preprint arXiv:1305.4825*, 2013.

G. Lecué and S. Mendelson. Performance of empirical risk minimization in linear aggregation. *Bernoulli*, 22(3):1520–1534, 2016.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer, 1991.

S. Lee and S. J. Wright. Manifold identification in dual averaging for regularized stochastic online learning. *J. Mach. Learn. Res.*, 13:1705–1744, 2012.

L. Lessard, B. Recht, and A. Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM J. Optim.*, 26(1):57–95, 2016.

E. S. Levitin and B. T. Polyak. Minimization methods in the presence of constraints. *Ž. Vyčisl. Mat. i Mat. Fiz.*, 6:787–823, 1966.

Y.-F. Li, I. W. Tsang, J. T.-Y. Kwok, and Z-H. Zhou. Tighter and convex maximum margin clustering. In *Proceedings of the International Conference on Artificial Intelligence and Statistics, (AISTATS)*, 2009.

C. H. Lim and S. Wright. Beyond the birkhoff polytope: Convex relaxations for vector permutation problems. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

H. Lin, J. Mairal, and Z. Harchaoui. A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

N. Littlestone. From on-line to batch learning. In *Proceedings of the International Conference on Learning Theory (COLT)*, 1989.

H. Lu, R. Freund, and Y. Nesterov. Relatively-smooth convex optimization by first-order methods, and applications. *arXiv preprint arXiv:1610.05708*, 2016.

Z. Q. Luo, W. K. Ma, A. C. So, Y. Ye, and S. Zhang. Semidefinite relaxation of quadratic optimization problems. *Signal Processing Magazine, IEEE*, 2010.

A. I. Lur'e and V.N. Postnikov. On the theory of stability of control systems. *Applied Mathematics and Mechanics*, 8(3):246–248, 1944.

A. M. Lyapunov. The general problem of the stability of motion. *Internat. J. Control*, 55(3):521–790, 1892. Translated by A. T. Fuller from Édouard Davaux's French translation (1907) of the 1892 Russian original.

Z. Ma and Y. Wu. Computational barriers in minimax submatrix detection. *Ann. Statist.*, 43(3):1089–1116, 06 2015.

O. Macchi. *Adaptive Processing: the Least-Mean-Squares Approach with Applications in Transmission*. John Wiley & Sons, 1995.

J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967.

E. Mammen. Estimating a smooth monotone regression function. *Ann. Statist.*, 19 (2):724–740, 1991.

E. Mammen and S. van de Geer. Locally adaptive regression splines. *Ann. Statist.*, 25(1):387–413, 1997.

B. Martinet. Breve communication. Régularisation d'inéquations variationnelles par approximations successives. *ESAIM: Modélisation Mathématique et Analyse Numérique*, 4:154–158, 1970.

P. Massart. *Concentration Inequalities and Model Selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, 2007.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Monographs on Statistics and Applied Probability. Chapman & Hall, second edition, 1989.

H. B. McMahan. Follow-the-regularized-leader and mirror descent: Equivalence theorems and l1 regularization. In *Proceedings of the International Conference on Artificial Intelligence and Statistics, (AISTATS)*, 2011.

S. Mendelson. Learning without concentration. *J. ACM*, 62(3), June 2015.

S. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, 2009.

M. Mézard and A. Montanari. *Information, Physics, and Computation*. Oxford Graduate Texts. Oxford University Press, 2009.

A. Moitra and G. Valiant. Settling the polynomial learnability of mixtures of Gaussians. In *Annual Symposium on Foundations of Computer Science (FOCS)*, 2010.

J.-J. Moreau. Fonctions convexes duales et points proximaux dans un espace Hilbertien. *C. R. Acad. Sci. Paris*, 255:2897–2899, 1962.

E. A. Nadaraya. On estimating regression. *Theory of Probability and Its Application*, 9:141–142, 1964.

A. V. Nazin. Information inequalities in a problem of gradient stochastic optimization and optimal realizable algorithms. *Avtomat. i Telemekh.*, (4):127–138, 1989.

A. S. Nemirovski and D. B. Yudin. Effective methods for the solution of convex programming problems of large dimensions. *Èkonom. i Mat. Metody*, 15(1):135–152, 1979.

A. S. Nemirovski, B. T. Polyak, and A. B. Tsybakov. The rate of convergence of non-parametric estimates of maximum likelihood type. *Problemy Peredachi Informatsii*, 21(4):17–33, 1985.

