Analysis and Acceleration of Gradient Descents and Gossip Algorithms

Raphaël Berthier

PSL UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'Ecole Normale Supérieure Ulm

Analyse et Accélération des Descentes de Gradient et des Algorithmes de Gossip

Soutenue par
Raphaël BERTHIER
Le 13 septembre 2021

Ecole doctorale n°386
Ecole Doctorale Sciences Mathématiques de Paris Centre

Spécialité
Mathématiques

Composition du jury :

Gabriel PEYRE Directeur de recherche, ENS Ulm	*Président*
Lorenzo ROSASCO Assistant professor, University of Genova	*Rapporteur*
Anatoli JUDITSKY Professeur, Université Grenoble-Alpes	*Rapporteur*
Prateek JAIN Ingénieur de recherche, Google Research	*Examinateur*
Francis BACH Directeur de recherche, ENS Ulm	*Directeur de thèse*
Pierre GAILLARD Chargé de recherche, Inria Grenoble	Co-encadrant

ENS PSL

RÉSUMÉ

Motivés par l'intérêt récent porté à l'apprentissage statistque et au calcul distribué, nous étudions l'optimisation convexe stochastique et les algorithmes de gossip en parallèle. Cette étude jointe est rendue possible grâce à des relations rigoureuses qui sont faites entre les structures de problèmes d'optimisation et leurs équivalents pour les algorithmes de gossip. La forte convexité d'un problème d'optimisation correspond au trou spectral entre les deux plus petites valeurs propres du Laplacien pour les algorithmes de gossip. Les conditions de capacité et de source d'un problème de moindres carrés, qui décrivent les lois de puissance des valeurs propres et de la projection de l'optimum sur les vecteurs propres, correspondent à la dimension spectrale du graphe pour les algorithmes de gossip.
Voir ci-dessous pour un résumé plus long.

MOTS CLÉS

optimisation, gossip, stochastique, distribué, aynchrone

ABSTRACT

Motivated by the recent interest in statistical learning and distributed computing, we study stochastic convex optimization and gossip algorithms in parallel. This joint study is enabled by rigorous relationships that are made between the structures of optimization problems and their equivalents for gossip algorithms. The strong convexity of an optimization problem corresponds to the spectral gap between the two smallest eigenvalues of the graph Laplacian for gossip algorithms. The capacity and source conditions of a least-squares problem, that describe power-law scalings for the eigenvalues and for the projection of the optimum against the eigenvectors, correspond to the spectral dimension of the graph for gossip algorithms.
See below for a longer abstract.

KEYWORDS

optimization, gossip, stochastic, distributed, asynchronous

Analysis and Acceleration of Gradient Descents and Gossip Algorithms

Raphaël Berthier

Inria - Département d'informatique de l'ENS
PSL Research University, Paris, France

PhD thesis under the supervision of Francis Bach and Pierre Gaillard

September 13, 2021

Abstract. Motivated by the recent interest in statistical learning and distributed computing, we study stochastic convex optimization and gossip algorithms in parallel. This joint study is enabled by rigorous relationships that are made between the structures of optimization problems and their equivalents for gossip algorithms. The strong convexity of an optimization problem corresponds to the spectral gap between the two smallest eigenvalues of the graph Laplacian for gossip algorithms. The capacity and source conditions of a least-squares problem, that describe power-law scalings for the eigenvalues and for the projection of the optimum against the eigenvectors, correspond to the spectral dimension of the graph for gossip algorithms.

In this common framework, our first contribution is to study the convergence rates of naive algorithms: stochastic gradient descent and the simple gossip algorithm. We largely focus on obtaining non-parametric rates in the noiseless case, typical of interpolation problems.

As the naive methods prove to be suboptimal, we propose two new techniques to accelerate them.

First, we propose so-called continuized accelerations to tackle the problem of asynchrony in accelerating distributed algorithms, like gossip algorithms. We model this asynchrony by assuming that communications and gradient steps happen at random times, and we adapt classical accelerations to this setting. Interestingly, the resulting continuized framework gives an insightful perspective even on the classical centralized acceleration of Nesterov.

Second, we propose an acceleration of gossip algorithms—called the Jacobi Polynomial Iteration—depending on the spectral dimension of the communication network. This contrasts with previous accelerations based on the spectral gap; taking into account the spectral dimension brings a significant improvement on large networks, in the non-asymptotic regime. This acceleration is derived in two different ways: using parallel techniques in optimization called polynomial-based iterative methods, or through its scaling on large graphs to a partial differential equation that mixes quickly.

Résumé. Motivés par l'intérêt récent porté à l'apprentissage statistque et au calcul distribué, nous étudions l'optimisation convexe stochastique et les algorithmes de gossip en parallèle. Cette étude jointe est rendue possible grâce à des relations rigoureuses qui sont faites entre les structures de problèmes d'optimisation et leurs équivalents pour les algorithmes de gossip. La forte convexité d'un problème d'optimisation correspond au trou spectral entre les deux plus petites valeurs propes du Laplacien pour les algorithmes de gossip. Les conditions de capacité et de source d'un problème de moindres carrés, qui décrivent les lois de puissance des valeurs propres et de la projection de l'optimum sur les vecteurs propres, correspondent à la dimension spectrale du graphe pour les algorithmes de gossip.

Dans ce cadre commun, notre première contribution est d'étudier les vitesses de convergence des algorithmes naïfs : la descente de gradient stochastique et l'algorithme de gossip simple. On se concentre principalement sur l'obtention de taux non-paramétriques dans le cas sans bruit additif, qui est typique des problèmes d'interpolation.

Comme les méthodes naïves se révèlent être sous-optimales, nous proposons deux techniques pour les accélérer.

Premièrement, nous proposons des accélérations dites continuisées pour résoudre le problème de l'asynchronie dans l'accélération des algorithmes distribués tels que les algorithmes de gossip. Nous modélisons cette asynchronie en faisant l'hypothèses que les communications et les pas de gradient ont lieu à des instants aléatoires, et nous adaptons les accélérations classiques à ce scénario. Curieusement, ce cadre continuisé apporte une perspective intuitive même pour l'accélération centralisée classique de Nesterov.

Deuxièmement, nous proposons une accélération des algorithmes de gossip, appelée itération des polynômes de Jacobi, qui dépend de la dimension spectrale du réseau de communication. Cela contraste avec les accélérations précédentes basées sur le trou spectral ; prendre en compte la dimension spectrale apporte une amélioration significative sur des grands réseaux, dans un régime non-asymptotique. Cette accélération est construite de deux manières différentes : en utilisant des techniques parallèles en optimisation appelées méthodes itératives par polynômes, et au travers de sa limite d'échelle sur des grands graphes vers une équation aux dérivées partielles qui mélange rapidement.

Remerciements

Francis et Pierre, un grand merci pour votre encadrement bienveillant tout au long de cette expérience. Vous avez su me guider astucieusement tout en me laissant la dose d'exploration, de liberté et d'erreurs qui permettent de se construire une curiosité, un esprit critique et une modestie (respectivement). J'espère tirer le meilleur de votre enseignement et imiter votre pédagogie patiente et discrète à l'avenir.

Ensuite, je remercie tous les membres du jury—Anatoli Judistky, Lorenzo Rosasco, Prateek Jain et Gabriel Peyré—qui me font l'honneur d'examiner ma thèse. Le précieux retour des rapporteurs m'a permis de mieux appréhender comment mes travaux pouvaient être reçus, et d'améliorer de nombreux points de ce manuscrit.

Merci à Andrea Montanari et Gérard Ben Arous qui m'ont offert des expériences exceptionnelles de recherche à l'étranger. Ces voyages ont forgé ma motivation.

Merci à l'équipe Willow–Sierra–Dyogene de l'Inria, où j'ai pu trouver d'excellents collaborateurs, des collègues extraordinaires pour demander un coup de pouce scientifique ou jouer avec une belle idée, mais aussi des complices pour m'initier à l'escalade, randonner dans les calanques à Marseille, se baigner dans une piscine de fluide non-Newtonien, mettre le plus de piment possible dans un banane plantain, etc. Vous restez le groupe qui a le plus déliré sur le jeu du "hip" (sans les dents), et pour moi ça veut dire beaucoup.

J'ai aussi eu l'immense chance de rencontrer un autre cluster de l'Inria—la "team admin" dans ma tête—qui a su apporter de la légèreté et une autre forme de gossip à une thèse avec ses inévitables hauts et bas. Je m'excuse encore une fois auprès de Julien et Antoine pour les nombreuses défaites qu'ils ont dû essuyer contre Loucas et moi au babyfoot ; l'humiliation n'était pas intentionnelle. Mais bref, merci à tous d'avoir été aussi joyeux et généreux.

Enfin, merci à mes (autres) amis et à ma famille, pour votre présence et votre soutien constant.

Acknowledgements. This thesis was funded by Inria and the DGA.

Contents

Vue d'ensemble de la thèse et des contributions

Ce long résumé suppose que le lecteur a des connaissances en optimisation, statistiques et algorithmes de gossip. Le lecteur peut choisir de lire d'abord l'introduction (Chapitre 1), puis de revenir à ce résumé.

L'optimisation du premier ordre recherche le minimum d'une fonction à partir de requêtes du gradient en des points choisis. Les algorithmes de gossip sont des sous-routines qui diffusent l'information à travers les réseaux dans les algorithmes distribués décentralisés. Cette thèse étudie les taux de convergence et les accélérations des algorithmes d'optimisation convexe (stochastique) et des algorithmes de gossip en parallèle. Les experts savent largement que certaines techniques peuvent être transposées d'un domaine à l'autre ; cependant, le Chapitre 1 introductif rend le parallèle plus rigoureux et détaillé.

Nous discutons et illustrons les similarités formelles entre les structures des deux problèmes : une analyse et une conception algorithmique communes sont possibles. Nous soulignons que les algorithmes de gossip correspondent à une classe spéciale d'algorithmes d'optimisation stochastique, appelée optimisation stochastique sans bruit, où les gradient stochastiques sont observés sans aucun bruit additif. Nous formalisons la relation entre l'hypothèse de forte convexité en optimisation convexe et l'hypothèse de trou spectral en gossip ; cette relation sous-tend de nombreuses contributions au problème de gossip. Nous introduisons un nouveau parallèle similaire entre les conditions de capacité et de source en optimisation et une hypothèse de dimension spectrale en gossip.

Cependant, une différence importante limite le parallèle entre les deux domaines : les algorithmes de gossip sont contraints d'utiliser uniquement des informations locales dans le réseau, ce qui a des conséquences importantes que nous détaillons.

Dans ce cadre commun, le Chapitre 2 étudie les taux de convergence des méthodes naïves, c'est-à-dire la descente de gradient stochastique pour l'optimisation des moindres carrés et l'algorithme de gossip simple pour les algorithmes de gossip. Sous une hypothèse de forte convexité / de trou spectral, l'analyse est simple et est donnée uniquement pour mettre les résultats en perspective. La contribution principale de ce chapitre est d'étudier les taux de convergence de la descente de gradient stochastique sous conditions de source et de capacité ; cela correpond aux taux de convergence de l'algorithme de gossip simple sous une hypothèse de dimension spectrale. Cela a été publié dans l'article de conférence suivant :

> R. Berthier, F. Bach, P. Gaillard. Tight Nonparametric Convergence Rates for Stochastic Gradient Descent under the Noiseless Linear Model, 2020, *Advances in Neural Information Processing Systems (NeurIPS)*.

La présentation se concentre sur le cas sans bruit, et les taux dans le cas avec bruit additif sont donnés comme extension.

Les algorithmes naïfs se révèlent être sous-optimaux, par conséquence le reste de cette thèse se concentre sur la question de l'accélération.

Le Chapitre 3 revisite l'accélération classique de Nesterov, mais en prenant en compte le caractère asynchrone des implémentations distribuées. Nous concevons l'accélération de Nesterov "continuisée", une variante proche de l'accélération de Nesterov où les pas de gradients sont réalisés à

des instants aléatoires. De manière intéressante, cette variante donne une perspective intuitive sur l'itération d'origine de Nesterov [1983]. Cette variante continuisée bénéficie du meilleur des cadres continu et discret : en tant que processus continu, on peut utiliser le calcul différentiel pour analyser la convergence et obtenir des expressions analytiques pour les paramètres ; mais une discrétisation du processus continuisé peut être calculée exactement avec des taux de convergence similaires à ceux de l'accélération originale de Nesterov. Nous montrons que la discrétisation a la même structure que l'accélération de Nesterov, mais avec des paramètres aléatoires. Nous étendons l'accélération continuisée à l'accélération stochastique, dans le cas particulier des gradients stochastiques sans bruit. Dans ce cas, notre capacité à accélérer dépend des scores de levier ("leverage scores" en anglais) ; l'accélération obtenue réalise des garanties de performance similaires à l'accélération de Jain et al. [2018]. Pour finir, nous décrivons une application aux algorithmes de gossip asynchrone ; ce n'est pas une contribution de l'auteur mais de Even et al. [2020], qui motiva la généralisation proposée ici. La plupart du contenu de ce chapitre est disponible dans la pré-publication :

> M. Even, R. Berthier, F. Bach, N. Flammarion, P. Gaillard, H. Hendrikx, L. Massoulié, A. Taylor. A Continuized View on Nesterov Acceleration for Stochastic Gradient Descent and Randomized Gossip, 2021, preprint.

Alors que dans le Chapitre 3 nous apprenons à accélérer dans des cadres asynchrones, le Chapitre 4 est une contribution orthogonale où l'on ignore l'asynchronie des communications dans les algorithmes de gossip. Nous concevons une accélération des algorithmes de gossip synchrone en fonction de la dimension spectrale du graphe. Notre méthode montre des améliorations importantes par rapport aux algorithmes existants dans le régime non-asymptotique. Notre approche consiste en un point de vue polynomial sur les algorithmes de gossip, ainsi qu'une approximation de la mesure spectrale des graphes avec une mesure de Jacobi. Nous montrons l'efficacité de cette approche avec des simulations et des garanties de performance sur des graphes variés, comme les grilles et les réseaux aléatoires de percolation. Le contenu de ce chapitre a été publié dans l'article de journal suivant :

> R. Berthier, F. Bach, P. Gaillard. Accelerated Gossip in Networks of Given Dimension using Jacobi Polynomial Iterations, 2020, *SIAM Journal on Mathematics of Data Science (SIMODS)*.

Le Chapitre 5 étudie les algorithmes de gossip sur les réseaux (dans le sens géométrique du terme). Dans ce cas, nous concevons et analysons les algorithmes de gossip au travers de leur limite d'échelle vers des équations aux dérivées partielles (EDPs). Alors que l'algorithme de gossip simple converge vers l'équation de la chaleur, la méthode accélérée du Chapitre 4 converge vers l'équation d'Euler–Poisson–Darboux, une équation avec une dérivée du second-ordre en temps et qui homogénise plus vite. Ce point de vue EDP donne une perspective différente sur l'accélération, qui été conçue par des méthodes algébriques sur les polynômes. Ce chapitre est une version préliminaire d'un travail joint avec Mufan Li.

Overview of the thesis and of the contributions

This long summary assumes that the reader knows about optimization, statistics and gossip algorithms. The reader may choose to read the introduction (Chapter 1) first, and then come back to this summary.

First-order optimization seeks the minimum of a function from the query of its gradient at chosen points. Gossip algorithms are subroutines that diffuse information throughout networks in distributed decentralized algorithms. This thesis studies the convergence rates and accelerations of convex (stochastic) optimization algorithms and of gossip algorithms in parallel. Experts largely know that some techniques can be brought from one field to the other; however, the introductory Chapter 1 makes the parallel more rigorous and detailed.

We discuss and illustrate the formal similarities between the structures of both problems: common analysis and algorithmic design are possible. We underline that gossip algorithms correspond to a special class of stochastic optimization algorithms, called noiseless stochastic optimization, where the stochastic gradients are observed without any additive noise. We formalize the relationship between the strong convexity assumption in convex optimization and the spectral gap assumption in gossip; this relationship underlies many contributions to the gossip problem. We introduce a new similar parallel between capacity and source conditions in optimization and a spectral dimension assumption in gossip.

However, an important difference limits the parallel between the two fields: gossip algorithms are constrained to use only local information in the network, with important consequences that we detail.

In this common framework, Chapter 2 studies the convergence rates of naive methods, namely stochastic gradient descent for least-squares optimization and simple gossip for gossip algorithms. Under a strong convexity / spectral gap assumption, the analysis is simple and given only to put the results in perspective. The main contribution of this chapter is to study the convergence rates of stochastic gradient descent under source and capacity conditions; they correspond to the convergence rates of the simple gossip algorithm under a spectral dimension assumption. It was published in the following conference article:

> R. Berthier, F. Bach, P. Gaillard. Tight Nonparametric Convergence Rates for Stochastic Gradient Descent under the Noiseless Linear Model, 2020, *Advances in Neural Information Processing Systems (NeurIPS)*.

The exposition focuses on the noiseless case, while rates under additive noise are given as an extension.

Naive algorithms are shown to be sub-optimal, thus the rest of the thesis focuses on the question of acceleration.

Chapter 3 revisits the classical Nesterov acceleration, but taking into account the asynchrony of distributed implementations. We design the "continuized" Nesterov acceleration, a close variant of Nesterov acceleration where gradient steps are taken at random times. Interestingly, this variant gives an insightful perspective on the original iteration by Nesterov [1983]. This continuized variant benefits from the best of the continuous and the discrete frameworks: as a continuous process, one

can use differential calculus to analyze convergence and obtain analytical expressions for the parameters; but a discretization of the continuized process can be computed exactly with convergence rates similar to those of Nesterov original acceleration. We show that the discretization has the same structure as Nesterov acceleration, but with random parameters. We extend the continuized acceleration to stochastic acceleration, in particular to the case of noiseless stochastic gradients. In this case, our ability to accelerate depends on the leverage scores; the resulting acceleration achieves performance guarantees similar to an acceleration of Jain et al. [2018]. Finally, an application to asynchronous gossip algorithms is described; note that it is not a contribution of the author but of Even et al. [2020], that motivated the generalization proposed here. Most of the content of this chapter is available in the preprint (under review):

> M. Even, R. Berthier, F. Bach, N. Flammarion, P. Gaillard, H. Hendrikx, L. Massoulié, A. Taylor. A Continuized View on Nesterov Acceleration for Stochastic Gradient Descent and Randomized Gossip, 2021, preprint.

While in Chapter 3 we learn to accelerate in asynchronous settings, Chapter 4 is an orthogonal contribution where we ignore the asynchrony of communications in gossip algorithms. We design an acceleration of synchronous gossip algorithms depending on the spectral dimension of the graph. Our method shows an important improvement over existing algorithms in the non-asymptotic regime. Our approach stems from a polynomial-based point of view on gossip algorithms, as well as an approximation of the spectral measure of the graphs with a Jacobi measure. We show the power of the approach with simulations and performance guarantees on various graphs, such as grids and random percolation bonds. The content of this chapter is published in the following journal article:

> R. Berthier, F. Bach, P. Gaillard. Accelerated Gossip in Networks of Given Dimension using Jacobi Polynomial Iterations, 2020, *SIAM Journal on Mathematics of Data Science (SIMODS)*.

Chapter 5 studies gossip algorithms on lattices. In this case, we design and analyze gossip algorithms through their large-scale limit to partial differential equations (PDE). While the simple gossip algorithm scales to the heat equation, the accelerated method of Chapter 4 scales to the Euler–Poisson–Darboux equation, an equation with a second-order derivative in time and that homogenizes faster. This PDE point of view gives a different perspective on the acceleration, that was designed through algebraic methods on polynomials. This chapter is a preliminary version of joint work with Mufan Li.