A. S. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, 19(4):1574–1609, 2009.

A. S. Nemirovsky and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience Series in Discrete Mathematics. John Wiley & Sons, 1983.

Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Dokl. Akad. Nauk SSSR*, 27(2):372–376, 1983.

Y. Nesterov. *Introductory Lectures on Convex Optimization*, volume 87 of *Applied Optimization*. Kluwer Academic, 2004. A basic course.

Y. Nesterov. Smoothing technique and its applications in semidefinite optimization. *Math. Program.*, 110(2, Ser. A):245–259, 2007.

Y. Nesterov. Primal-dual subgradient methods for convex problems. *Math. Program.*, 120(1, Ser. B):221–259, 2009.

Y. Nesterov. Gradient methods for minimizing composite functions. *Math. Program.*, 140(1, Ser. B):125–161, 2013.

Y. Nesterov and A. Nemirovskii. *Interior-Point Polynomial Algorithms in Convex Programming*, volume 13 of *SIAM Studies in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), 1994.

A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems (NIPS)*. 2002.

A. Nitanda. Stochastic proximal gradient descent with acceleration techniques. In *Advances in Neural Information Processing Systems (NIPS)*. 2014.

G. Niu, B. Dai, L. Shang, and M. Sugiyama. Maximum volume clustering: a new discriminative clustering approach. *J. Mach. Learn. Res.*, 14(1):2641–2687, 2013.

J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, second edition, 2006.

B. O'Donoghue and E. Candès. Adaptive restart for accelerated gradient schemes. *Found. Comput. Math.*, pages 1–18, 2013.

R. Oldenburger. Infinite powers of matrices and characteristic roots. *Duke Math. J.*, 6:357–361, 1940.

R. I. Oliveira. The lower tail of random quadratic forms with applications to ordinary least squares. *Probab. Theory Related Fields*, 166(3-4):1175–1194, 2016.

J. M. Ortega and W. C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*, volume 30 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), 2000.

R. Ostrovsky, Y. Rabani, L. J. Schulman, and C. Swamy. The effectiveness of Lloyd-type methods for the k-means problem. In *Annual Symposium on Foundations of Computer Science (FOCS)*, 2006.

A. M. Ostrowski. *Solution of Equations and Systems of Equations*. Second edition. Pure and Applied Mathematics, Vol. 9. Academic Press, 1966.

O. Perron. Über stabilität und asymptotisches verhalten der integrale von differentialgleichungssystemen. *Math. Z.*, 29(1):129–160, 1929.

W. M. Flinders Petrie. Sequences in prehistoric remains. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 29(3/4):pp. 295–301, 1899.

B. T. Polyak. Gradient methods for minimizing functionals. *Ž. Vyčisl. Mat. i Mat. Fiz.*, 3:643–653, 1963.

B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

B. T. Polyak. A general method for solving extremal problems. *Dokl. Akad. Nauk SSSR*, 174:33–36, 1967.

B. T. Polyak. Convergence and convergence rate of iterative stochastic algorithms. I. General case. *Avtomat. i Telemeh.*, (12):83–94, 1976.

B. T. Polyak. *Introduction to Optimization*. Translations Series in Mathematics and Engineering. Optimization Software, 1987.

B. T. Polyak. A new method of stochastic approximation type. *Avtomatika i telemekhanika*, 51(7):98–107, 1990.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.*, 30(4):838–855, 1992.

V.-M. Popov. Absolute stability of nonlinear systems of automatic control. *Automat. Remote Control*, 22:857–875, 1961.

M. Raginsky and A. Rakhlin. Information-based complexity, feedback and dynamics in convex programming. *IEEE Trans. Inform. Theory*, 57(10):7036–7056, 2011.

A. Rakhlin and K. Sridharan. Online non-parametric regression. In *Proceedings of the International Conference on Learning Theory (COLT)*, 2014.

A. Rakhlin, O. Shamir, and K. Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the Conference on Machine Learning (ICML)*, 2012.

I. Rish and G. Grabarnik. *Sparse Modeling: Theory, Algorithms, and Applications*. CRC press, 2014.

H. Robbins and S. Monro. A stochastic approxiation method. *Ann. Math. Statist*, 22 (3):400–407, 1951.

H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing Methods in Statistics (Proc. Sympos., Ohio State Univ.)*, pages 233–257. Academic Press, 1971.

T. Robertson, F. T. Wright, and R. L. Dykstra. *Order Restricted Statistical Inference*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Ltd., Chichester, 1988.

R. T. Rockafellar. *Convex Analysis*. Princeton Mathematical Series, No. 28. Princeton University Press, 1970.

D. Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

J. Sacks. Asymptotic distribution of stochastic approximation procedures. *Ann. Math. Statist.*, 29:373–405, 1958.

M. Schmidt, N. Le Roux, and F. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

M. Schmidt, N. Le Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *Math. Program.*, 162(1-2, Ser. A):83–112, 2017.

P. H. Schönemann. A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 31(1), 1966.

N. B. Shah, S. Balakrishnan, and M. J. Wainwright. Feeling the bern: Adaptive estimators for bernoulli probabilities of pairwise comparisons. *arXiv preprint arXiv:1603.06881*, 2016.

N. B. Shah, S. Balakrishnan, A. Guntuboyina, and M. J. Wainwright. Stochastically transitive models for pairwise comparisons: Statistical and computational issues. *IEEE Trans. Inf. Theor.*, 63(2):934–959, 2017.

S. Shalev-Shwartz. *Online Learning: Theory, Algorithms, and Applications*. PhD thesis, 2007.

S. Shalev-Shwartz. Online learning and online convex optimization. *Found. Trends Mach. Learn.*, 4(2), 2012.

S. Shalev-Shwartz. Sdca without duality, regularization and individual convexity. In *Proceedings of the Conference on Machine Learning (ICML)*, 2016a.

S. Shalev-Shwartz. Stochastic optimization for machine learning. Slides of presentation at "Optimization Without Borders 2016", 2016b.

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

S. Shalev-Shwartz and S. M. Kakade. Mind the duality gap: Logarithmic regret algorithms for online optimization. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

S. Shalev-Shwartz and Y. Singer. Online learning meets optimization in the dual. In *Learning Theory*, volume 4005 of *Lecture Notes in Comput. Sci.*, pages 423–437. Springer, 2006.

S. Shalev-Shwartz and Y. Singer. A primal-dual perspective of online learning algorithms. *Mach. Learn.*, 69(2-3), 2007.

S. Shalev-Shwartz and N. Srebro. Svm optimization: Inverse dependence on training set size. In *Proceedings of the Conference on Machine Learning (ICML)*, 2008.

S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *J. Mach. Learn. Res.*, 14:567–599, 2013.

S. Shalev-Shwartz and T. Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. In *Proceedings of the Conference on Machine Learning (ICML)*, 2014.

S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Stochastic convex optimization. In *Proceedings of the International Conference on Learning Theory (COLT)*, 2009.

O. Shamir. The sample complexity of learning linear predictors with the squared loss. *J. Mach. Learn. Res.*, 16(1):3475–3486, 2015.

O. Shamir. Without-replacement sampling for stochastic gradient methods. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

A. Shapiro, D. Dentcheva, and A. Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory, Second Edition*. Society for Industrial and Applied Mathematics (SIAM), 2014.

N. Z. Shor. An application of the method of gradient descent to the solution of the network transportation problem. *Notes, Scientific seminar on theory and application of cybernetics and operations research, Academy of Sciences U.S.S.R.*, pages 9–17, 1962.

N. Z. Shor, K. C. Kiwiel, and A. Ruszczyński. *Minimization Methods for Non-Differentiable Functions*. Springer, 1985.

J.M. Shoung and C.H. Zhang. Least squares estimators of the mode of a unimodal regression function. *Ann. Statist.*, 29(3), 2001.

K. Sridharan, S. Shalev-Shwartz, and N. Srebro. Fast rates for regularized objectives. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

P. Stein. Some general theorems on iterants. *J. Research Nat. Bur. Standards*, 48: 82–83, 1952.

Q. F. Stout. Unimodal regression via prefix isotonic regression. *Comput. Statist. Data Anal.*, 53(2):289–297, 2008.

W. Su, S. Boyd, and E. Candès. A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

T. Suzuki. Dual averaging and proximal gradient descent for online alternating direction multiplier method. In *Proceedings of the Conference on Machine Learning (ICML)*, 2013.

E. Takimoto and M. K. Warmuth. The minimax strategy for gaussian density estimation. pp. In *Proceedings of the International Conference on Learning Theory (COLT)*, 2000.

G. Temple. The general theory of relaxation methods applied to linear systems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 169(939):476–500, 1939.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

J. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 2012.

P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. *unpublished*, 2008.