Notations

Important conventions

n, k	non-negative integers, number of iterations. But also, rarely, k is a kernel.
t	non-negative real number, time index. But also, rarely, function defining a translation invariant kernel.
f	objective function of an optimization problem, energy function of a gossip problem.
$\mathcal{H}$	Hilbert space with scalar product $\langle ., . \rangle$ and norm $\|.\|$. The Hilbert space $\mathcal{H}$ is the set over which objective functions f are optimized. The elements of $\mathcal{H}$ are denoted with an x, or φ when they are functions.
x_*	minimizer of f.
m	dimension of $\mathcal{H}$ when it is finite, or number of vertices in a graph.
$x(1), \dots, x(m)$	components of a vector $x \in \mathbb{R}^m$.
N	number of functions in a finite sum structure, number of samples of a learning problem, or number of edges in a graph.
a	input of a learning problem.
b	output of a learning problem.
Σ	Hessian, or covariance operator.
$G = (\mathcal{V}, \mathcal{E})$	graph G, with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$.
u, v, w	vertices of a graph.
$v \sim w$	there is an edge between v and w. This is equivalent to $\{v, w\} \in \mathcal{E}$.
$\mathcal{L}$	Laplacian of a graph.
A	adjacency matrix of a graph.
D	degree matrix of a graph.
W	gossip matrix on a graph.
d	spectral dimension of a graph, typically the dimension of a lattice.
L	smoothness parameter.
μ	strong convexity parameter, or spectral gap.
α	often, regularity of the optimum or of the features.

Classical mathematical notations

∇	gradient of a function. When there are several variables, we write ∇_y to denote the gradient in the variable y.
∇^2	Hessian of a function.
Δ	Laplacian of a function.
$\nabla\cdot$	divergence of a function.
$\partial_t, \partial_{tt}$	first-order and second-order partial derivatives in the variable t.
$\preccurlyeq$	positive semi-definite order.
O	big-O notation.
o	small-o notation.
Θ	big-theta notation. $u = \Theta(v)$ if and only if $u = O(v)$ and $v = O(u)$.
$\xi \sim \mathcal{P}$	the random variable ξ has law $\mathcal{P}$.
$\mathbb{E}$	expectation.
$\mathbb{E}_{\xi\sim\mathcal{P}}$	partial expectation over ξ with law $\mathcal{P}$. Sometimes this notation is used for the full expectation $\mathbb{E}$, only to remind which variables are random in the expression.
$\mathcal{N}(x, \Sigma)$	(multivariate) Gaussian distribution with mean x and (co)variance Σ
$\mathrm{Unif}(.)$	uniform law on a finite set.
$.^\top$	transpose of a matrix or an operator.
$\otimes$	when $a \in \mathcal{H}$, $a \otimes a$ is the operator defined by the formula $(a \otimes a)x = \langle a, x\rangle a$. In the finite dimensional case, $a \otimes a = aa^\top$.
$\mathbb{1}$	vector with every coordinate equal to 1 (with dimension clear from the context). Also, $\mathbb{1}_A$ is the indicator of a random event A, i.e., takes value 1 if A holds and 0 otherwise.
a.s.	almost surely.
i.i.d.	independent identically distributed.
Tr	trace of a matrix or of a linear operator.
x_+	positive part of a real number x. If $x \geqslant 0$, $x_+ = x$ and if $x < 0$, $x_+ = 0$.
$\lfloor s \rfloor$	integer part of a real number s.

CHAPTER 1

Introduction

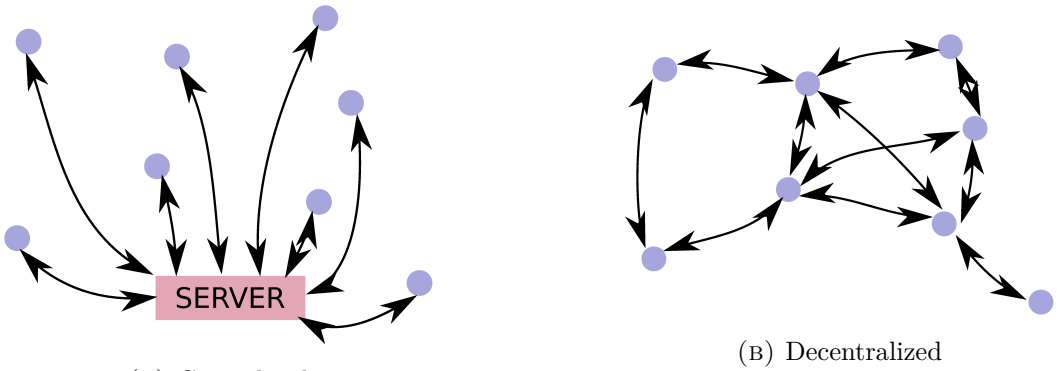
1.1. Context and motivations

Context. We ask machines more and more: to detect spams and frauds, to recognize the content of images, to understand speech, to advise in medical diagnoses, to personalize teaching and marketing to users [Jordan and Mitchell, 2015]. This has been enabled by a steady progress in computational power, memory capacity, and sensor quality, as classically illustrated by Moore's law [Moore, 2005]. Simultaneously, this technology becomes cheaper, making it accessible for more and more users and applications. An intense algorithmic research learns to use at best these new resources. As a consequence, computer science is constantly evolving: new trends emerge thanks to the new possibilities enabled by an improving technology. Let us describe a few trends that we are interested in.

First, learning algorithms have been revolutionized by the *big data* era: more data collection, more data storage, and more computational power to process the data. Modern datasets now contain a large number of observations, often high-dimensional: to give an order of magnitude, the classical COCO dataset of Lin et al. [2014] contains $300,000$ images, each one of them composed of more than $100,000$ pixels. Statistics and theoretical computer science are struggling to explain some of the recent progress of practitioners in this large scale setting. Many observations challenge the classical statistical wisdom, including the bias-variance tradeoff [Ma et al., 2018, Bartlett et al., 2021]. The success of neural networks—the dominant machine learning technique on complex, large-scale data—is not explained by the current learning theories. As a consequence, the big data era has stimulated a burst of theoretical research.

Further, large scale machine learning motivates a particular interest in first-order stochastic optimization, that allows to train a machine learning model in an online fashion (accessing the observations one after the other) and with only a few passes over the dataset [Bottou et al., 2018]. When the dataset is large, these properties of stochastic optimization methods are crucial to obtain reasonable algorithmic running times. This thesis largely deals with first-order (stochastic) optimization, with a special focus on models that are suitable for high-dimensional problems.

Finally, because of the amount of computations and/or data required, algorithms are now distributed on several machines. Distributed computing can be seen as a special case of parallel computing, where there is limited communication between machines, no memory sharing, and computations are asynchronous. These difficulties generate research in order to adapt classical single-machine algorithms to distributed settings, see for instance [Assran et al., 2020]. Among distributed networks, one opposes *centralized* and *decentralized* networks, see Figure 1.1. Centralized networks are the most common: a master node (say, a server) distributes the tasks and aggregates the information and the computations of the other nodes, the workers. In decentralized networks, there is no distinguished node that aggregates information. Decentralized networks are receiving increasing interest for their flexibility, robustness to node/links failures, and scalability. This thesis contributes to the analysis and the design of algorithms for a simple decentralized problem called the gossip problem.



(A) Centralized

(B) Decentralized

FIGURE 1.1. Cartoon picture of centralized and decentralized distributed networks.

Role of scientific theory. As the work presented in this thesis is mostly theoretical, we recall the role of scientific theory in the progress described above. Theory aims at explaining and predicting observed phenomena. In our field, this means providing practitioners with insights on the limitations of existing algorithms, and with potential candidates for faster or more precise algorithms.

However, such insights usually come with important simplifying assumptions in order to enable a theoretical study. Modeling is the art of making the assumptions that remove enough irrelevant real-world details for theory to be possible, but keep the core properties of the problem. This requires numerous discussions with experimenters. When a satisfying model is not found, toy problems model only a subset of the features of the original problem. Their formulation and resolution give intermediate steps for theoretical research. For instance, the averaging problem, or the synchronous gossip framework, both introduced below, are toy problems of decentralized computing.

Given the numerous simplifications made, theory rarely hopes to make precise quantitative predictions. The goal is only to make rough qualitative predictions. Is the algorithm stable? Will it converge to a solution? How long will it take before it outputs a solution: about 1 second, 1 day or longer than the age of the universe? What are the natural suggestions to improve such algorithm?

However, models come with an important blessing: many real-world problems with different appearances exhibit a common mathematical structure when simplified, as illustrated by the note of McKean [2003]. This diagonal perspective on science enables to make giant leaps. For instance, computer science and in particular machine learning have used many ideas coming from statistical physics [Mezard and Montanari, 2009, Zdeborová and Krzakala, 2016]. The notion of optimal transport was originally created by Monge [1781] to study the optimal movement of a resource to specified locations; it is now a large theory able, e.g., to analyze toy models of neural networks [Chizat and Bach, 2018], and a set of computation schemes used in imaging sciences and machine learning (see the review by Peyré and Cuturi [2019]). Quadratic optimization, including this thesis, has benefited from the theory of orthogonal polynomials, originally created to express solutions of equations appearing in physics [Fischer, 1996]. Finally, and more importantly to us, this thesis solidifies a more modest parallel between stochastic first-order methods and gossip algorithms.

Stucture of the rest of the introduction. First, in Section 1.2, we present convex optimization with deterministic or stochastic gradients, along with some applications to statistical learning. Second, in Section 1.3, we present the averaging problem and gossip algorithms. The materials of these sections is classical; the goal is only to clarify our objects of interests and some terminology.

In Section 1.4, we explain the relations that can be made between convex optimization and gossip algorithms.

Finally, in Section 1.5, we describe equivalent sets of assumptions in convex optimization and in gossip algorithms: strong convexity in optimization is equivalent to a spectral gap assumption in gossip algorithms, and capacity and source conditions are equivalent to a spectral dimension assumption. In each setting, we discuss the rates and accelerations. This section articulates the rest of the thesis.

1.2. Convex optimization, gradient descents and kernel methods

1.2.1. Optimization with deterministic gradients. In computer science, a large number of problems seek the *best* element x in a set $\mathcal{H}$ of options. To give an unambiguous meaning to the word "best", one can have a cost function $f : \mathcal{H} \to \mathbb{R}$ defined on the set $\mathcal{H}$. The goal of optimization theory is to minimize f on the set $\mathcal{H}$, i.e., to find an element $x \in \mathcal{H}$ with minimal or near minimal cost $f(x)$:

$$\min_{x \in \mathcal{H}} \quad f(x)\,.$$

Depending on the nature of the problem, the function f can be named differently than "cost" function. In this thesis, we refer to f as the "risk" function when dealing with learning problems, and as the "energy" function when dealing with gossip algorithms. Some other fields maximize $-f$ rather than minimizing f; for instance, reinforcement learning maximizes "rewards", economic theory maximizes "utility" functions. These various terminologies and conventions have no importance for the theory.

Given the extreme generality of this problem formulation, optimization theory has to be an enormous field. Algorithms and their performance differ depending on the way the function f can be accessed and on the properties of the function f and the set $\mathcal{H}$. In this thesis, we make several assumptions that set our work within a specific branch of optimization.

- *(unconstrained, first-order optimization)* We assume that the set $\mathcal{H}$ is a Hilbert space endowed with a scalar product $\langle ., . \rangle$. We denote $\|.\|$ the associated norm. For most applications of this thesis, $\mathcal{H} = \mathbb{R}^m$ endowed with the canonical scalar product, but the generalization to infinite-dimensional Hilbert spaces is important for applications to kernel methods. The function f is assumed to be differentiable and accessible through the computations of its gradient $\nabla f(x)$ at a chosen points x, called *queries*. In most applications, the queries are the most costly part of algorithms (in terms of time, or computational effort). Thus optimization methods seek the minimum of f in a minimal number of queries.
- *(convex optimization)* We assume that the function f is convex, i.e., for all $x, y \in \mathcal{H}$,

  $$f(y) \geqslant f(x) + \langle \nabla f(x), y - x \rangle\,.$$

  This assumption is ubiquitous in optimization to have algorithms converging to a global minimizer of the function f. The power of the convexity assumption is that it gives a global lower bound on the function f from only local information (the function value $f(x)$ and the gradient $\nabla f(x)$) on the function at a point x. We also make the following similar assumption, that gives an upper bound from local information.
- *(smooth optimization)* We assume that there exists a constant $L \geqslant 0$ such that f is L-smooth, i.e., for all $x, y \in \mathcal{H}$,

  $$f(y) \leqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2\,.$$

While in general, optimization problems can be extremely hard to solve, unconstrained smooth convex optimization problems can be solved using a simple algorithm, called *gradient descent*. It is

an *iterative method*, meaning that it starts from an initial guess $x_0 \in \mathcal{H}$ of the minimum of f and repeatedly attempts to improve it. As the gradient of f at x_0 points towards the direction of local maximal increase of f, gradient descents moves in the opposite direction to decrease f: it defines $x_1 = x_0 - \gamma_0 \nabla f(x_0)$, where γ_0 is a step-size. It then computes the gradient at x_1, takes a new gradient step, and so on:

$$x_{n+1} = x_n - \gamma_n \nabla f(x_n)\,, \qquad n \geqslant 0\,. \tag{1.1}$$

The step-sizes $\gamma_n \geqslant 0$ need to be chosen by the algorithmic designer.

Gradient descent is sometimes referred to as a "naive" method because of its simplicity, but also because it requires to know few properties of the function f. It is easy to understand, versatile, and widely used. A folklore result is that if the step-sizes are small enough, then gradient descent finds solutions $x \in \mathcal{H}$ with value $f(x)$ arbitrarily close to the optimal value.

> **Theorem 1.1** ([Nesterov, 2003, Corollary 2.1.2])**.** Let $f : \mathcal{H} \to \mathbb{R}$ be a convex, L-smooth function, minimized at a point x_*. Consider the iterates $(x_n)_{n\in\mathbb{N}}$ of gradient descent (1.1), with constant step-sizes equal to $\gamma = \frac{1}{L}$. Then
>
> $$f(x_n) - f(x_*) \leqslant \frac{2L\|x_0 - x_*\|^2}{n+4}\,.$$

In short, the excess cost $f(x_n) - f(x_*)$ is dominated by $1/n$ as $n \to \infty$. Given this theorem, one could think unconstrained convex smooth optimization is trivially solved: one simply has to run gradient descent for a sufficient number of iterations. It turns out this is not the case, for at least two reasons.

- First, the theorem states only inequalities. One can show faster convergence rates for gradient descent by assuming more properties on the function f, for instance strong convexity or a source condition (introduced below).
- Second, there exists variants of gradient descent that exhibit a better convergence rate; they are called *accelerated methods*. For instance, the celebrated Nesterov acceleration, presented below in (3.1)-(3.3), has an excess cost $f(x_n) - f(x_*)$ dominated by $1/n^2$ under the same assumptions as Theorem 1.1. In practice, we rather want to find a solution $x \in \mathcal{H}$ that achieves a given bound on the excess risk $f(x) - f(x_*) \leqslant \varepsilon$. Gradient descent achieves the bound in a number of iterations $n = O(\varepsilon^{-1})$ while Nesterov acceleration requires $O(\varepsilon^{-1/2})$ iterations: this is faster when ε is small. The rate $O(1/n^2)$ for Nesterov acceleration is *worst-case optimal* among all possible methods that access gradients and linearly combine them [Nesterov, 2003, Nemirovskij and Yudin, 1983]. This means that we can not hope to find a better method than Nesterov acceleration for minimizing general convex smooth functions. However, it is possible to design better acceleration for specific function classes. A huge literature aims at designing accelerations tailor-made for functions satisfying specific properties. See, e.g., [d'Aspremont et al., 2021] and references therein.

In this thesis, we are interested in improving over Theorem 1.1 in both directions above. In Section 1.5, we introduce the different properties of the function f that we are interested in. We give an overview of the convergence rate of naive algorithms and of the possible accelerations.

Acceleration techniques. Accelerated methods are iterations that closely resemble the gradient descent iteration (1.1). For instance, Polyak's heavy ball method [Polyak, 1964] is of the form

$$x_1 = x_0 - \gamma \nabla f(x_0)\,, \qquad x_{n+1} = x_n - \gamma \nabla f(x_n) + \beta(x_n - x_{n-1})\,, \qquad n \geqslant 1\,, \tag{1.2}$$

where $\beta \in [0, 1]$ is a new parameter that requires to be tuned. See Figure 1.2 for a toy comparison between the heavy ball method and gradient descent. The new term $\beta(x_n - x_{n-1})$ is called the

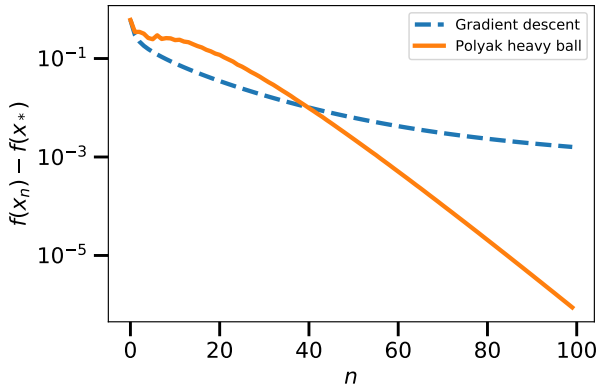


FIGURE 1.2. Comparison between the heavy ball method (1.2) and gradient descent (1.1), initialized from $x_0 = 0$, in optimizing the function $f(x) = \frac{1}{2N}\sum_{k=1}^{N}(b_k - \langle a_k, x\rangle)^2 + \frac{\lambda}{2}\|x\|^2$, where $x \in \mathbb{R}^{60}$, $N = 50$, $b_1, \dots, b_N$ are i.i.d. $\mathcal{N}(0,1)$ and independently $a_1, \dots, a_N$ are i.i.d. $\mathcal{N}(0, \mathrm{Id}_{60})$. The parameters of both algorithms were tuned by hand in order to maximize the observed asymptotic rate of convergence.

momentum. It is frequent that accelerated methods have some form of momentum terms. They can also take the form of iterations over several variables. For instance, denoting $v_n = x_n - x_{n-1}$, Polyak's heavy ball method (1.2) can be rewritten as

$$\begin{aligned} v_{n+1} &= -\gamma \nabla f(x_n) + \beta v_n \,, \\ x_{n+1} &= x_n + v_{n+1} \,. \end{aligned}$$

Let us try to give an informal high-level idea motivating momentum-based method. Gradient descent has a slow rate of convergence because in the directions where the Hessian of the optimization problem is small, the gradient is not large enough for gradient descent to make sufficient progress. Gradient descent then needs to make a large number of steps, while the gradient repeatedly points in the same direction. Accelerated methods add a momentum to solve this issue: if the gradient points in the same direction from one iterate to the next, the point x_n gains inertia (i.e., the discrete speed $x_n - x_{n-1}$ increases) to converge faster to the minimum. This intuition is best understood by looking at the scaling limit of gradient descent and accelerated gradient descent as the step-size γ converges to 0 and the number of iterations grows, see [Su et al., 2014] for instance. Gradient descent converges to the gradient flow equation $\partial_t x_t = \nabla f(x_t)$ while accelerated methods typically converge to second-order ODEs; for instance, in an appropriate scaling, Polyak's heavy ball method (1.2) converges to the ODE $\partial_{tt} x_t = \nabla f(x_t) - \kappa \partial_t x_t$ (where κ depends on the scaling of the parameters).

In the continuous-time limit, it is easier to show that accelerated methods enjoy fast convergence. However, this intuition does not readily design and prove convergence for discrete accelerated methods. Indeed, discretizing the continuous-time ODEs is not straightforward as it introduces stability and approximation errors that must be controlled, see [Zhang et al., 2018, Shi et al., 2019, Sanz-Serna and Zygalakis, 2020]. As a consequence, the ODE point of view on acceleration is mostly an intuition and rarely a method for building accelerations. Instead, accelerated methods are often built in a less intuitive way: they are tuned in order to make their proof of convergence work. Let us describe two techniques that we use in this thesis.

- *Lyapunov techniques* design accelerated methods so that a certain function is Lyapunov, i.e., a decreasing function of the number of iterations. This is the guiding principle of Nesterov acceleration [Nesterov, 2003], that we mimic in Chapter 3.
- *Polynomial-based accelerations* [Fischer, 1996] are specialized in the case of quadratic functions (see Definition 1.1). The strategy is to see iterations as polynomials in the Hessian of the problem, and to choose a wise sequence of polynomials to have fast convergence. This is the basis of the famous Chebyshev acceleration (reviewed in [d'Aspremont et al., 2021, Chapter 2]). Polynomial-based iterations are the core source of inspiration for Chapter 4.

1.2.2. Optimization with stochastic gradients. Stochastic optimization is a generalization of the framework of the previous section in which we can not query directly the gradient $\nabla f(x)$, but we can compute a stochastic estimation of it. More precisely, it is possible to compute a random quantity $g(x, \xi)$ where x is chosen, the random variable ξ is generated according to some law $\mathcal{P}$, and $\mathbb{E}_{\xi\sim\mathcal{P}} g(x, \xi) = \nabla f(x)$. The random variable ξ is independently re-generated from one query to another. Again, each query is costly (in terms of time, computation, or statistical data), thus we seek algorithms that minimize f in a minimal number of queries of stochastic gradients $g(x, \xi)$.

All the questions sketched in the previous section can be transposed in this framework. The naive algorithm, called *stochastic gradient descent* [Robbins and Monro, 1951], replaces the deterministic steps of gradient descent by stochastic steps:

$$x_{n+1} = x_n - \gamma_n g(x_n, \xi_{n+1}) \,, \qquad n \geqslant 0 \,, \tag{1.3}$$

where the random variables $\xi_1, \xi_2, \dots$ are independent identically distributed (i.i.d.) according to $\mathcal{P}$. In this setting, the typical questions we address are: does stochasticity in our estimation of the gradients hurt convergence? What are the convergence rates of stochastic gradient descent? How can we accelerate?

In the remainder of this section and the following, we list examples of stochastic optimization problems that are central to this thesis. The objective is two-fold: first, it gives concrete examples of (stochastic) optimization problems that we focus on in this thesis; and second, it serves as an introduction to an important discussion on the different natures of stochastic gradients, exposed in Section 1.4.3.

> **Example 1.1** (Additive noise model)**.** In the *additive noise model*, we assume that our computations of the gradient are perturbed by a centered additive term ξ, namely we can compute
>
> $$g(x, \xi) = \nabla f(x) + \xi \,,$$
>
> where ξ is distributed according to some law $\mathcal{P}$ satisfying $\mathbb{E}\xi = 0$ and independent of x. This model can be used to study the effect of small measurement errors or computations errors. It is also appreciated by the mathematical simplicity brought by the additive structure. As such, it is often used as a simplification of other stochastic gradients introduced below. However, in Section 1.4.3, we discuss the limitations of additive noise in understanding all behaviors of stochastic gradient descents.
>
> A wide literature studies how to obtain the fastest convergence for stochastic optimization under additive noise. This usually involves decaying step-sizes γ_n and averaging of the iterates, see [Polyak and Juditsky, 1992] for instance.

In a large number of applications, the cost function $f(x)$ is itself the expectation of a random function $f_\xi(x)$, of which we only observe random instances: $f(x) = \mathbb{E}_{\xi\sim\mathcal{P}} f_\xi(x)$. In this case, an unbiased gradient is given by the gradient of the random function: $g(x, \xi) = \nabla f_\xi(x)$. For instance, the additive noise model fits this framework with $f_\xi(x) = f(x) + \langle \xi, x \rangle$.

Example 1.2 (Finite sum [Bertsekas, 2011])**.** Finite sums are functions of the form

$$f(x) = \frac{1}{N}\sum_{k=1}^{N} f_k(x)\,.$$

The function f can be seen as the expectation of $f_k(x)$ where $k = \xi$ is uniform in $\{1,\dots,N\}$. The empirical risk of a supervised learning problem and the energy function of an averaging problem, both introduced below, are of this form. Computing the gradient of a finite sum involves the potentially costly computation of the gradients of all of its components f_k. In order to circumvent this computation, the stochastic gradient strategy uses the gradient $\nabla f_k(x)$ of a randomly selected function in order to approximate the full gradient $\nabla f(x)$:

$$x_{n+1} = x_n - \gamma_n \nabla f_{k_{n+1}}(x_n)\,,$$

where $k_1, k_2, \dots$ are i.i.d. uniform in $\{1,\dots,N\}$.

We continue with a more sophisticated example where $f(x)$ is not the expectation of a random function $f_\xi(x)$.

Example 1.3 (Coordinate gradient descent [Tseng and Yun, 2009, Nesterov, 2012, Wright, 2015])**.** In this example, we assume that we are in the finite-dimensional case $x \in \mathbb{R}^m$. We denote $x(1),\dots,x(m)$ the coordinates of x and $e_1,\dots,e_m$ the canonical basis of $\mathbb{R}^m$. In some applications, in order to avoid the costly computation of the full gradient $\nabla f(x)$, coordinate gradient descent strategies prefer to compute the derivative $\frac{\partial f}{\partial x(i)}(x)$ with respect to one coordinate $x(i)$ only, changing the selected coordinate i at each iteration. Different selection strategies are possible: cycle through the coordinates, select them randomly, or even more sophisticated Markov chain strategies [Sun et al., 2020]. The most convenient for the mathematical analysis is the random case where at each iteration $n+1$, a coordinate i_{n+1} is selected, independently of the past choices $i_1,\dots,i_n$, and, for simplicity, uniformly in $\{1,\dots,m\}$. In this setting, the naive method is coordinate gradient descent. It is the iteration

$$x_{n+1}(i_{n+1}) = x_n(i_{n+1}) - \tilde{\gamma}_n \frac{\partial f}{\partial x(i_{n+1})}(x_n)\,, \tag{1.4}$$

$$x_{n+1}(j) = x_n(j)\,, \qquad j \neq i_{n+1}\,, \qquad\qquad n \geqslant 0\,, \tag{1.5}$$

where $\tilde{\gamma}_n$ denote the step-sizes of the algorithm. In this setting where coordinates are randomly sampled, coordinate gradient descent (1.4)-(1.5) is a special case of stochastic gradient descent (1.3), where $\xi = i \sim \mathrm{Unif}(\{1,\dots,m\})$, $g(x,\xi) = m\langle \nabla f(x), e_i\rangle e_i$ and $\gamma_n = \tilde{\gamma}_n/m$. Note that indeed, the stochastic gradient $g(x,\xi)$ is unbiased, i.e.,

$$\mathbb{E}_{\xi\sim\mathcal{P}} g(x,\xi) = m\,\mathbb{E}_{i\sim\mathrm{Unif}(\{1,\dots,m\}}\left[\langle \nabla f(x), e_i\rangle e_i\right] = \nabla f(x)\,.$$

1.2.3. Quadratics, least-squares linear regression and kernel methods. A large part of this thesis specializes in the case where f is a quadratic function.

Definition 1.1 (Quadratic function)**.** The function f is said to be *quadratic* if it is of the form

$$f(x) = \frac{1}{2}\langle x, \Sigma x\rangle + \langle \tau, x\rangle + c\,,$$

where Σ is a bounded operator on $\mathcal{H}$ (in the finite-dimensional case, a matrix), $\tau \in \mathcal{H}$ and $c \in \mathbb{R}$.

As the function f is convex, its Hessian Σ is a positive semi-definite (p.s.d.) operator (in the finite-dimensional case, a p.s.d. matrix). Below, we frequently assume that there exists a minimizer $x_* \in \mathcal{H}$ of f, in which case we have the simpler formula

$$f(x) = \frac{1}{2} \langle x - x_*, \Sigma(x - x_*) \rangle + f(x_*) \, .$$

This is equivalent to assuming that τ is in the image space of the operator Σ.

Remark 1.1. The quadratic function f is L-smooth if and only $\Sigma \preccurlyeq L \, \mathrm{Id}$, where $\preccurlyeq$ denotes the positive semi-definite order.

We now continue with examples of stochastic optimization problems on quadratic functions.

Example 1.4 (Least-squares linear regression [Legendre, 1806])**.** In statistics, a fundamental task aims at understanding the relation between an output variable b from an input variable a. For instance, computer-aided medicine wants to predict the probability b that a patient will develop a given disease given some medical information a on the patient (DNA information, blood tests, etc). Computer vision wants to understand the object b represented in an image (e.g., a cat or a dog?) from the array a of the colors of the pixels.

The relation between the variables (a, b) is learned from empirical observations (a_1, b_1), $\ldots$, (a_N, b_N), called *samples*, obtained through experiments or observations of the environment. The ultimate goal of statistics is to be able to predict the output b corresponding to a new input a. There is thus a *generalization* challenge: we want the computer to learn the relation between a and b sufficiently well to be able to predict the output b even for outputs a that are not in the database $a_1, \ldots, a_n$.

The input a and the output b can have various formats, but it is convenient for us to assume that a is a vector in a Hilbert space $\mathcal{H}$ and b is a real number. In this case, it is frequent to assume a linear relation between the output b and the input a: we assume that a relation $b = \langle a, x \rangle$ should hold for some $x \in \mathcal{H}$, at least approximately.

We find the parameter $x \in \mathcal{H}$ by fitting to the database: we seek the minimizer of the empirical risk

$$f(x) = \frac{1}{2N} \sum_{k=1}^{N} \left(b_k - \langle a_k, x \rangle \right)^2 \, . \tag{1.6}$$

Once a point $x \in \mathcal{H}$ with a low empirical risk is found, it is possible to predict the output b corresponding to a new input a with the predictor $\widehat{b} = \langle a, x \rangle$.

With this strategy, one has reduced the statistical problem of learning the relationship between a and b to the optimization problem of minimizing f. What is the structure of the optimization problem? The empirical risk f is a convex and smooth function. It is quadratic with Hessian $\Sigma = \frac{1}{N} \sum_{k=1}^{N} a_k \otimes a_k$. (If $a \in \mathcal{H}$, $a \otimes a$ is the operator defined by the formula $(a \otimes a)x = \langle a, x \rangle a$. In the finite dimensional case, $a \otimes a = a a^\top$.) Finally, the empirical risk f is a finite-sum in the sense of Example 1.2, thus one can use the stochastic gradient descent for finite sums in this special case:

$$x_{n+1} = x_n + \gamma_n (b_{k_{n+1}} - \langle a_{k_{n+1}}, x_n \rangle) a_{k_{n+1}} \, , \qquad n \geqslant 0 \, . \tag{1.7}$$

where $k_1, k_2, \ldots$ are i.i.d. uniform in $\{1, \ldots, N\}$. Here, the computational advantage of stochastic gradient descent is that computing the gradient of one component of the sum requires to read only one sample (a_k, b_k) of the database, while the computation of a full gradient $\nabla f(x)$ would require reading the full database.

Many variants of this strategy exist, for instance in the case of binary classification, where the output b can take only two possible values (e.g., a cat or a dog). In this case, we can arbitrarily fix the two values to be encoded as $b = +1$ and $b = -1$. A common strategy is then to seek a relation of the form $b = \text{sign}\langle a, x\rangle$, i.e., a composition of a linear regression and the sign function. However, minimizing the empirical risk

$$f(x) = \frac{1}{2N} \sum_{k=1}^{N} \mathbb{1}_{\{b_k = \text{sign}\langle a_k, x\rangle\}}$$

is harder as the function f is not convex (it is piecewise constant). One needs to use convex surrogates of this quantity; for instance, logistic regression minimizes the empirical risk

$$f(x) = \frac{1}{2N} \sum_{k=1}^{N} \log\left(1 + e^{-b_k \langle a_k, x\rangle}\right) .$$

Once a point $x \in \mathcal{H}$ with a low empirical logistic risk is found, it is possible to predict the output b corresponding to a new input a with the predictor $\widehat{b} = \text{sign}\langle a, x\rangle$.

In fact, minimizing the empirical risk as above can be a bad idea due to the potential *overfitting*: the computer finds a point $x \in \mathcal{H}$ that fits well for all points of the database ($b_k \approx \langle a_k, x\rangle$), but that does not generalize well, meaning that b is not well approximated by $\langle a, x\rangle$ for new samples (a, b). For an illustration of this phenomena, see Figure 1.3. This is a statistical problem, seemingly unrelated to the optimization problem of finding a minimizer of the empirical risk. To avoid overfitting, a common strategy is to *penalize* the empirical risk: the optimized function f is changed to a linear combination of the empirical risk and a penalization term controlling the complexity of the predictor x; typically, this penalization is the 2-norm or 1-norm of the vector x. Another strategy is to stop (stochastic) gradient descent early before convergence [Yao et al., 2007].

However, when the observations $(a_1, b_1), \dots, (a_N, b_N)$ are i.i.d. according to some law $\mathcal{P}$, another solution is to design stochastic optimization methods that control directly the generalization error.

Example 1.5 (Supervised learning)**.** In this thesis, the supervised learning setting refers to the least-squares linear regression setting above in the special case $(a_1, b_1), \dots, (a_N, b_N)$ are i.i.d. according to some law $\mathcal{P}$. In this setting, the statistical question of learning a linear relationship between a and b is well-defined: we seek a point $x \in \mathcal{H}$ that minimizes the so-called *population risk*

$$f(x) = \frac{1}{2}\mathbb{E}_{(a,b)\sim\mathcal{P}} \left(b - \langle a, x\rangle\right)^2 .$$

The optimization of the population risk matches more truthfully our statistical goal than the optimization of the empirical risk: we want to have a linear relation $b \approx \langle a, x\rangle$ that fits well for a new data sample $(a, b) \sim \mathcal{P}$ rather than on the dataset $(a_1, b_1), \dots, (a_N, b_N)$. The population risk is convex, smooth and quadratic with Hessian $\Sigma = \mathbb{E}_{(a,b)\sim\mathcal{P}} a \otimes a$. However it is impossible to compute its gradients as it involves an expectation over a law $\mathcal{P}$ unknown to the algorithm. As opposed to Examples 1.2-1.4 where (deterministic) gradient descent strategies were avoided because of their computational cost, here computing exactly the gradients of the population risk is impossible because we do not have the necessary information. However, we can extract from the samples sufficient information to have stochastic gradients of the population risk. The population risk is an expectation:

$$f(x) = \mathbb{E}_{(a,b)\sim\mathcal{P}} f_{(a,b)}(x) , \qquad f_{(a,b)}(x) = \frac{1}{2} \left(b - \langle a, x\rangle\right)^2 .$$

For functions of this form, we have seen in Section 1.2.2 that it is natural to build stochastic gradients

$$g(x,(a,b)) = \nabla f_{(a,b)}(x) = -(b - \langle a, x\rangle)a\,.$$

We can use the statistical samples $(a_1,b_1),\dots,(a_N,b_N)$ to generate such stochastic gradients. However, we need the randomness $\xi = (a,b) \sim \mathcal{P}$ to be independent from one iteration to another, thus we can use each sample at one iteration only. This gives the stochastic gradient iteration

$$x_{n+1} = x_n + \gamma_n(b_{n+1} - \langle a_{n+1}, x_n\rangle)a_{n+1}\,, \qquad 0 \leqslant n < N\,. \tag{1.8}$$

Note that here, the number of iterations is bounded by the number of data samples. Thus any accelerated method requiring less iterations would not only save time and computations, but also require less data collection.

The stochastic gradient descents (1.7) and (1.8) obtained for the minimization of the empirical risk and of the population risk are extremely similar; the only difference lies in the choice of the index of the data sample. When only one pass on the data is allowed, without replacement of the data samples, stochastic gradient descent can be seen as minimizing the population risk, while when multiple passes on the data are done, with replacement of the data samples, stochastic gradient descent is seen as minimizing the empirical risk.

Feature maps and kernel methods (see, e.g., [Hofmann et al., 2008] and references therein). In general, restricting ourselves to linear relations $b \approx \langle a, x\rangle$ is too restrictive to approximate potentially complex dependencies between the input a and the output b. Moreover, the input could be in a data format which is not naturally a Hilbert space. Let us now consider a non-structured input u which belongs to a set $\mathcal{U}$, not necessarily a Hilbert space. Again, we would like to understand the relationship between the input variable $u \in \mathcal{U}$ and the output variable $b \in \mathbb{R}$.

A natural strategy is to transform u through a *feature map* $\Psi : \mathcal{U} \to \mathcal{H}$ so that the feature vectors $a = \Psi(u)$ belongs to a Hilbert space $\mathcal{H}$. One can then apply linear regression to the transformed input a, as described in Examples 1.4-1.5. This can be done even if $\mathcal{U}$ is already a Hilbert space, in order to enlarge the expressive power of linear regression. For instance, if $u \in \mathbb{R}$, the linear maps in $a = (1, u, u^2) \in \mathbb{R}^3$ are the second-order polynomials in u, see Figure 1.3.

The kernel trick enables to perform the above feature map implicitly, without computing the feature vectors $a = \Psi(u)$. It only requires to be able to compute the dot products in feature space

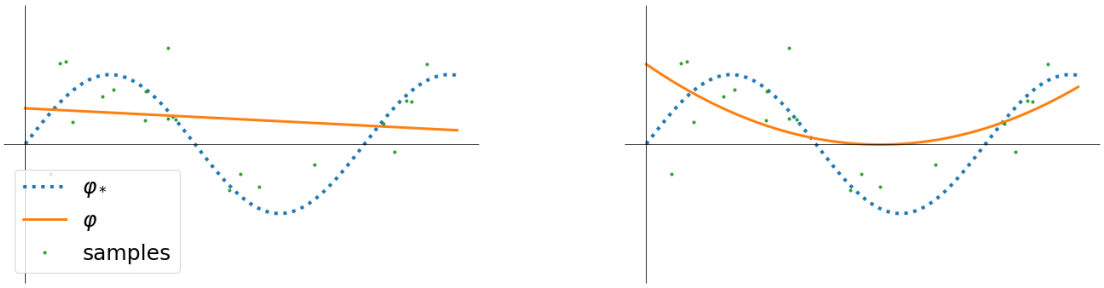
$$k(u,u') = \langle \Psi(u), \Psi(u')\rangle\,.$$

Definition 1.2 (implied by Theorem 1.2 below)**.** Let $k : \mathcal{U} \times \mathcal{U} \to \mathbb{R}$ be a symmetric function. Then the two following statements are equivalent:

- for all $m \geqslant 1$, for all $u_1,\dots,u_m \in \mathcal{U}$, the matrix $(k(u_i,u_j))_{1\leqslant i,j\leqslant m}$ is positive semi-definite, and
- there exists a Hilbert space $\mathcal{H}$ endowed with a scalar product $\langle .,.\rangle$ and a feature map $\Psi : \mathcal{U} \to \mathbb{R}$ such that for all $u,u' \in \mathcal{U}$, $k(u,u') = \langle \Psi(u), \Psi(u')\rangle$.