A. B. Tsybakov. Optimal rates of aggregation. In *Proceedings of the International Conference on Learning Theory (COLT)*, 2003.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2009.

Y. Z. Tsypkin and B. T. Polyak. Attainable accuracy of adaptation algorithms. *Dokl. Akad. Nauk SSSR*, 218:532–535, 1974.

B. C. Turnbull and S. K. Ghosh. Unimodal density estimation using bernstein polynomials. *Computational Statistics & Data Analysis*, 72:13–29, 2014.

M. M. Vainberg. On the convergence of the method of steepest descents for non-linear equations. *Soviet Math. Dokl.*, 1:1–4, 1960.

L. G. Valiant. A theory of the learnable. In *Proceedings of the Symposium on Theory of Computing (STOC)*, 1984.

S. van de Geer. Estimating a regression function. *Ann. Statist.*, 18(2):907–924, 1990.

S. van de Geer. The entropy bound for monotone functions. Statistics Technical Report 91-10, Leiden Univ., 1991.

S. van de Geer. Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.*, 21(1):14–44, 1993.

A. W. van der Vaart. *Asymptotic Statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 1998.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer, 1996.

R. van Handel. *Probability in High Dimension*. 2014.

V. N. Vapnik. *Statistical Learning Theory*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. Wiley, New-York, 1998.

V. N. Vapnik and A. Y. Chervonenkis. The uniform convergence of frequencies of the appearance of events to their probabilities. *Teor. Verojatnost. i Primenen.*, 16: 264–279, 1971.

V. N. Vapnik and A. Y. Chervonenkis. *Teoriya raspoznavaniya obrazov. Statisticheskie problemy obucheniya.* Izdat. "Nauka", 1974.

J.-P. Vial. Strong and weak convexity of sets and functions. *Math. Oper. Res.*, 8(2): 231–259, 1983.

J. von Neumann. Thermodynamik quantummechanischer Gesamheiten. *Gött. Nach*, (1):273–291, 1927.

J. von Neumann. Discussion of a maximum problem. volume VI of *Collected Works*, pages 89–95. Pergamon Press, 1963. Unpublished working paper from 1947.

V. Vovk. Competitive on-line statistics. *International Statistical Review*, 69(2):213–248, 2001.

F. Wang, B. Zhao, and C. Zhang. Linear time maximum margin clustering. *IEEE Transactions on Neural Networks*, 2010.

H. Wang, F. Nie, and H. Huang. Multi-view clustering and feature learning via structured sparsity. In *Proceedings of the Conference on Machine Learning (ICML)*, 2013.

G. S. Watson. Smooth regression analysis. *Sankhyā Ser. A*, 26:359–372, 1964.

Z. Wen, D. Goldfarb, and K. Scheinberg. Block coordinate descent methods for semidefinite programming. In *Handbook on Semidefinite, Conic and Polynomial Optimization*. Springer, 2012.

A. Wibisono, A. C. Wilson, and M. I. Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113(47), 2016.

A. I. Wilson, B. Recht, and M. I. Jordan. A Lyapunov analysis of momentum methods in optimization. *arXiv preprint arXiv:1611.02635v3*, 2016.

B. E. Woodworth and N. Srebro. Tight complexity bounds for optimizing composite objectives. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Trans. Signal Process.*, 57(7):2479–2493, 2009.

L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *J. Mach. Learn. Res.*, 11:2543–2596, 2010.

L. Xu, J. Neufeld, B. Larson, and D. Schuurmans. Maximum margin clustering. In *Advances in Neural Information Processing Systems (NIPS)*, 2004.

V. A. Yakubovich. The S-procedure in nonlinear control theory. *Vestnik Leningrad. Univ.*, (1):62–77, 1971.

J. Ye, Z. Zhao, and M. Wu. Discriminative k-means for clustering. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

C.-H. Zhang. Risk bounds in isotonic regression. *Ann. Statist.*, 30(2):528–555, 04 2002.

K. Zhang, I. W. Tsang, and J. T. Kwok. Maximum margin clustering made practical. *IEEE Transactions on Neural Networks*, 2009.

T. Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the Conference on Machine Learning (ICML)*, 2004.

T. Zhang. Data dependent concentration bounds for sequential prediction algorithms. In *Proceedings of the International Conference on Learning Theory (COLT)*, 2005.

Y. Zhang and L. Xiao. Stochastic primal-dual coordinate method for regularized empirical risk minimization. In *Proceedings of the Conference on Machine Learning (ICML)*, 2015.