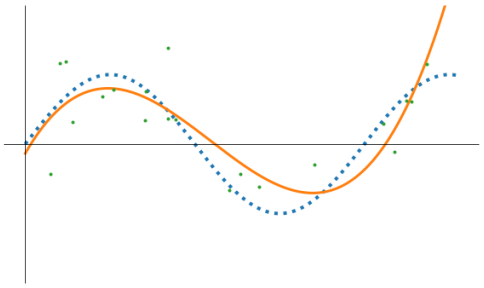
If these conditions hold, k is called a *positive definite kernel* on $\mathcal{U}$.

The above result illustrates well the spirit of RKHS theory: there exists simple conditions on a function $k : \mathcal{U} \times \mathcal{U} \to \mathbb{R}$ ensuring that it corresponds to a scalar product in a Hilbert space after a feature map. The knowledge of the feature map Ψ is not important to us as iterations can be written in terms of k directly (see Examples 1.6 and 1.7).

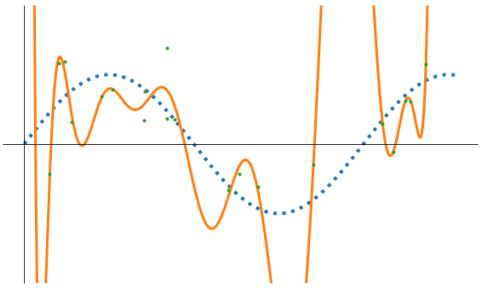


(A) order 1 (linear regression)

(B) order 2



(C) order 3



(D) order 15

FIGURE 1.3. Regressions using polynomials of order 1, 2, 3, and 15. The samples $(u_1, b_1), \ldots, (u_{20}, b_{20})$ were independently generated, with $u \sim \mathrm{Unif}([0,1])$ and $b = \varphi_*(u) + \xi$ where ξ is an independent additive noise. We took $\varphi_*(u) = \sin(8u)$ and ξ a centered Gaussian random variable with variance 0.25. For each plot, we show the polynomial φ minimizing the empirical risk $f(\varphi) = \frac{1}{40}\sum_{k=1}^{20}(b_k - \varphi(u_k))^2$. This is done by performing a linear regression over the feature vector $\Psi(u) = (1, u)$, $\Psi(u) = (1, u, u^2)$, $\Psi(u) = (1, u, u^2, u^3)$, or $\Psi(u) = (1, u, u^2, \ldots, u^{15})$. For low-order polynomials, the quality of the approximation improves as we enrich the function class. However, the interpolation with a polynomial of order 15 suffers from overfitting.

In this setting, it is convenient to think of an element $x \in \mathcal{H}$ as a function φ_x defined on $\mathcal{U}$ by the formula

$$\varphi_x(u) = \langle x, \Psi(u) \rangle .$$

Actually, the reproducing kernel Hilbert space (RKHS) theory defines the Hilbert space $\mathcal{H}$ in Definition 1.2 directly as a space of functions.

Theorem 1.2 (Moore-Aronszajn [Aronszajn, 1950])**.** Let $k : \mathcal{U} \times \mathcal{U} \to \mathbb{R}$ be a positive definite kernel. Then there exists a unique Hilbert space $\mathcal{H}$ of functions on $\mathcal{U}$ endowed with a scalar product $\langle ., . \rangle$, such that

- for all $u \in \mathcal{U}$, $k(u, .) \in \mathcal{H}$, and
- (reproducing property) for all $\varphi \in \mathcal{H}$, for all $u \in \mathcal{U}$, $\langle \varphi, k(u, .) \rangle = \varphi(u)$.

In the RKHS terminology, these properties mean that $\mathcal{H}$ is a reproducing kernel Hilbert space with reproducing kernel k.

Note that k can then be interpreted as the dot product in the feature $\Psi(u) = k(u, .) \in \mathcal{H}$. Indeed, using the reproducing property,

$$\langle \Psi(u), \Psi(u') \rangle = \langle k(u, .), k(u', .) \rangle = k(u, u') \,.$$

We now illustrate how stochastic gradient descent can be "kernelized" [Ying and Pontil, 2008, Tarrès and Yao, 2014, Rosasco and Villa, 2015, Dieuleveut and Bach, 2016].

Example 1.6 (Supervised learning, with kernels)**.** We adapt the supervised learning setting of Example 1.5 to the kernel setting. We now want to learn the relationship between an input $u \in \mathcal{U}$ and an output $b \in \mathbb{R}$, following some joint law $(u, b) \sim \mathcal{P}_{(u,b)}$. We have access to i.i.d. samples $(u_1, b_1), \dots, (u_N, b_N) \sim \mathcal{P}_{(u,b)}$. Let k be a positive definite kernel on $\mathcal{U}$.

The kernel k implicitly defines a reproducing kernel Hilbert space $\mathcal{H}$. We want to perform a linear regression in the features $a_1 = k(u_1, .), \dots, a_N = k(u_N, .) \in \mathcal{H}$, as it is done in Example 1.5. The elements of $\mathcal{H}$ are now functions on $\mathcal{U}$; we thus denote them with the symbol φ rather than x. The stochastic gradient iteration (1.8) writes

$$\begin{aligned} \varphi_{n+1} &= \varphi_n + \gamma_n (b_{n+1} - \langle a_{n+1}, \varphi_n \rangle) a_{n+1} \,, \\ &= \varphi_n + \gamma_n (b_{n+1} - \langle k(u_{n+1}, .), \varphi_n) k(u_{n+1}, .) \\ &= \varphi_n + \gamma_n (b_{n+1} - \varphi_n(u_{n+1})) k(u_{n+1}, .) \,, \qquad 0 \leqslant n < N \,. \end{aligned}$$

This is a stochastic gradient descent, on the RKHS $\mathcal{H}$, of the risk function

$$f(\varphi) = \frac{1}{2} \mathbb{E}_{(u,b) \sim \mathcal{P}_{(u,b)}} \left(b - \varphi(u) \right)^2 \,, \tag{1.9}$$

where the scalar product structure used (defining the gradients) is the one of the Hilbert space $\mathcal{H}$.

Note that similarly, it is possible to "kernelize" the stochastic gradient descent for the minimization of the empirical risk from Example 1.4. We now turn to an important special case of Example 1.6.

Example 1.7 (Function interpolation from values at random points)**.** Let $\varphi_* : \mathcal{U} \to \mathbb{R}$ be a function that we want to regress from the observation of its values at random points $(u_1, \varphi_*(u_1)), \dots, (u_N, \varphi_*(u_N))$, where $u_1, \dots, u_N$ are i.i.d. from some law $\mathcal{P}_u$. We seek a function φ that minimizes the $L^2(\mathcal{P}_u)$-distance to φ_*:

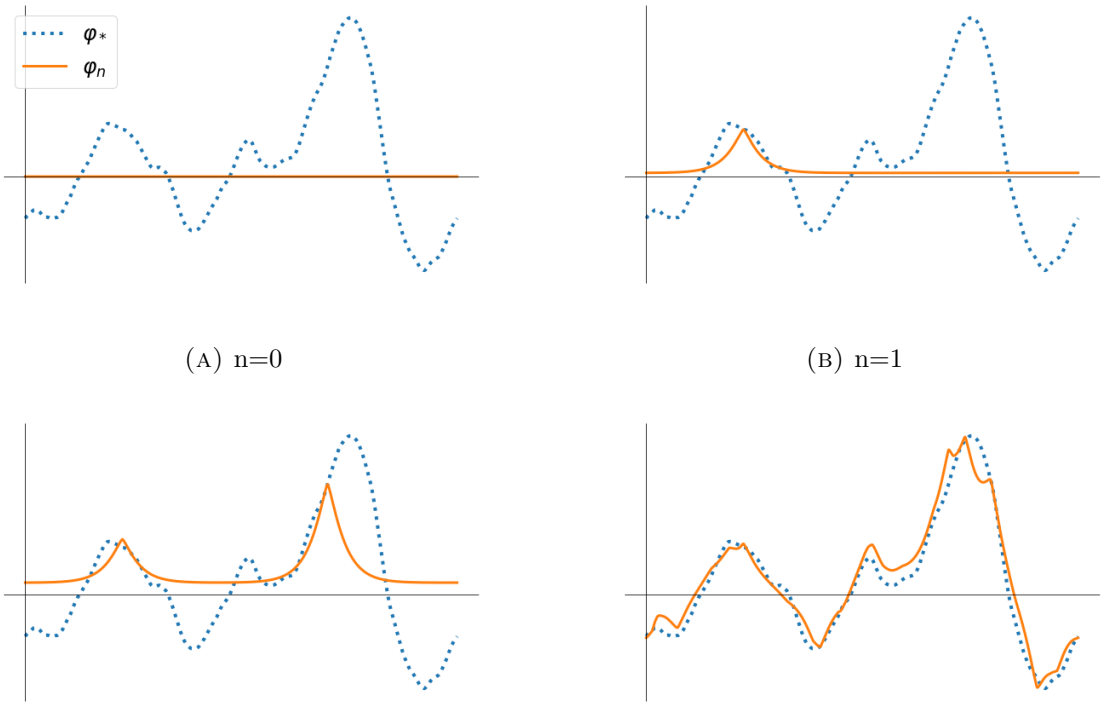
$$f(\varphi) = \frac{1}{2} \left\| \varphi_* - \varphi \right\|^2_{L^2(\mathcal{P}_u)} = \frac{1}{2} \mathbb{E}_{u \sim \mathcal{P}_u} \left(\varphi_*(u) - \varphi(u) \right)^2 \,.$$

This is a special case of (1.9) with $b = \varphi_*(u)$. Following Example 1.6, we choose a positive definite kernel k on $\mathcal{U}$ and we run the associated stochastic gradient descent

$$\varphi_{n+1} = \varphi_n + \gamma_n (\varphi_*(u_{n+1}) - \varphi_n(u_{n+1})) k(u_{n+1}, .) \,, \qquad 0 \leqslant n < N \,. \tag{1.10}$$

This update rule corrects φ_n so that $\varphi_{n+1}(u_{n+1})$ is closer to the observed value $\varphi_*(u_{n+1})$ than $\varphi_n(u_{n+1})$. Points u near u_{n+1}, in the sense that $k(u_{n+1}, u)$ is large, are also updated in the same direction.

In Figure 1.4, we illustrate the interpolation of a function φ_* on $\mathcal{U} = [0, 1]$ using stochastic gradient descent (1.10) with a translation invariant kernel.



(A) n=0 (B) n=1

(C) n=2 (D) n=40

FIGURE 1.4. Interpolation of a function φ_* on $\mathcal{U} = [0,1]$ (seen as periodic, i.e., a torus) from the observation of its values at random points, using kernel stochastic gradient descent (1.10) initialized from $\varphi_0 = 0$. Here, the chosen kernel k is translation invariant, meaning that there exists a function $t : [0,1] \to \mathbb{R}_{\geqslant 0}$ such that $k(u,u') = t(u-u')$. The function φ_* is in blue and the iterates φ_n are in orange.

We now interpret the covariance operator in the kernel setting.

Remark 1.2. Denote $\mathcal{P}_u$ the marginal of u under $\mathcal{P}_{(u,b)}$. In the above example, the covariance operator $\Sigma : \mathcal{H} \to \mathcal{H}$ is defined by the formula: for $\varphi \in \mathcal{H}$,

$$\begin{aligned}\Sigma\varphi &= \mathbb{E}\left[a \otimes a\right]\varphi = \mathbb{E}_{u\sim\mathcal{P}_u}\left[k(u,.) \otimes k(u,.)\right]\varphi = \mathbb{E}_{u\sim\mathcal{P}_u}\left[\langle k(u,.),\varphi\rangle\, k(u,.)\right] \\ &= \mathbb{E}_{u\sim\mathcal{P}_u}\left[k(u,.)\varphi(u)\right]\,.\end{aligned} \tag{1.11}$$

In words, Σ is the integral operator associated to the kernel k (see, e.g., [Cucker and Zhou, 2007]). In particular,

$$\langle\varphi,\Sigma\varphi\rangle = \mathbb{E}_{u\sim\mathcal{P}_u}\left[\langle\varphi,k(u,.)\varphi(u)\rangle\right] = \mathbb{E}_{u\sim\mathcal{P}_u}\left[\varphi(u)^2\right] = \|\varphi\|^2_{L^2(\mathcal{P}_u)}\,.$$

Thus, we can define the norm associated to the RKHS $\mathcal{H}$ in terms of the covariance operator:

$$\|\varphi\|^2 = \|\Sigma^{-1/2}\varphi\|^2_{L^2(\mathcal{P}_u)}\,. \tag{1.12}$$

One should not be misled by this equation: by definition, the RKHS and its norm $\|.\|$ depend only on the kernel k and not on the distribution of the input $\mathcal{P}_u$.

We finish this section by using Equation (1.12) in order to compute an important example of a RKHS.

Example 1.8 (Sobolev spaces)**.** Assume further $\mathcal{U} = [0,1]^d$ (seen as periodic, i.e., a torus) and that k is a translation-invariant kernel: $k(u,u') = t(u-u')$ where t is a square-integrable 1-periodic function on $[0,1]^d$. The kernel k is positive-definite if and only if the Fourier transform of t is positive [Wahba, 1990]. This imposes, in particular, that t is maximal at 0. In this setting, we relate the RKHS to the Sobolev smoothness of the function t.

A function $\varphi \in L^2([0,1]^d)$ with Fourier series $\widehat{\varphi}$ belongs to the Sobolev space H^s_{per} if

$$\|\varphi\|^2_{H^s_{\mathrm{per}}} = \sum_{v\in\mathbb{Z}^d} |\widehat{\varphi}(v)|^2 \left(1+|v|^2\right)^s < \infty\,.$$

Assume that the Fourier series of t satisfies a power-law decay: there exists $c, C > 0$ such that:

$$c\left(1+|v|^2\right)^{-s/2-d/4} \leqslant \widehat{t}(v) \leqslant C\left(1+|v|^2\right)^{-s/2-d/4}\,, \qquad v\in\mathbb{Z}^d\,. \tag{1.13}$$

This condition does not cover C^∞ kernel, including the Gaussian kernel; it is relevant for less regular kernels, that have a power decay in Fourier. This condition is satisfied, for instance, by the Wendland functions [Wendland, 2004, Theorem 10.35], or in dimension $d = 1$ by the kernels corresponding to splines of order s, see [Wahba, 1990] or [Pillaud-Vivien et al., 2018]. The latter can be computed using the polylogarithm or—for special values of s—the Bernoulli polynomials. We have $t \in H^{s'}_{\mathrm{per}}$ if and only if $s' < s$, thus s measures the Sobolev smoothness of t.

We now compute the RKHS norm associated to the kernel k. We first use (1.12) where we choose $\mathcal{P}_u$ to be the uniform law on $\mathcal{U} = [0,1]^d$, and then Parseval formula:

$$\|\varphi\|^2 = \left\langle \varphi, \Sigma^{-1}\varphi\right\rangle_{L^2([0,1]^d)} = \sum_{v\in\mathbb{Z}^d} \widehat{\varphi}(v)\overline{\widehat{\Sigma^{-1}\varphi}(v)}\,. \tag{1.14}$$

Here, $\overline{z}$ denotes the conjugate of a complex number z. Further, using again that we choose $\mathcal{P}_u$ to be the uniform law on $\mathcal{U} = [0,1]^d$, we can rewrite (1.11) as

$$\left(\Sigma\varphi\right)(u') = \int_{[0,1]^d} \mathrm{d}u\, k(u,u')\varphi(u) = \int_{[0,1]^d} \mathrm{d}u\, t(u'-u)\varphi(u)\,.$$

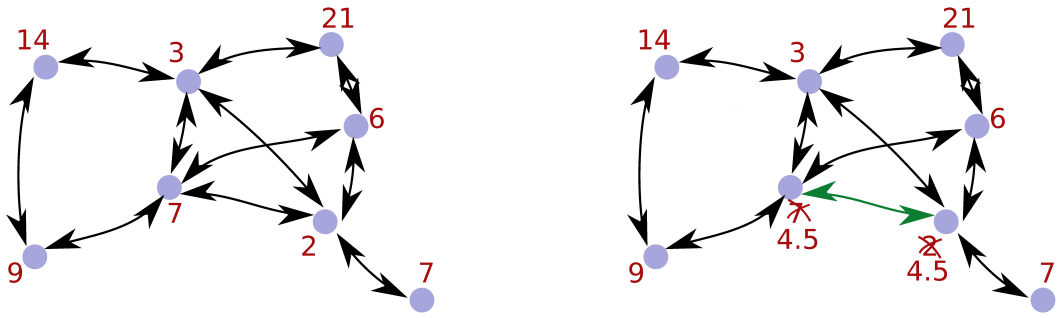
In words, Σ is the convolution by t, thus the multiplication in Fourier space by $\widehat{t}$. Thus, back to (1.14), we obtain

$$\|\varphi\|^2 = \sum_{v\in\mathbb{Z}^d} |\widehat{\varphi}(v)|^2 \widehat{t}(v)^{-1} \asymp \sum_{v\in\mathbb{Z}^d} |\widehat{\varphi}(v)|^2 \left(1+|v|^2\right)^{s/2+d/4} = \|\varphi\|^2_{H^{s/2+d/4}_{\mathrm{per}}}\,.$$

In words, the RKHS associated to a translation invariant kernel satisfying (1.13) is equivalent to the Sobolev space $H^{s/2+d/4}_{\mathrm{per}}$.

1.3. The averaging problem and gossip algorithms

Gossip algorithms are subroutines that diffuse information throughout networks in distributed decentralized algorithms. Here, we first introduce the most pure gossip problem, the averaging problem. We present its naive solution, that we call the *simple gossip algorithm*, and a simplification, called *synchronous simple gossip*. We finish with a discussion on the importance of the averaging problem to distributed computing.



(A) Before the update (B) After the update

FIGURE 1.5. Update of the values x_t in the network, when the green edge is activated.

Averaging problem. Let $G = (\mathcal{V}, \mathcal{E})$ be a finite undirected connected graph with vertex set $\mathcal{V}$ of cardinal m and edge set $\mathcal{E}$ of cardinal N. This graph represents a network of agents $\mathcal{V}$ (computers, sensors, etc) connected through communication links $\mathcal{E}$. We assign to each agent $v \in \mathcal{V}$ a real value $x_0(v)$, called an observation. The goal of the averaging problem, or gossip problem, is to design an iterative procedure allowing each agent to know the average $\bar{x} = \frac{1}{m}\sum_{v\in\mathcal{V}} x_0(v)$ of the initial observations in the network, as quickly as possible, using only local communications.

To pose the problem, we need to define properly the communication model. We assume the time t to be continuous, i.e., t is a non-negative real number. We generate a Poisson point measure $\mathrm{d}N(t,e) = \sum_{n\geqslant 1} \delta_{(T_n,\{v_n,w_n\})}$ on $\mathbb{R}_{\geqslant 0} \times \mathcal{E}$ with intensity measure $\mathrm{d}t \otimes \mu_{\mathcal{E}}$, where $\mathrm{d}t$ is the Lebesgue measure and $\mu_{\mathcal{E}}$ is the counting measure on $\mathcal{E}$. (An introduction to Poisson point measures is given in Section 3.A.1.) The times T_n are the moments where an edge is activated, and $\{v_n, w_n\}$ is the activated edge: the agents v_n and w_n can thus communicate at time T_n. The Poisson point measure assumption implies that edges are activated independently of one another, and independently of the past.

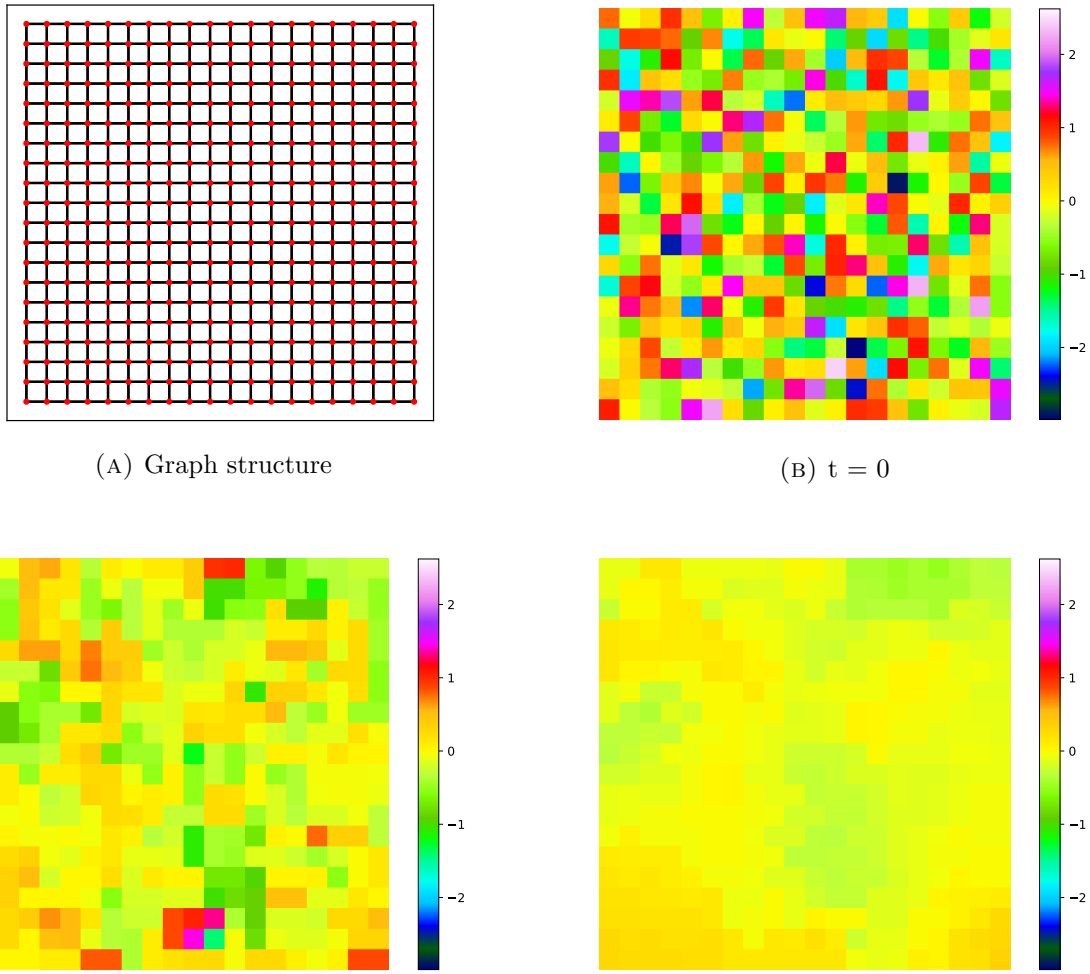
Simple gossip [Boyd et al., 2006]. We call *simple gossip* the naive algorithm for the averaging problem. Each agent $v \in \mathcal{V}$ keeps an estimate $x_t(v)$ of the average $\bar{x}$, initialized at its observation $x_0(v)$. When an edge $\{v, w\}$ is activated, the two agents average their estimates, see Figure 1.5. More precisely, the estimates $x_t = (x_t(v))_{v\in\mathcal{V}}$ remains constant between the activation times $(T_n)_{n\geqslant 1}$. At the activation time T_n, denote $\{v_n, w_n\}$ the associated edge. Then

$$x_{T_n}(v_n) = x_{T_n}(w_n) = \frac{x_{T_n-}(v_n) + x_{T_n-}(w_n)}{2}\,, \qquad x_{T_n}(v) = x_{T_n-}(v)\,, \quad v \neq v_n, w_n\,.$$

Note that in this thesis, all stochastic processes $t \mapsto x_t$ defined on $\mathbb{R}_{\geqslant 0}$ are "càdlàg" by convention, i.e., right continuous with well-defined left-limits x_{t-} (see Definition 3.5 in Appendix 3.A). A realization of the simple gossip procedure on a two-dimensional grid is shown in Figure 1.6.

The questions of the averaging problem are similar to those for the optimization problems of Section 1.2: what is the rate of convergence of the naive method, simple gossip? How does it depend on the graph structure? Can we accelerate simple gossip, i.e., can we modify simple gossip to achieve faster convergence rates?

Synchronous gossip. To a large extent, the difficulty in answering the above questions is due to the randomness in the sampling of an edge at each activation time T_n. The synchronous gossip problem is a simplified setting where at each activation time T_n, all edges are synchronously activated. The activations times still form a Poisson point measure $\mathrm{d}N(t) = \sum_{n\geqslant 1} \delta_{T_n}$ with intensity $\mathrm{d}t$. By opposition, the classical gossip problem is sometimes referred to as *asynchronous* gossip.



(A) Graph structure

(B) t = 0

(C) t = 2

(D) t=8

FIGURE 1.6. One realization of the simple gossip process on a two-dimensional grid. The observations $(x_0(v))_{v\in\mathcal{V}}$ are i.i.d. standard Gaussian random variables. The figures (B)-(D) display the vector $x_t = (x_t(v))_{v\in\mathcal{V}}$ using a color scale, where the display in the image corresponds to the display in the original two-dimensional grid.

In synchronous gossip algorithms, agents average their estimate $x(v)$ of $\bar{x}$ with the estimates of their neighbors. The weights in this local averaging operation is given by a gossip matrix.

Definition 1.3. A *gossip matrix* $W = (W_{v,w})_{v,w\in\mathcal{V}}$ on the graph G is a matrix with entries indexed by the vertices of the graph satisfying the following properties:

- W is non-negative: for all $v, w \in \mathcal{V}$, $W_{v,w} \geqslant 0$.
- W is supported by the graph G: for all distinct vertices v, w such that $W_{v,w} > 0$, then $v \sim w$, that is $\{v, w\}$ must be an edge of G.
- W is stochastic: for all $v \in \mathcal{V}$, $\sum_{w:w\sim v} W_{v,w} = 1$.
- W is symmetric: for all $v, w \in \mathcal{V}$, $W_{v,w} = W_{w,v}$.

If W is a gossip matrix and $x = (x(v))_{v\in\mathcal{V}}$ is a set of values stored by the agents v, the product Wx is interpreted as the computation by each agent v of a weighted average of the values $x(w)$ of its neighbors w in the graph (and of its own value $x(v)$). Note that we do not need the symmetry assumption on W to interpret W as an averaging operation. This assumption is usual in gossip frameworks as it allows one to use the spectral theory for W, on which our analysis of Chapters 4 and 5 relies heavily. It appears, for instance, in the works of Boyd et al. [2006], Cao et al. [2006], Rebeschini and Tatikonda [2017].

In a d-regular graph G (every node has degree $\deg v = d$), a typical gossip matrix is $W = A/d = (\mathbb{1}_{\{\{v,w\}\in\mathcal{E}\}}/d)_{v,w\in\mathcal{V}}$ where A is the adjacency matrix of the graph. The operation Wx is then interpreted as the computation by each agent of the uniform average of the values of its neighbors. However, if the graph is not regular, this operation of local uniform averaging does not correspond to a symmetric gossip matrix; it is not guaranteed that the average throughout the network is invariant through this operation. Instead, if the graph has all vertices of degree bounded by some quantity $d_{\max}$, a natural gossip matrix is

$$W = I + \frac{1}{d_{\max}}(A - D)\,. \tag{1.15}$$

Here, D is the degree matrix, i.e., the diagonal matrix such that $D_{v,v} = \deg v$.

Once a gossip matrix W on the graph G is chosen, the *synchronous simple gossip* iteration is constant between activation times, and at activation time T_n, for all $v \in \mathcal{V}$,

$$x_{T_n}(v) = \sum_{w:w\sim v} W_{v,w} x_{T_n-}(w)\,,$$

or more compactly, with the notation $\tilde{x}_n := x_{T_n}$,

$$\tilde{x}_n = W\tilde{x}_{n-1}\,. \tag{1.16}$$

Thus, synchronous simple gossip mathematically boils down to the power iteration of the matrix W.

Motivations for studying the averaging problem. This paragraph is a non-exhaustive list of related work, biased by the author's interests.

First, from a theoretical perspective, the simple gossip process can be seen as a prototype interacting particle system, among epidemic processes, voter models, token processes, etc., see [Aldous, 2013]. However, the linear structure of the interactions in the simple gossip process makes the analysis simpler than in other interacting particle systems; this property is key for all the results we show in this thesis.

Further, gossip algorithms are often used as primitives in more complex distributed decentralized algorithms [Dimakis et al., 2010]. For instance, in distributed optimization (e.g., [Assran et al., 2020, Nedic et al., 2010, Scaman et al., 2017, Hendrikx et al., 2019]), one seeks to minimize a function $f(x)$ that is a sum of functions $f_v(x)$, where the gradient of $f_v(x)$ is accessible to agent v only:

$$f(x) = \sum_{v\in\mathcal{V}} f_v(x)\,. \tag{1.17}$$

Problems of this form appear for example in least-square regression (Example 1.4), when the data is distributed to the agents: the full empirical risk $f(x)$ is a linear combination of the empirical risk $f_v(x)$ of each agent on its subset of the data. For problems of the form (1.17), the general strategy is to have each agent keep an estimate $x(v)$ of the minimizer of f. Algorithms alternate gradient steps $x(v) \leftarrow x(v) - f_v(x(v))$, where agents minimize their local cost, and gossip steps, where agents average their estimates $x(v)$ to converge to a common minimizer of f.

Similarly, distributed bandit problems also typically alternate arm pulls and gossip steps of the empirical rewards [Szorenyi et al., 2013, Landgren et al., 2016, Korda et al., 2016].

Finally, as the averaging problem is a prototype distributed decentralized problem, we expect progress on it to inspire progress on similar algorithms. For instance, after a long time t of running the simple gossip algorithm, the sign of $x_t(v) - x_{t-1}(v)$ can be used to perform a distributed clustering of the nodes [Becchetti et al., 2018, 2020]. Although this has not been studied yet, it could be interesting to accelerate this procedure. Another example is given in Section 4.6: we adapt our work on the averaging problem to the network localization problem [Barooah and Hespanha, 2008], in which the agents try to estimate some quantity $(x(v))_{v\in\mathcal{V}}$ defined over the graph, from noisy relative measurements over the edges of the graph:

$$\xi(v,w) = x(v) - x(w) + \eta(v,w)\,, \qquad \{v,w\} \in \mathcal{E}\,.$$

1.4. Similarities and differences between convex optimization and the averaging problem

In this thesis, our inspiration to analyze and accelerate gossip algorithms is crucially based on a view of gossip algorithms as (stochastic) gradient descents on a so-called *energy function*. We start by formalizing the parallel in Section 1.4.1. In the following Sections 1.4.2-1.4.4, we bring important nuances to this parallel. In particular, in Section 1.4.3, we explain that the stochastic gradients corresponding to gossip algorithms have a special property, that we call being *noiseless* stochastic gradients. We advocate for a better understanding of noiseless stochastic optimization.

1.4.1. Gossip algorithms are gradient descents on the energy function. We start by showing that asynchronous simple gossip corresponds to a stochastic gradient descent. We continue by showing that its simplification, synchronous simple gossip, corresponds to a deterministic gradient descent.

Asynchronous gossip. For a vector $x = (x(v))_{v\in\mathcal{V}}$, define the energy function

$$f(x) = \frac{1}{2N} \sum_{\{v,w\}\in\mathcal{E}} (x(v) - x(w))^2\,.$$

Recall that N is the cardinal of the edge set $\mathcal{E}$. This is a convex, smooth, finite sum:

$$f(x) = \frac{1}{N} \sum_{\{v,w\}\in\mathcal{E}} f_{\{v,w\}}(x)\,, \qquad\qquad f_{\{v,w\}}(x) = \frac{1}{2}\,(x(v) - x(w))^2\,.$$

Following Example 1.2, we build a stochastic gradient for f by taking the gradient of a uniformly sampled function of the sum. The partial derivatives of $f_{\{v,w\}}$ are

$$\frac{\partial f_{\{v,w\}}}{\partial x(v)} = x(v) - x(w)\,, \qquad \frac{\partial f_{\{v,w\}}}{\partial x(w)} = x(w) - x(v)\,, \qquad \frac{\partial f_{\{v,w\}}}{\partial x(u)} = 0\,, \qquad u \neq v, w\,.$$

Thus, a gradient step on the function $f_{\{v,w\}}$ with step-size $\frac{1}{2}$ is

$$\begin{aligned}
x(v) &\leftarrow x(v) - \frac{1}{2}\frac{\partial f_{\{v,w\}}}{\partial x(v)} = x(v) - \frac{1}{2}(x(v) - x(w)) = \frac{x(v) + x(w)}{2}\,,\\
x(w) &\leftarrow x(w) - \frac{1}{2}\frac{\partial f_{\{v,w\}}}{\partial x(w)} = x(w) - \frac{1}{2}(x(w) - x(v)) = \frac{x(v) + x(w)}{2}\,,\\
x(u) &\leftarrow x(u) - \frac{1}{2}\frac{\partial f_{\{v,w\}}}{\partial x(u)} = x(u)\,, \qquad u \neq v, w\,.
\end{aligned}$$

In words, taking a gradient step on $f_{\{v,w\}}$ with step-size $1/2$ corresponds to averaging on the edge $\{v,w\}$. Thus *stochastic gradient descent on the finite sum $f(x)$, with step-size $1/2$, corresponds to the simple gossip algorithm* (up to a time numbering, to be discussed in the next section).

Remark 1.3. This does not mean that the goal of gossip algorithms is simply to find a minimizer of the energy function f. The energy function f is indeed minimal at $x_* = \bar{x}\mathbb{1}$, where $\mathbb{1}$ denotes the vector with each component equal to 1. But the energy function is minimal at all constant vectors (and only at constant vectors, as the graph G is connected). However, the average of the vector x remains invariant through the stochastic gradient dynamics, so that if the algorithm converges to a minimum of f, it is necessarily $x_* = \bar{x}\mathbb{1}$.

Remark 1.4. The energy function corresponds to the empirical risk of a least-squares regression problem. Indeed, denote $(e_v)_{v\in\mathcal{V}}$ the canonical basis of $\mathbb{R}^{\mathcal{V}}$. Then

$$f(x) = \frac{1}{2N} \sum_{\{v,w\}\in\mathcal{E}} \langle e_v - e_w, x\rangle^2 \ .$$

This is a special case of the least-squares structure (1.6), with $a_{\{v,w\}} = e_v - e_w$ and $b_{\{v,w\}} = 0$. The energy function is a quadratic function with Hessian

$$\Sigma = \frac{1}{N} \sum_{\{v,w\}\in\mathcal{E}} (e_v - e_w)(e_v - e_w)^\top = \frac{1}{N} (D - A) \ ,$$

where again D and A are respectively the degree and adjacency matrices of the graph. In short, the Hessian of the energy function is proportional to the graph Laplacian $\mathcal{L} = D - A$: this explains that the properties of the graph Laplacian are important in showing the convergence of gossip algorithms, see Chapter 2.

Synchronous gossip. Let W be a gossip matrix on the graph G. For a vector $x = (x(v))_{v\in\mathcal{V}}$, define the energy function

$$f_W(x) = \frac{1}{2} \sum_{\{v,w\}\in\mathcal{E}} W_{v,w} \left(x(v) - x(w)\right)^2 = \frac{1}{2} \langle x, (\mathrm{Id} - W)x\rangle \ .$$

A gradient step on the function f_W with step-size 1 is

$$x \leftarrow x - \nabla f_W(x) = x - (\mathrm{Id} - W)x = Wx \, .$$

This is exactly a synchronous gossip step. Thus *gradient descent on the function $f_W(x)$, with step-size 1, corresponds to the synchronous simple gossip algorithm.* This idea can also be found in Scaman et al. [2017] in the more sophisticated case of distributed optimization.

Remark 1.5. In the synchronous case, the Hessian of the energy function is $\mathrm{Id} - W$: this explains that the properties of the gossip matrix W are important in showing the convergence of synchronous gossip algorithms, see Chapter 4.

1.4.2. Time and iteration counter. A first difference between gossip algorithms and gradient descents is that the former are indexed by a continuous time parameter $t \in \mathbb{R}_{\geqslant 0}$ while the latter are indexed by a discrete number of iterations $n \in \mathbb{N}$.

This difference is only superficial when studying naive algorithms. In order to make the parallels of the previous section perfect, one simply has to discretize the continuous-time gossip process at the activation times $(T_n)_{n\geqslant 0}$ (with the convention $T_0 = 0$). The discrete process $(x_{T_n})_{n\geqslant 0}$ is rigorously the stochastic gradient descent of the energy function with step-size $1/2$. Thus the studies in discrete time and continuous-time are exactly the same, up to a time re-scaling which is the Poisson counting process on the positive half-line, of intensity $N = |\mathcal{E}|$ (see Figure 1.7). For large times / iteration numbers, $T_n \approx Nn$ by concentration of sums of random variables, thus the randomness of this re-scaling can be neglected. This discussion justifies that discrete time is often

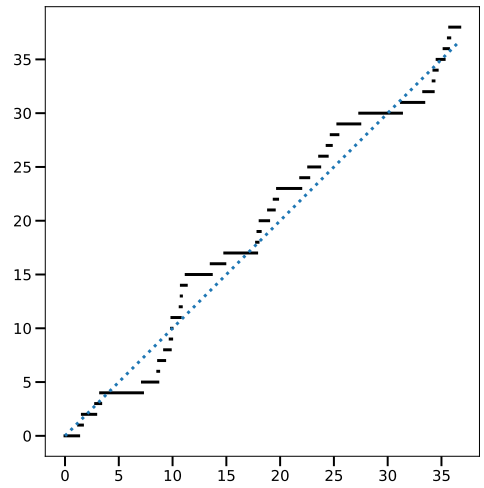


FIGURE 1.7. A realization of the Poisson counting process of intensity 1 (black, plain line), compared with the identity function (blue, dotted line).

used for convenience when studying the simple gossip algorithm (see Chapter 2 or [Boyd et al., 2006], for instance).

On the contrary, the difference between discrete and continuous time is more important when accelerating algorithms [Hendrikx et al., 2019, Loizou et al., 2019]. Indeed, many acceleration algorithms tune their parameters as a function of the number of iterations n (for instance, see Nesterov acceleration in the convex case, see Theorem 3.1.(1)). In the gossip problem, this number n of iteration is the number of past activations in the network, unknown to a particular node. As a consequence, many accelerated optimization algorithms are not implementable in a distributed fashion when translated as gossip algorithms.