M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Conference on Machine Learning (ICML)*, 2003.

List of Figures

List of Tables

Résumé

De multiples problèmes en apprentissage automatique consistent à minimiser une fonction lisse sur un espace euclidien. Pour l'apprentissage supervisé, cela inclut les régressions par moindres carrés et logistique. Si les problèmes de petite taille sont résolus efficacement avec de nombreux algorithmes d'optimisation, les problèmes de grande échelle nécessitent en revanche des méthodes du premier ordre issues de la descente de gradient.

Dans ce manuscrit, nous considérons le cas particulier de la perte quadratique. Dans une première partie, nous nous proposons de la minimiser grâce à un oracle stochastique. Dans une seconde partie, nous considérons deux de ses applications à l'apprentissage automatique : au partitionnement de données et à l'estimation sous contrainte de forme.

La première contribution est un cadre unifié pour l'optimisation de fonctions quadratiques non-fortement convexes. Celui-ci comprend la descente de gradient accélérée et la descente de gradient moyennée. Ce nouveau cadre suggère un algorithme alternatif qui combine les aspects positifs du moyennage et de l'accélération.

La deuxième contribution est d'obtenir le taux optimal d'erreur de prédiction pour la régression par moindres carrés en fonction de la dépendance au bruit du problème et à l'oubli des conditions initiales. Notre nouvel algorithme est issu de la descente de gradient accélérée et moyennée.

La troisième contribution traite de la minimisation de fonctions composites, somme de l'espérance de fonctions quadratiques et d'une régularisation convexe. Nous étendons les résultats existants pour les moindres carrés à toute régularisation et aux différentes géométries induites par une divergence de Bregman.

Dans une quatrième contribution, nous considérons le problème du partitionnement discriminatif. Nous proposons sa première analyse théorique, une extension parcimonieuse, son extension au cas multi-labels et un nouvel algorithme ayant une meilleure complexité que les méthodes existantes.

La dernière contribution de cette thèse considère le problème de la sériation. Nous adoptons une approche statistique où la matrice est observée avec du bruit et nous étudions les taux d'estimation minimax. Nous proposons aussi un estimateur computationellement efficace.

Mots Clés

Optimisation convexe, accélération, moyennage, gradient stochastique, régression par moindres carrés, approximation stochastique, algorithme dual moyenné, descente miroire, partitionnement discriminatif, relaxation convexe, parcimonie, sériation statistique, apprentissage de permutation, estimation minimax, contraintes de forme.

Abstract

Many problems in machine learning are naturally cast as the minimization of a smooth function defined on a Euclidean space. For supervised learning, this includes least-squares regression and logistic regression. While small problems are efficiently solved by classical optimization algorithms, large-scale problems are typically solved with first-order techniques based on gradient descent.

In this manuscript, we consider the particular case of the quadratic loss. In the first part, we are interested in its minimization when its gradients are only accessible through a stochastic oracle. In the second part, we consider two applications of the quadratic loss in machine learning: clustering and estimation with shape constraints.

In the first main contribution, we provided a unified framework for optimizing non-strongly convex quadratic functions, which encompasses accelerated gradient descent and averaged gradient descent. This new framework suggests an alternative algorithm that exhibits the positive behavior of both averaging and acceleration.

The second main contribution aims at obtaining the optimal prediction error rates for least-squares regression, both in terms of dependence on the noise of the problem and of forgetting the initial conditions. Our new algorithm rests upon averaged accelerated gradient descent.

The third main contribution deals with minimization of composite objective functions composed of the expectation of quadratic functions and a convex function. We extend earlier results on least-squares regression to any regularizer and any geometry represented by a Bregman divergence.

As a fourth contribution, we consider the the discriminative clustering framework. We propose its first theoretical analysis, a novel sparse extension, a natural extension for the multi-label scenario and an efficient iterative algorithm with better running-time complexity than existing methods.

The fifth main contribution deals with the seriation problem. We propose a statistical approach to this problem where the matrix is observed with noise and study the corresponding minimax rate of estimation. We also suggest a computationally efficient estimator whose performance is studied both theoretically and experimentally.

Keywords

Convex optimization, acceleration, averaging, stochastic gradient, least-squares regression, stochastic approximation, dual averaging, mirror descent, discriminative clustering, convex relaxation, sparsity, statistical seriation, permutation learning, minimax estimation, shape constraints.