Even for accelerated optimization algorithms where the parameters do not depend on the number of iterations (like Nesterov acceleration in the strongly convex case, see Theorem 3.1.(2)), a similar second problem often prevents them to translate into implementable gossip algorithms: when an edge is activated, they often require to perform an update at all nodes of the network, even for some far-away nodes that are not aware of the activation.

Because of these two problems, the design of accelerated methods for gossip algorithms does not follow directly from acceleration of stochastic gradient descents. Some authors simply ignore the problem and study the acceleration in discrete time, see [Cao et al., 2006] for instance. In Chapter 4, we also study the acceleration of gossip algorithms in discrete time, but in the synchronous setting, where the number of past activations is known to each node. Remarkably, Even et al. [2020] designs an implementable acceleration of the asynchronous gossip algorithm in continuous time (under different graph assumptions than Chapter 4). Their work mimics Nesterov's acceleration of coordinate gradient descent [Nesterov, 2012], but using the elapsed time t as an approximation for the number of iterations n. Their proof technique controls the fluctuations due to the randomness of the activation process.

In Chapter 3, we take the parallel between optimization and gossip in the opposite direction: we study how classical optimization algorithms are interpreted in continuous time, with gradient steps taken at random times. In this so-called *continuized* framework, we show that Nesterov acceleration has a close variant benefiting from the best of the continuous and the discrete frameworks: as a continuous process, one can use differential calculus to analyze convergence and obtain analytical expressions for the parameters; but a discretization of the continuized process can be computed

exactly with convergence rates similar to those of Nesterov original acceleration. We show that the discretization has the same structure as Nesterov acceleration, but with random parameters. To sum up, a modeling assumption natural in gossip algorithms brings a new insightful perspective on Nesterov acceleration.

1.4.3. Additive and multiplicative noises in stochastic gradient descents. There are different sources of randomness of stochastic gradients. Archetypal examples are:

- Additive noise from Example 1.1: $g(x,\xi) = \nabla f(x) + \xi$. The gradient is fully observed but perturbed by additive noise.
- Coordinate stochastic gradients from Example 1.3: $g(x,\xi) = m\langle \nabla f(x), e_i\rangle e_i$, $\xi = i \sim \mathrm{Unif}(\{1,\dots,m\})$. Only one component of the gradient is observed, but without noise.

This second source of randomness is often called *multiplicative noise* [Dieuleveut et al., 2017], but this terminology can be misleading: this randomness comes from a *sampling* process, and is completely different from the usual picture of noise as a small additive perturbation.

This terminology of "additive" and "multiplicative" noise is best understood in the least-squares supervised learning setting of Example 1.5:

$$\begin{aligned} f(x) &= \frac{1}{2}\mathbb{E}_{(a,b)\sim\mathcal{P}} \left(b - \langle a, x\rangle\right)^2 , \\ \nabla f_{(a,b)}(x) &= -\left(b - \langle a, x\rangle\right) a \,. \end{aligned}$$

Assume x_* is a minimum of f. Then

$$\begin{aligned} \nabla f_{(a,b)}(x) &= -\left(b - \langle a, x_*\rangle\right) a + \langle a, x - x_*\rangle a \\ &= \nabla f_{(a,b)}(x_*) + a \otimes a (x - x_*) \,. \end{aligned}$$

Thus the stochastic gradient $\nabla f_{(a,b)}(x)$ can be divided into two parts:

- an additive part $\nabla f_{(a,b)}(x_*)$, independent of x and centered as $\mathbb{E}_{(a,b)\sim\mathcal{P}} \nabla f_{(a,b)}(x_*) = \nabla f(x_*) = 0$, and
- a multiplicative part $a \otimes a(x - x_*)$, where the randomness of a acts multiplicatively on the error $x - x_*$. The error $x - x_*$ is shrunk by stochastic gradient descent only in the sampled direction a.

The two sources of randomness have different effects. The additive noise makes stochastic gradient descent deviate from the optimum, even if initialized exactly at the optimum. Variance reduction techniques (decreasing step-sizes, averaging [Polyak and Juditsky, 1992], ...) are thus required to obtain convergence. On the contrary, if the optimal regressor x_* is able to perfectly predict the output from the input, namely $b = \langle a, x_*\rangle$ almost surely, then there is no additive noise. The stochastic gradient $\nabla f_{(a,b)}(x) = a \otimes a(x - x_*)$ is then almost surely 0 at the optimum x_*, and no variance reduction is needed.

> **Definition 1.4** (Noiseless linear model)**.** We say that we are in the *noiseless linear model* if there exists an optimum $x_* \in \mathcal{H}$ such that $b = \langle a, x_*\rangle$ almost surely.

Note that noiseless gradients are still stochastic; indeed, "noiseless" is a shorthand for "without additive noise" and the multiplicative noise remains. A similar meaning for the notions of "noiseless" and "noisy" can be found in [Kearns and Vazirani, 1994] or more recently in [Varre et al., 2021, Bordelon and Pehlevan, 2021, Cui et al., 2021].

Noiseless stochastic gradients are relevant modelings in several situations.

- In *coordinate gradient descent*, the stochastic gradient $g(x,\xi) = m\langle \nabla f(x), e_i\rangle e_i$ is almost surely zero at optimum. Thus coordinate gradient descent is a noiseless problem.

- Consider the minimization of the empirical risk of Example 1.4 in the *overparameterized regime*: the number N of samples is smaller than the number m of parameters describing the model $x \in \mathbb{R}^m$. In this regime, under mild assumptions, there exists a model x_* achieving a zero empirical risk:

$$0 = f(x_*) = \frac{1}{2N} \sum_{k=1}^{N} \left(b_k - \langle a_k, x_* \rangle\right)^2 , \tag{1.18}$$

  i.e., $b_k = \langle a_k, x_* \rangle$ for all $k = 1, \dots, N$. While traditional statistics warn against the risk of overfitting when perfectly fitting the datapoints in an overparameterized regime, recent practical and theoretical work have shown that in some situations, the optimum x_* found by (stochastic) gradient descent generalizes well. This is reviewed by Bartlett et al. [2021]. This motivates understanding the performance of stochastic optimization in this regime. Remarkably, the stochastic gradient descent (1.7) on the empirical risk is noiseless.
- Consider the *function interpolation* problem of Example 1.7:

$$f(\varphi) = \frac{1}{2} \left\|\varphi_* - \varphi\right\|^2_{L^2(\mathcal{P}_u)} = \frac{1}{2} \mathbb{E}_{u \sim \mathcal{P}_u} \left(\varphi_*(u) - \varphi(u)\right)^2 ,$$
$$\nabla f_u(\varphi) = (\varphi(u) - \varphi_*(u)) k(u, .) .$$

  The stochastic gradients $f_u(\varphi)$ where $u \sim \mathcal{P}_u$ are noiseless: they are zero almost surely at the optimum φ_*.

  Supervised learning problems where the output b is completely determined by the input u under $\mathcal{P}_{(u,b)}$ can be seen as function interpolation problems: then there exists a function φ_* such that $b = \varphi_*(u)$. This assumption is relevant for some basic vision or sound recognition tasks, where there is no ambiguity of the output b given the input u, but the rule determining the output from the input can be complex. An example from Jun et al. [2019, Section 6] is the classification of images of cats versus dogs. For typical images, the output is unambiguous; humans indeed achieve a near-zero error. In sound recognition, one could think of the recovery of the melody from a tune, an unambiguous (but tremendously complex!) task. In these problems, the difficulty of learning does not come from the ambiguity of the output given the input (the additive noise), but from the fact that we observe the optimal mapping φ_* at few sampled points only.

  Note that there is a potential confusion between this function interpolation example and the previous example of minimizing the empirical risk in the overparameterized regime. This is largely due to the fact that (1.18) is called the *interpolation regime*. However, note that the interpolation regime appears in any overparameterized learning problem, even under noisy models where b is not determined by a.
- *Gossip algorithms* translate into stochastic gradient descents with noiseless gradients. Indeed, simple gossip does not move from optimum when initialized at optimum $x_* = \bar{x}\mathbb{1}$. It corresponds to a least-square regression with $a_{\{v,w\}} = e_v - e_w$ and $b_{\{v,w\}} = 0$ (see Remark 1.4): the optimal predictor x_* achieves a zero error.

To conclude this section, one should be skeptical about the ability of the additive noise model to explain many behaviors of stochastic gradient descents. At the extreme opposite, the noiseless stochastic model, where there is pure multiplicative noise, is more mathematically challenging, but relevant for many optimization problems and the gossip problem. This common structure explains some bridges in the literature: for instance Even et al. [2020] builds an acceleration of gossip algorithms inspired from the acceleration of coordinate gradient descent [Nesterov, 2012]; Chapter 2 treats function interpolation and gossip algorithms in a common framework.

1.4.4. Local computation contraint in gossip. We end this section with one more constraint making some optimization techniques unfeasible when translated as gossip algorithms. It is a simple remark that turns out to be important in Chapter 4.

In gossip algorithms, the estimator $x_t(v)$ of a node v can be computed only from past information $x_s(v)$, $s < t$ and local information $x_t(w)$ from neighbors w of the node v. This contrasts with centralized implementations of (stochastic) gradient descents where any operation on the full vector x_n is allowed. As a consequence, some techniques in centralized optimization are unimplementable in gossip.

- Preconditioning techniques [Axelsson, 1996] involve the multiplication by a typically dense matrix. In gossip, performing such a multiplication implies that all pairs of nodes (not only neighboring ones) communicate: it is not implementable. As a consequence, gossip problems with a badly conditioned graph Laplacian must remain so.
- The conjugate gradient technique involves computing scalar products between iterates (see, e.g., [Golub and Van Loan, 2013, Section 11.3] or [Nesterov, 2003, Section 1.3.2] for an introduction). When translated into a gossip algorithm (called the parameter-free polynomial iteration in Chapter 4), it requires computing sums over the whole network. This non-local operation makes the algorithm impractical.

1.5. Problem structures, convergence analyses and acceleration

In this section, we describe two possible assumptions that we can make on the (deterministic or stochastic) optimization problems of Section 1.2: Section 1.5.1 introduces the *strong convexity* assumption and Section 1.5.2 introduces the *capacity and source conditions*. Each notion translates into a related assumption in gossip algorithms: a *spectral gap* and a *spectral dimension* assumption respectively. The link between the former pair of assumptions is well known and used in the literature, see [Hendrikx et al., 2019, Loizou et al., 2019] for instance; here, we present it only as an illustration of the parallel of Section 1.4.1 at play in a simple setting. On the contrary, the link between the spectral dimension and the capacity and source condition is novel; it was introduced by Berthier et al. [2020].

For each of these settings, and for both deterministic and stochastic optimization, we give an overview of the rate of convergence of the naive algorithms (deterministic, stochastic gradient descent or the simple gossip algorithm), and of the accelerations. This articulates the rest of the thesis. To sum up, in Figure 1.8, we organize the materials of this thesis in an array, clarifying the type of assumptions for each result.

Finally, in Section 1.5.3, we nuance this binary picture: we illustrate that both assumptions can be simultaneously true; in this case, they explain the behavior of the algorithm on different time scales.

1.5.1. Strong convexity and spectral gap. We start this subsection by introducing the strong convexity assumption, the spectral gap assumption, and their equivalence. We continue by giving the rate of convergence of (deterministic) gradient descent and its acceleration under this assumption. To finish, we describe the similar picture with stochastic gradients.

Definition 1.5 (strong convexity)**.** Let $\mu > 0$. The function f is said to be *μ-strongly convex* if for all $x, y \in \mathcal{H}$,

$$f(y) \geqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2 .$$

The function f is said to be *strongly convex* if it is μ-strongly convex for some $\mu > 0$.

Remark 1.6. If f is strongly convex, f has a unique minimizer x_* on $\mathcal{H}$.

		Strong convexity or spectral gap	Capacity, source condition or spectral dimension
Deterministic	**Naive**		Theorem 4.1.(1)
	Acceleration	Section 3.1-3.3	Most of Chapter 4, Chapter 5
Stochastic	**Naive**	Chapter 2 (Theorems 2.1, 2.6)	Chapter 2 (Theorems 2.2-2.5, 2.7-2.8)
	Acceleration	Section 3.4-3.5	Largely open, some elements in Section 3.4.2

FIGURE 1.8. Array describing the type of assumptions made in the different parts of this thesis. Note that Section 4.5 does not fit this table as it studies acceleration under the joint assumption of a spectral gap and a spectral dimension.

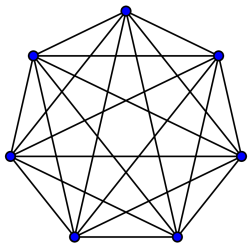


FIGURE 1.9. K_7, the complete graph on 7 vertices.

Remark 1.7. Let f be a quadratic function with Hessian Σ (see Definition 1.1). Then f is μ-strongly convex if and only if $\Sigma \succcurlyeq \mu \operatorname{Id}$.

From Section 1.4.1, synchronous simple gossip can be seen as gradient descent on the energy function $f_W(x) = \frac{1}{2}\langle x, (\operatorname{Id} - W)x\rangle$, where W is the gossip matrix. It is natural to ask whether this function is smooth or strongly convex. This turns out to be related to the following notion.

Definition 1.6 (Spectral gap)**.** Denote $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_m$ the real eigenvalues of the symmetric matrix W. As W is stochastic, $W\mathbb{1} = \mathbb{1}$; we can take $\lambda_1 = 1$, that corresponds to the eigenvector $\mathbb{1} = (1, \ldots, 1)$. According to the Perron-Frobenius theorem, all eigenvalues must be smaller than 1 in magnitude. We define:

(1) the *spectral gap* $\mu = 1 - \lambda_2$ as the distance between the two largest eigenvalues of W,
(2) the *absolute spectral gap* $\tilde{\mu} = \min(1 - \lambda_2, \lambda_m + 1)$ as the difference between the moduli of the two largest eigenvalues of W in magnitude.

Example 1.9 (complete graph)**.** Let K_m denote the complete graph on m vertices, i.e., the graph with vertex set $\mathcal{V} = \{1, \ldots, m\}$ and edge set $\mathcal{E} = \{\{v, w\} \,|\, v, w \in \mathcal{V}, v \neq w\}$ (see Figure 1.9). It is naturally endowed with the gossip matrix $W = \frac{1}{m-1}A$ where

$$A = \begin{pmatrix} 0 & 1 & 1 & \cdots & 1 \\ 1 & 0 & 1 & \cdots & 1 \\ 1 & 1 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & 0 \end{pmatrix} = \mathbb{1}\mathbb{1}^\top - \operatorname{Id}$$

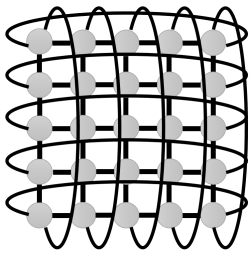


FIGURE 1.10. $\mathbb{T}^2_5$, the two-dimensional torus of side length 5 [Hemis62].

is the adjacency matrix of the graph. A has eigenvalue $m-1$ with multiplicity 1 and eigenvalue -1 with multiplicity $m-1$. Thus W has eigenvalue 1 with multiplicity 1 and eigenvalue $-1/(m-1)$ with multiplicity $m-1$. Thus $\mu = 1+1/(m-1)$ and $\tilde{\mu} = 1-1/(m-1)$.

It is more interesting for us to consider graphs where the spectral gap vanishes in the large graph limit.

Example 1.10 (torus)**.** Let $\mathbb{T}^d_\Lambda$ denote the d-dimensional torus of side length Λ, i.e., the graph with vertex set $\mathcal{V} = (\mathbb{Z}/\Lambda\mathbb{Z})^d$ and edge set $\mathcal{E} = \{\{v,w\} \,|\, v,w \in \mathcal{V}, \|v-w\|_2 = 1\}$ (see Figure 1.10). The torus is naturally endowed with the gossip matrix $W = \frac{1}{2d}A(\mathbb{T}^d_\Lambda) = \mathrm{Id} - \frac{1}{2d}\mathcal{L}(\mathbb{T}^d_\Lambda)$, where $A(\mathbb{T}^d_\Lambda)$ is the adjacency matrix, $\mathcal{L}(\mathbb{T}^d_\Lambda) = 2d\,\mathrm{Id} - A(\mathbb{T}^d_\Lambda)$ is the Laplacian, and $2d$ is the degree of each node.

We compute the eigenvalues of W. The eigenvalues of the Laplacian of the circle $\mathbb{T}^1_\Lambda$ are $2-2\cos\left(\frac{2\pi i}{\Lambda}\right)$, $i \in \mathbb{Z}, -\Lambda/2 < i \leqslant \Lambda/2$ [Chung, 1997, Example 1.5]. As $\mathbb{T}^d_\Lambda$ is the graph Cartesian product $\mathbb{T}^1_\Lambda \times \cdots \times \mathbb{T}^1_\Lambda$ (with d terms), the eigenvalues of the Laplacian of the torus $\mathbb{T}^d_\Lambda$ are the

$$2-2\cos\left(\frac{2\pi i_1}{\Lambda}\right) + \cdots + 2-2\cos\left(\frac{2\pi i_d}{\Lambda}\right), \qquad i_1, \dots i_d \in \mathbb{Z}, \quad -\frac{\Lambda}{2} < i_1, \dots, i_d \leqslant \frac{\Lambda}{2}.$$

Thus, the eigenvalues of $W = \mathrm{Id} - \frac{1}{2d}\mathcal{L}(\mathbb{T}^d_\Lambda)$ are the

$$\frac{1}{d}\left[\cos\left(\frac{2\pi i_1}{\Lambda}\right) + \cdots + \cos\left(\frac{2\pi i_d}{\Lambda}\right)\right], \qquad i_1, \dots i_d \in \mathbb{Z}, \quad -\frac{\Lambda}{2} < i_1, \dots, i_d \leqslant \frac{\Lambda}{2}.$$

As a consequence,

(1) the spectral gap of W is $\mu = \frac{1}{d}\left[1-\cos\left(\frac{2\pi}{\Lambda}\right)\right]$,
(2) if Λ is even, then the absolute spectral gap $\tilde{\mu}$ is 0, and if Λ is odd, then $\tilde{\mu} = \min\left(\mu, 1-\cos\left(\frac{\pi}{\Lambda}\right)\right)$.

The take-home message is that the spectral gap scales like Λ^{-2} (neglecting constants depending on the dimension) when the size of the graph Λ goes to infinity. Depending on the parity of Λ, the absolute spectral gap is either 0, or has the same order of magnitude Λ^{-2}.

We now clarify the relation between spectral gap assumptions and strong convexity.

Proposition 1.1. Define $f_W(x) = \frac{1}{2}\langle x, (\mathrm{Id} - W)x\rangle$. Then:

(1) Let μ be the spectral gap of W. Then f_W is μ-strongly convex and 2-smooth on $x_0 + \mathbb{1}^\perp$, the affine hyperplane orthogonal to the constant vector $\mathbb{1}$ and that passes through x_0.

(2) Let $\tilde{\mu}$ be the absolute spectral gap of W. Then f_W is $\tilde{\mu}$-strongly convex and $(2-\tilde{\mu})$-smooth on $x_0 + \mathbb{1}^\perp$.

PROOF. When restricted to $x_0 + \mathbb{1}^\perp$, the function f is quadratic with Hessian $I - W$ restricted and co-restricted to $\mathbb{1}^\perp$. The eigenvalues of this restricted operator are $1 - \lambda_2, \dots, 1 - \lambda_n$: they are all lower-bounded by $\mu \geqslant \tilde{\mu}$ and upper-bounded by $2 - \tilde{\mu} \leqslant 2$. Using Remarks 1.1 and 1.7, we obtain the stated result. □

Under a convexity assumption only, Theorem 1.1 states that iterates of gradient descent converge at a rate $O(1/n)$. However, under the strong convexity assumption, one can show a faster exponential rate of convergence.

Theorem 1.3 ([Nesterov, 2003, Theorem 2.1.15])**.** Let f be a μ-strongly convex and L-smooth function. Let $(x_n)_{n\in\mathbb{N}}$ be the iterates of gradient descent (1.1) with constant step-sizes $\gamma = \frac{2}{\mu+L}$.

$$\|x_n - x_*\|^2 \leqslant \left(\frac{L-\mu}{L+\mu}\right)^{2n} \|x_0 - x_*\|^2, \tag{1}$$

$$f(x_n) - f(x_*) \leqslant \frac{L}{2}\left(\frac{L-\mu}{L+\mu}\right)^{2n} \|x_0 - x_*\|^2. \tag{2}$$

In words, under the strong convexity assumption, gradient descent converges exponentially fast in a typical time of the order of the *condition number* L/μ of the function. This is also called a *linear* rate of convergence because the number of significant digits in the solution grows linearly in the iteration count. Slower rates of convergence are called *sublinear*, but in this work we prefer to be more specific and speak of *polynomial* rates of convergence for rates of the form $n^{-\alpha}$, $\alpha > 0$.

Using the interpretation of the synchronous gossip algorithm as a gradient descent on the energy function, we get a similar result.

Corollary 1.1. Let $(x_n)_{n\in\mathbb{N}}$ be the iterates of the synchronous gossip iteration with gossip matrix W and activation times $(T_n)_{n\in\mathbb{N}}$. Let $\tilde{\mu}$ be the absolute spectral gap of W. Then

$$\sum_{v\in\mathcal{V}} (x_{T_n}(v) - \bar{x})^2 \leqslant (1-\tilde{\mu})^{2n} \sum_{v\in\mathcal{V}} (x_0(v) - \bar{x})^2, \tag{1}$$

$$\frac{1}{2}\sum_{\{v,w\}\in\mathcal{E}} W_{v,w}(x_{T_n}(v) - x_{T_n}(w))^2 \leqslant (1-\tilde{\mu})^{2n} \sum_{v\in\mathcal{V}} (x_0(v) - \bar{x})^2. \tag{2}$$

In words, synchronous simple gossip converges exponentially fast in a typical time of the order of the inverse absolute spectral gap $1/\tilde{\mu}$.

PROOF. From Section 1.4.1, $(x_{T_n})_{n\in\mathbb{N}}$ are the iterates of gradient descent on the energy function $f_W(x) = \frac{1}{2}\langle x, (\mathrm{Id} - W)x\rangle$ with step-size $\gamma = 1$. As W is a gossip matrix, all iterates have the same average as x_0. Thus all iterates x_{T_n} belong to $x_0 + \mathbb{1}^\perp$, a subspace where f is $\tilde{\mu}$-strongly convex and $(2-\tilde{\mu})$-smooth by Proposition 1.1. Theorem 1.3 applies as the step-sizes are small enough: $1 = \gamma = \frac{2}{\tilde{\mu}+(2-\tilde{\mu})}$. Thus the iterates converge exponentially to the unique minimizer of f in $x_0 + \mathbb{1}^\perp$ which is $x_* = \bar{x}\mathbb{1}$. More precisely,

$$\sum_{v\in\mathcal{V}} (x_{T_n}(v) - \bar{x})^2 = \|x_{T_n} - x_*\|^2 \leqslant \left(\frac{(2-\tilde{\mu}) - \tilde{\mu}}{(2-\tilde{\mu}) + \tilde{\mu}}\right)^{2n} \|x_0 - x_*\|^2 = (1-\tilde{\mu})^{2n} \sum_{v\in\mathcal{V}} (x_0(v) - \bar{x})^2 \ .$$

This proves the first bound. The second bound is proved similarly using

$$\frac{1}{2}\sum_{\{v,w\}\in\mathcal{E}} W_{v,w}(x_{T_n}(v)-x_{T_n}(w))^2 = f_W(x_{T_n}) = f_W(x_{T_n}) - f_W(x_*)\,.$$

□

Remark 1.8. It is possible to show that the result of Corollary 1.1 is false if one replaces the absolute spectral gap $\tilde{\mu}$ by the spectral gap μ. And indeed, the proof mechanism above fails in this case: under a spectral gap assumption only, Proposition 1.1 states only that f_W is 2-smooth, and μ-strongly convex. The step-size $\gamma = 1$ corresponding to the gossip algorithm is too large for Theorem 1.3 to apply.

Example 1.11 (torus, continued)**.** We now interpret the result of Corollary 1.1 in the special case of the torus $\mathbb{T}^d_\Lambda$ from Example 1.10.

If Λ is even, then the absolute spectral gap $\tilde{\mu}$ is 0 and the result of Corollary 1.1 is vacuous. And indeed, the iterates of simple synchronous gossip may not converge. In this case, the graph $\mathbb{T}^d_\Lambda$ is bipartite: its vertex set $\mathcal{V}$ can be divided into two disconnected subsets $\mathcal{V}_1$ and $\mathcal{V}_2$. Moreover, because of our choice of gossip matrix, a node $v \in \mathcal{V}_1$ only averages values from nodes in $\mathcal{V}_2$ (it does not put any weight on itself), and conversely. As a consequence, the observations $x_0(v)$ from nodes $v \in \mathcal{V}_2$ are averaged, and similarly for the observations in $\mathcal{V}_1$, but these partial averages never mix. An algorithmic solution is to change the gossip matrix so that nodes average the values of their neighbor with their own running value: $W \leftarrow \frac{1}{2}(W + \mathrm{Id})$. Through this operation, the spectral gap is divided by two, but the absolute spectral gap of the new gossip matrix is automatically equal to its spectral gap.

In the case where Λ is odd, $\Lambda \to \infty$, then the absolute spectral gap $\tilde{\mu}$ scales like Λ^{-2}. Corollary 1.1 states that synchronous simple gossip converges in a typical time Λ^2. In comparison, any gossip algorithm requires information to diffuse from one end of the graph to the other one; thus any algorithm requires a number of iterations equal to the diameter of the graph, here of the order of Λ. This is an illustration of the suboptimality of the simple gossip algorithm: while we could hope for a convergence is $O(\Lambda)$ steps, simple gossip requires $O(\Lambda^2)$ iterations. This is also called a *diffusive* rate of convergence [Rebeschini and Tatikonda, 2017]: after n iterations, while information could have reached nodes at distance n, it is typically spread on nodes at distance $\sqrt{n}$; this is similar to heat diffusion, or equivalently, to the mixing time of a random walk on the graph [Boyd et al., 2006]. In Chapter 5, we push this comparison further by showing that on large grids, the simple gossip algorithm scales like a heat diffusion.

Acceleration. Motivated by the above example, we now turn to the question of acceleration. Many methods exist to accelerate in the strongly convex case with deterministic gradients; of interest to us are Nesterov acceleration (3.1)-(3.3), Polyak's heavy ball method (1.2) and the Chebyshev acceleration (reviewed in [d'Aspremont et al., 2021, Chapter 2]). All of these accelerations achieve exponential convergence in a typical time $O(L/\mu)$, that is, the square root of the condition number. For Nesterov acceleration, the accelerated convergence rate was proved for all L-smooth and μ-strongly convex functions; for Polyak's heavy ball method and Chebyshev acceleration, the known proofs of the accelerated global convergence rate only apply to quadratic functions with these properties.

When the problem is ill-conditioned, the improvement from $O(L/\mu)$ to $O(\sqrt{L/\mu})$ can be significant. When transposed to gossip algorithms on the torus $\mathbb{T}^d_\Lambda$, these algorithms accelerate from the diffusive typical time $O(\Lambda^2)$ to the optimal order of magnitude $O(\Lambda)$. In Figure 4.3, the performance of the shift-register method is typical of these accelerations based on the spectral gap.

Stochastic case. Above, we presented rates and accelerations for optimization with deterministic gradients, or synchronous gossip. We now mention extensions to the stochastic case, or asynchronous gossip.

The analysis of stochastic gradient descent under strong convexity is well-known, see [Bottou et al., 2018] for instance. In Theorems 2.1 and 2.6, we derive again those results in the least-squares setting, only for the sake of comparison with our contributions. The take-home message is that, in the noiseless case, fixed step-size stochastic gradient descent converges exponentially fast in a typical time $O(R_0/\mu)$, where R_0 is the maximal square norm of the features a and μ is the strong convexity parameter. The parallel result for gossip algorithms was given by Boyd et al. [2006]. In the noisy case, fixed step-size stochastic gradient descent does not converge anymore. However, it reaches exponentially fast a region with a low sub-optimality gap, that is proportional to the optimal risk $f(x_*)$ and the step-size γ.

Accelerating stochastic gradient descent is an active research topic. In the least-square case, Jain et al. [2018] show that in general it is not possible to design an acceleration converging in time $O(\sqrt{R_0/\mu})$. The authors design an acceleration whose improvement over stochastic gradient descent depends on a new quantity, the *statistical condition number*. In Section 3.4.2, we design an acceleration achieving a similar performance as theirs, using a method closely similar to Nesterov acceleration.

1.5.2. Source, capacity conditions and spectral dimension. We start this subsection by introducing the source and capacity condition, the spectral dimension assumption, and their equivalence. We then give an overview of convergence rates and accelerations of gradient descents and gossip algorithms under these assumptions.

The definitions of this section are less stable from one work to another. Even in this thesis, it is convenient in Chapter 4 to take a definition of the spectral dimension different from the one of this section. However, the heuristic picture remains the same; we give here the definitions that we use in Chapter 2.

We now specialize to the least-squares supervised learning problem of Example 1.5:

$$f(x) = \frac{1}{2}\mathbb{E}_{(a,b)\sim\mathcal{P}}(b - \langle a, x\rangle)^2 \,.$$

We denote x_0 the initialization of the algorithms and we assume that there exists a minimizer $x_* \in \mathcal{H}$ of $f(x)$. Recall that we then have the formula

$$f(x) = \frac{1}{2}\left\langle x - x_*, \Sigma(x - x_*)\right\rangle + f(x_*)\,,$$

where $\Sigma = \mathbb{E}_{(a,b)\sim\mathcal{P}} a \otimes a$ is the covariance operator of the features a. We do not assume that the linear operator Σ is invertible. Throughout this thesis, we use the following convenient notation: if α is a positive real and x a vector,

$$\left\|\Sigma^{-\alpha/2}x\right\|^2 = \langle x, \Sigma^{-\alpha}x\rangle := \inf\left\{\|x'\|^2 \,\middle|\, x' \text{ such that } x = \Sigma^{\alpha/2}x'\right\}\,,$$

with the convention that it is equal to ∞ when $x \notin \Sigma^{\alpha/2}(\mathcal{H})$.

> **Definition 1.7** (source condition)**.** A source condition is the assumption that there exists a positive number α_1 such that $x_* - x_0 \in \Sigma^{\alpha_1/2}(\mathcal{H})$, i.e., $\|\Sigma^{-\alpha_1/2}(x_* - x_0)\|^2 < \infty$. The exponent α_1 is called the regularity of the optimum.

Note that in this definition, the source condition does not only depend on the objective function f, but also on the initialization x_0.

Definition 1.8 (capacity condition)**.** A capacity condition is the assumption that there exists a positive number α_2 such that $a \in \Sigma^{\alpha_2/2}(\mathcal{H})$ almost surely (a.s.), and that there exists a constant R_{α_2} such that $\|\Sigma^{-\alpha_2/2}a\|^2 \leqslant R_{\alpha_2}$ a.s. The exponent α_2 is called the regularity of the features.

The source condition is classical in the non-parametric kernel literature [Caponnetto and De Vito, 2007, Yao et al., 2007]. The capacity condition is assumed in this form by Pillaud-Vivien et al. [2018]. It implies that

$$\mathrm{Tr}(\Sigma^{1-\alpha_2}) = \mathbb{E}[\mathrm{Tr}(aa^\top \Sigma^{-\alpha_2})] = \mathbb{E}[a^\top \Sigma^{-\alpha_2} a] \leqslant R_{\alpha_2}\,.$$

This last condition is sometimes stated under the form of a given decay of the eigenvalues of Σ; it is related to the *effective dimension* of the problem [Caponnetto and De Vito, 2007].

Example 1.12 (Sobolev spaces, continued)**.** Consider Example 1.7: we interpolate a function φ_* on $\mathcal{U}$ from the observation of its value at random points using a kernel k defining a RKHS $\mathcal{H}$.

We set ourselves in the Sobolev case of Example 1.8: we choose $\mathcal{U} = [0,1]^d$, the kernel $k(u,u') = t(u-u')$ is translation-invariant and satisfies the Fourier decay (1.13), so that the RKHS $\mathcal{H}$ is equivalent to the Sobolev space $H^{s/2+d/4}_{\mathrm{per}}$. We assume that $\varphi_* \in \mathcal{H} = H^{s/2+d/4}_{\mathrm{per}}$: this is the so-called *attainable case*.

Assume that the input u is uniform in $[0,1]^d$. From (1.11), the covariance operator is the convolution by t: it is thus diagonalized in Fourier space where it is multiplication by $\widehat{t}$. Its eigenvalues $\widehat{t}(v)$ converge to 0 as $|v| \to \infty$, thus no strong convexity condition can hold.

Instead, we show that source and capacity conditions hold for the problem of minimizing the risk

$$f(\varphi) = \frac{1}{2}\mathbb{E}_{(u,b)\sim\mathcal{P}_{(u,b)}}\left(b - \varphi(u)\right)^2\,,$$

and are related to the Sobolev smoothness of the functions φ_* and t. The computations are similar to those in Example 1.8.

(1) (source condition) Choose the initialization $\varphi_0 = 0$. Then from (1.12) and Parseval identity,

$$\|\Sigma^{-\alpha_1/2}(\varphi_* - \varphi_0)\|^2 = \left\langle \varphi_*, \Sigma^{-\alpha_1-1}\varphi_* \right\rangle^2_{L^2([0,1]^d)} = \sum_{v\in\mathbb{Z}^d} \widehat{\varphi_*}(v)\overline{\widehat{\Sigma^{-\alpha_1-1}\varphi_*}(v)}\,.$$

Further, Σ is the convolution by t, thus the multiplication in Fourier space by $\widehat{t}$. Thus

$$\begin{aligned}
\|\Sigma^{-\alpha_1/2}(\varphi_* - \varphi_0)\|^2 &= \sum_{v\in\mathbb{Z}^d} |\widehat{\varphi}_*(v)|^2 \widehat{t}(v)^{-1-\alpha_1} \\
&\asymp \sum_{v\in\mathbb{Z}^d} |\widehat{\varphi}_*(v)|^2 \left(1+|v|^2\right)^{(\alpha_1+1)(s/2+d/4)} \\
&= \|\varphi_*\|^2_{H^{(\alpha_1+1)(s/2+d/4)}_{\mathrm{per}}}\,.
\end{aligned}$$

Thus, if φ_* is in the Sobolev space H^r_{per}, then the function interpolation problem satisfies the source condition with

$$\alpha_1 = \frac{2r}{s+d/2} - 1\,.$$

(2) (capacity condition) Similar computations give:

$$\begin{aligned}
\|\Sigma^{-\alpha_2/2}k(u,.)\|^2 &= \left\langle t(u-.), \Sigma^{-\alpha_2-1}t(u-.)\right\rangle_{L^2([0,1]^d)} \\
&= \sum_{v\in\mathbb{Z}^d} \widehat{t(u-.)}(v)\overline{\widehat{\Sigma^{-\alpha_2-1}t(u-.)}(v)} \\
&= \sum_{v\in\mathbb{Z}^d} |\widehat{t(u-.)}(v)|^2 \widehat{t}(v)^{-\alpha_2-1} \\
&= \sum_{v\in\mathbb{Z}^d} \widehat{t}(v)^{-\alpha_2+1} \\
&= \sum_{v\in\mathbb{Z}^d} \left(1+|v|^2\right)^{(s/2+d/4)(\alpha_2-1)}\,.
\end{aligned}$$

Thus the function interpolation problem satisfies the capacity condition for all α_2 such that

$$\alpha_2 < 1 - \frac{d}{s+d/2}\,.$$

From Section 1.4.1, simple gossip can be seen as a stochastic gradient descent on the energy function $f(x) = \frac{1}{2N}\sum_{\{v,w\}\in\mathcal{E}} \langle e_v - e_w, x\rangle^2 = \frac{1}{2N}\langle x, \mathcal{L}x\rangle$, where $\mathcal{L}$ is the Laplacian of the graph. It is natural to ask what are the capacity and source condition satisfied by this stochastic optimization problem. It turns out to be related to the following notion of spectral dimension.

Definition 1.9 (spectral dimension)**.** Let $v \in \mathcal{V}$ be a vertex. As $\mathcal{L}$ is a bounded positive semi-definite operator, there exists a unique measure σ_v, called the *spectral measure of* $\mathcal{L}$ *at* v, such that for all continuous real functions f,

$$\langle e_v, f(\mathcal{L})e_v\rangle = \int \mathrm{d}\sigma_v(\lambda) f(\lambda)\,.$$

If $0 = \lambda_1 < \lambda_2 \leqslant \ldots \leqslant \lambda_m$ are the eigenvalues of $\mathcal{L}$ and $u_1 = \mathbb{1}, u_2, \ldots, u_m$ are the corresponding normalized eigenvectors, then

$$\sigma_v(\mathrm{d}\lambda) = \sum_{i=1}^m (u_i(v))^2 \delta_{\lambda_i}(\mathrm{d}\lambda)\,.$$

We say that G is of spectral dimension $d \geqslant 0$ with constant $V > 0$ if

$$\forall v \in \mathcal{V}\,, \quad \forall E \in (0,\infty)\,, \quad \sigma_v((0,E]) \leqslant V^{-1}E^{d/2}\,.$$

A typical example motivating this definition is the following.

Example 1.13 (torus, continued)**.** Let $\mathbb{T}^d_\Lambda$ denote the d-dimensional torus of side length Λ, whose spectral gap is studied in Example 1.10. The torus $\mathbb{T}^d_\Lambda$ is of spectral dimension d with some constant $V(d)$ that depends on the dimension d but not on the side length Λ.

PROOF OF EXAMPLE 1.13. The graph $\mathbb{T}_\Lambda^d$ is invariant by translation, thus the spectral measure σ_v is the same for all vertices $v \in \mathcal{V}$. Thus

$$|\mathcal{V}|\sigma_v(\mathrm{d}\lambda) = \sum_{w\in\mathcal{V}} \sigma_w(\mathrm{d}\lambda) = \sum_{w\in\mathcal{V}}\sum_{i=1}^m u_i(w)^2\delta_{\lambda_i} = \sum_{i=1}^m \left(\sum_{w\in\mathcal{V}} u_i(w)^2\right)\delta_{\lambda_i} = \sum_{i=1}^m \delta_{\lambda_i}\,.$$

Thus

$$\sigma_v((0,E]) = \frac{1}{\Lambda^d}\left|\{1 < i \leqslant m | \lambda_i \leqslant E\}\right|\,.$$

From Example 1.10, the eigenvalues of the Laplacian of the torus $\mathbb{T}_\Lambda^d$ are the

$$2-2\cos\left(\frac{2\pi i_1}{\Lambda}\right) + \cdots + 2 - 2\cos\left(\frac{2\pi i_d}{\Lambda}\right)\,, \qquad i_1,\dots i_d \in \mathbb{Z}, \quad -\frac{\Lambda}{2} < i_1,\dots,i_d \leqslant \frac{\Lambda}{2}\,.$$

For $y \in [-\pi,\pi]$, $1-\cos(y) \geqslant \frac{2}{\pi^2}y^2$. Thus

$$\begin{aligned} 2-2\cos\left(\frac{2\pi i_1}{\Lambda}\right) + \cdots + 2 - 2\cos\left(\frac{2\pi i_d}{\Lambda}\right) &\leqslant E \\ \Rightarrow \frac{4}{\pi^2}\left[\left(\frac{2\pi i_1}{\Lambda}\right)^2 + \cdots + \left(\frac{2\pi i_d}{\Lambda}\right)^2\right] &\leqslant E \\ \Leftrightarrow i_1^2 + \cdots + i_d^2 &\leqslant \frac{E\Lambda^2}{16}\,. \end{aligned}$$

We need to count the number of integer points in the Euclidean ball centered at 0 and of radius $\sqrt{E}\Lambda/4$ in $\mathbb{R}^d$. This problem is famously known as Gauss circle problem. For our purposes, a crude estimate suffices: there exists a constant $C(d)$, depending only on the dimension d, such that for all radius R, the number of integer points in the ball of radius R is smaller than $1 + C(d)R^d$. This leads to the final estimate

$$\begin{aligned} \sigma_v((0,E]) &= \frac{1}{\Lambda^d}\left|\left\{(i_1,\dots,i_d) \in \left(\mathbb{Z}\cap\left(-\frac{\Lambda}{2},\frac{\Lambda}{2}\right]\right)^d \setminus \{0\} \text{ such that } \right.\right. \\ &\qquad\qquad \left.\left. 2-2\cos\left(\frac{2\pi i_1}{\Lambda}\right) + \cdots + 2 - 2\cos\left(\frac{2\pi i_d}{\Lambda}\right) \leqslant E\right\}\right| \\ &\leqslant \frac{1}{\Lambda^d}\left|\left\{(i_1,\dots,i_d) \in \mathbb{Z}^d\setminus\{0\} \,\middle|\, i_1^2 + \cdots + i_d^2 \leqslant \frac{E\Lambda^2}{16}\right\}\right| \\ &\leqslant \frac{1}{\Lambda^d}C(d)\left(\frac{E\Lambda^2}{16}\right)^{d/2} = \frac{C(d)}{4^d}E^{d/2}\,. \end{aligned}$$

This proves the example with $V(d) = 4^d/C(d)$. □

Similar spectral dimension results were proved for supercritical percolation bonds in [Mathieu and Remy, 2004] and for the random geometric graphs in [Avrachenkov et al., 2019].

We now conclude our digression by showing that a spectral dimension assumption on the network graph G of a gossip problem implies source and capacity conditions for the associated stochastic least-squares problem.

Proposition 1.2. Assume that the graph G is of spectral dimension d with constant V. Let $\delta_{\max}$ denote the maximal degree of the nodes in the graph. Assume further that the initial observation $x_0 : \mathcal{V} \to \mathbb{R}$ is the indicator of some distinguished vertex $v_* \in \mathcal{V}$: $x_0(v_*) = 1$ and

$x_0(v) = 0$ if $v \neq v_*$. Recall from Section 1.4.1 that the gossip problem can be seen as the least-square problem

$$f(x) = \frac{1}{2}\mathbb{E}_{(a,b)\sim\mathcal{P}}(b - \langle a, x\rangle)^2$$

with $b = 0$, $a = e_v - e_w$ where the edge $\{v, w\}$ is uniformly sampled, and $x \in x_0 + \mathbb{1}^\perp$, where $x_* = \bar{x}\mathbb{1}$ is the unique minimizer.

This problem, initialized from x_0, satisfies

(1) (source condition) for any $\alpha_1 < d/2$, the optimum has regularity α_1, and

$$\|\Sigma^{-\alpha_1/2}(x_* - x_0)\|^2 \leqslant N^{\alpha_1} V^{-1} \delta_{\max}^{d/2-\alpha_1} \frac{d}{d - 2\alpha_1}\,,$$

where again N denotes the number of edges in the graph, and

(2) (capacity condition) for any $\alpha_2 < d/2$, the features have regularity α_2 with associated constant

$$R_{\alpha_2} = 2N^{\alpha_2} V^{-1} \delta_{\max}^{d/2-\alpha_2} \frac{d}{d - 2\alpha_2}\,.$$

PROOF. (1) Let $0 = \lambda_1 < \lambda_2 \leqslant \ldots \leqslant \lambda_m$ be the eigenvalues of $\mathcal{L}$ and $u_1 = \mathbb{1}, u_2, \ldots, u_m$ be the corresponding normalized eigenvectors. Then

$$\|\Sigma^{-\alpha_1/2}\,(x_* - x_0)\,\|^2 = N^{\alpha_1}\,\langle x_* - x_0, \mathcal{L}^{-\alpha_1}(x_* - x_0)\rangle = N^{\alpha_1} \sum_{i=2}^{m} \lambda_i^{-\alpha_1}\,\langle x_* - x_0, u_i\rangle^2\,.$$

First, as x_* is a constant vector, $\langle x_*, u_i\rangle$ is zero for all $i \geqslant 2$. Second, $x_0 = e_{v_\star}$. Thus

$$\begin{aligned}
\|\Sigma^{-\alpha_1/2}(x_* - x_0)\|^2 &= N^{\alpha_1} \sum_{i=2}^{m} \lambda_i^{-\alpha_1} u_i(v_\star)^2 \\
&= N^{\alpha_1} \int_{(0,\infty)} \mathrm{d}\sigma_{v_\star}(\lambda)\,\lambda^{-\alpha_1} \\
&= N^{\alpha_1} \int_{(0,\infty)} \mathrm{d}\sigma_{v_\star}(\lambda) \int_0^\infty \mathrm{d}s\,\mathbb{1}_{\{s\leqslant\lambda^{-\alpha_1}\}} \\
&= N^{\alpha_1} \int_0^\infty \mathrm{d}s \int_{(0,\infty)} \mathrm{d}\sigma_{v_\star}(\lambda)\,\mathbb{1}_{\{\lambda\leqslant s^{-1/\alpha_1}\}} \\
&= N^{\alpha_1} \int_0^\infty \mathrm{d}s\,\sigma_{v_\star}((0, s^{-1/\alpha_1}])\,.
\end{aligned}$$

The graph G is of spectral dimension d with constant V, thus $\sigma_{v_\star}((0, s^{-1/\alpha_1}]) \leqslant V^{-1} s^{-\frac{d}{2\alpha_1}}$. However, if $s < \delta_{\max}^{-\alpha_1}$, it is better to use a more naive bound. As all eigenvalues of $\mathcal{L}$ are smaller or equal than $\delta_{\max}$, $\sigma_{v_\star}((0, s^{-1/\alpha_1}]) \leqslant \sigma_{v_\star}((0, \delta_{\max}]) \leqslant V^{-1}\delta_{\max}^{d/2}$. Then

$$\|\Sigma^{-\alpha_1/2}\,(x_* - x_0)\,\|^2 \leqslant N^{\alpha_1}\left[\int_0^{\delta_{\max}^{-\alpha_1}} \mathrm{d}s\,V^{-1}\delta_{\max}^{d/2} + \int_{\delta_{\max}^{-\alpha_1}}^\infty \mathrm{d}s\,V^{-1}s^{-\frac{d}{2\alpha_1}}\right]$$

Note that this last integral is finite because we take $\alpha_1 < d/2$. We obtain

$$\|\Sigma^{-\alpha_1/2}\,(x_* - x_0)\,\|^2 = N^{\alpha_1} V^{-1} \delta_{\max}^{d/2-\alpha_1} \frac{d}{d - 2\alpha_1}\,.$$

(2) Let $\{v, w\} \in \mathcal{E}$. As $\|\Sigma^{-\alpha_2/2}.\|$ is a norm, by the triangle inequality,

$$\|\Sigma^{-\alpha_2/2}(e_v - e_w)\|^2 = \|\Sigma^{-\alpha_2/2}\,[(x_* - e_w) - (x_* - e_v)]\,\|^2$$

$$\leqslant \left(\|\Sigma^{-\alpha_2/2}(x_* - e_w)\| + \|\Sigma^{-\alpha_2/2}(x_* - e_v)\|\right)^2$$
$$\leqslant 2\left(\|\Sigma^{-\alpha_2/2}(x_* - e_w)\|^2 + \|\Sigma^{-\alpha_2/2}(x_* - e_v)\|^2\right).$$

We bound the two quantities as above. We obtain

$$R_{\alpha_2} = \sup_{\{v,w\}\in\mathcal{E}} \|\Sigma^{-\alpha_2/2}(e_v - e_w)\|^2 \leqslant 2N^{\alpha_2} V^{-1} \delta_{\max}^{d/2-\alpha_2} \frac{d}{d-2\alpha_2}.$$

□

Deterministic gradient descents and synchronous gossip. Gradient descent on a quadratic function with deterministic gradients is easily studied as it boils down to the power iteration of a matrix. We have $\|x_n - x_*\|^2 = O(n^{-\alpha_1})$ and $f(x_n) - f(x_*) = O(n^{-\alpha_1-1})$, where α_1 is the regularity of the optimum. For synchronous simple gossip, this translates in a rate of convergence $\|x_t - x_*\|^2 = O(t^{-d/2})$ (up to potential log factors, due to the fact that the source condition is not exactly $d/2$ in Proposition 1.2). This rate is another effect of the diffusivity phenomena, already seen in Example 1.11. If the initialization is the indicator $x_0 = e_{v_*}$ of some distinguished vertex, roughly speaking, the mass of the 1 diffuses on all vertices in a ball of radius $O(\sqrt{t})$ around v_* in a time t. In a graph of spectral dimension d, this ball contains $\Theta(\sqrt{t}^d)$ vertices, all getting a similar mass $\Theta(\sqrt{t}^{-d})$. This results in an error $\|x_t\|^2 = \Theta(t^{d/2}(\sqrt{t}^{-d})^2) = \Theta(t^{-d/2})$. (We consider here a very large graph to avoid border effect; then $\bar{x} \approx 0$.)

This naturally raises the question of acceleration under a source condition, for instance to solve the diffusivity problem in gossip. The acceleration of the optimization of quadratic functions is solved by the conjugate gradient algorithm (see, e.g., [Golub and Van Loan, 2013, Section 11.3] or [Nesterov, 2003, Section 1.3.2] for an introduction), that finds the best linear combination of the past gradient descent iterates, in a online fashion that adapts to the spectrum of the Hessian. In particular, it adapts to a potential source condition, even if unknown. However, as we have seen in Section 1.4.4, the conjugate gradient algorithm translates into a theoretical gossip algorithm which is not implementable as it involves non-local computations.

On quadratic functions, many accelerated methods—including the conjugate gradient methods—can be derived and analyzed through the polynomials in the Hessian that they compute. Some of them, called *inner-product free*, require only local computations when translated as gossip algorithms: they can be used to build accelerated gossip algorithms. The polynomials are chosen to satisfy some minimization property. For instance, Nemirovsky [1991, 1992] proposed a variant of the Chebyshev acceleration that achieves acceleration under a source condition by considering some minimax families of polynomials. In Chapter 4, we use some families of orthogonal polynomials—that minimize some norm—to accelerate gossip algorithms under a spectral dimension assumption. If d is the spectral dimension of the graph, we achieve rates $\|x_t - x_*\|^2 = \Theta(t^{-d})$. In the above discussion, this means that we overcome the diffusivity problem: information reaches distance $\Theta(t)$ in the graph at time t.

Stochastic gradient descents and asynchronous gossip. It is the subject of Chapter 2 to study of the rates of stochastic gradient descent in this setting. A capacity condition on the stochastic gradients is required, along with the source condition: it ensures that the source condition is maintained from one iterate to the next. We prove that stochastic gradient descent converges at the rates $\|x_n - x_*\|^2 = O(n^{-\min(\alpha_1,\alpha_2)})$ and $f(x_n) - f(x_*) = O(n^{-\min(\alpha_1,\alpha_2)-1})$, where α_1 and α_2 denote the regularities of the optimum and of the features respectively. For gossip algorithms, this translates in the rate $\|x_t - x_*\|^2 = O(t^{-d/2})$ (up to log factors), as in the deterministic case.

We now turn to the acceleration of stochastic gradient descent under capacity and source conditions. The case with additive noise has been well studied (see for instance [Rosasco and

Villa, 2015, Dieuleveut and Bach, 2016, Pillaud-Vivien et al., 2018]), with optimality reached by averaged stochastic gradient descent in many cases. To the best of our knowledge, acceleration in the noiseless case has not been studied; we can expect faster rates to be possible. In Section 3.4.2, we provide an acceleration in the noiseless case when the regularities $\alpha_1 = \alpha_2 = 1$. However, we leave the question open for other regularities α_1, α_2.

1.5.3. Co-existence of both settings. For high-dimensional least-squares problems, or for gossip algorithms on large graphs, both a strong convexity condition and source and capacity conditions can simultaneously hold. Of course, the faster exponential rates implied by the strong convexity are asymptotically tighter. Nevertheless, for a number of iterations smaller than the typical time of exponential convergence, the exponential bound is non-informative. In this phase, the polynomial rates based on the capacity and source conditions are tighter; they govern the qualitative behavior of the algorithm.

As an illustration, consider in Figure 1.11 the performance of the simple gossip algorithm on the two-dimensional torus $\mathbb{T}^1_{200}$. The associated gossip problem satisfies both a spectral dimension condition (Example 1.13) and a spectral gap condition (Example 1.10). The spectral dimension dictates a polynomial convergence $O(t^{-d/2}) = O(t^{-1/2})$ in a first phase, best seen as a line with slope $-1/2$ in the plot on the left where both axes are logarithmic. The spectral gap dictates an exponential converge in a second phase, best seen as a line in the plot on the right where only the y-axis is logarithmic.

This two-phase phenomena can be explained as follows: the smaller the eigenvalue of the Hessian, the harder it is to optimize in the direction of the corresponding eigenvector. Thus, the long-time error of algorithms is governed by the small eigenvalues of the Hessian; the error in the other directions is negligible. The typical size of the eigenvalues that govern the error depends on the number of iterations (or the elapsed time, for gossip algorithms). In a first phase, the group of the smallest eigenvalues matters: the source and capacity condition explain the behavior. In a second phase, only the smallest eigenvalue matters: the strong convexity explains the behavior.

When accelerating algorithms, we do not only want asymptotically faster algorithms, but algorithms that are faster in both phases. We need to design accelerations jointly in the strong convexity and in the capacity and source conditions of our problems. This is done, in the particular case of synchronous gossip algorithms, in Section 4.5. This largely improves over accelerations based only on the spectral gap.

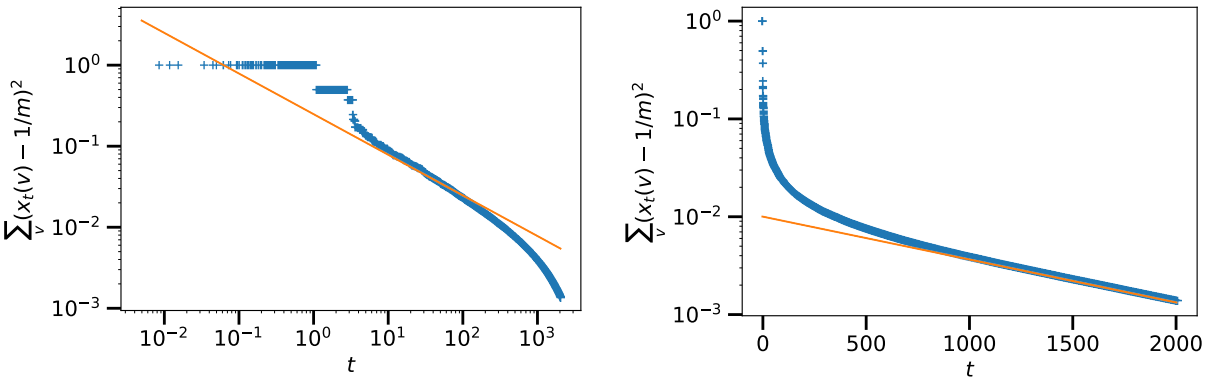


FIGURE 1.11. Rate of convergence of one realization of simple gossip algorithm on the torus $\mathbb{T}^1_{200}$, initialized from the indicator of a distinguished vertex. Blue crosses indicate the values of $\sum_{v\in\mathcal{V}}(x_t(v) - \bar{x})^2$ at the activation times. Both plots show the same values, but with a different scale for the x-axis. In both plots, we added an orange line to emphasize a specific part of the convergence: polynomial on the left, exponential on the right.

CHAPTER 2

Stochastic Gradient Descent and the Simple Gossip Algorithm

We remind that the contents of this chapter were published in the following conference article:

> R. Berthier, F. Bach, P. Gaillard. Tight Nonparametric Convergence Rates for Stochastic Gradient Descent under the Noiseless Linear Model, 2020, *Advances in Neural Information Processing Systems (NeurIPS)*.

In this chapter, we set ourselves in the least-squares supervised learning problem of Example 1.5: we seek to minimize

$$f(x) = \frac{1}{2}\mathbb{E}_{(a,b)\sim\mathcal{P}} \left(b - \langle a, x\rangle\right)^2 . \tag{2.1}$$

from i.i.d. samples $(a_1, b_1), (a_2, b_2), \cdots \sim \mathcal{P}$. We analyze the performance of the stochastic gradient descent (SGD) algorithm

$$x_{n+1} = x_n + \gamma(b_{n+1} - \langle a_{n+1}, x_n\rangle)a_{n+1} , \qquad n \geqslant 0 . \tag{2.2}$$

Motivated by the applications enumerated in Section 1.4.3, we mostly set ourselves in under the noiseless linear model (Definition 1.4): we assume that there exists a ground-truth linear relation $b = \langle x_*, a\rangle$ between the feature vector a and the output $b \in \mathbb{R}$. The feature vector a may be itself a non-linear transformation of the inputs $\Psi(u)$, explicitly computed through a feature map $a = \Psi(u)$ or implicitly defined through a positive-definite kernel $k(u, u')$ [Hofmann et al., 2008]. Note that in the noiseless model, there is still the randomness of the sampling of $a_1, \ldots, a_n$, sometimes called multiplicative noise [Dieuleveut et al., 2017]. Given those inputs, the outputs $b_1, \ldots, b_n$ are deterministic: there is no additive noise, and thus the noiseless linear model we consider in this paper is a simplification of problems with low additive noise. In Section 2.2, we describe how the results are perturbed in the presence of additive noise.

We analyze the behavior of an extremely naive algorithm: SGD, with no step-size decay nor averaging, no explicit regularization, and a single pass on the data. Remarkably, under the noiseless linear model, the iterates of SGD converge to the optimum x_* and the generalization error of SGD vanishes as the number of samples increases. Assuming strong convexity, the convergence shown to be exponential. This result is already known in the literature, see for instance [Bottou et al., 2018, Theorem 4.6]; we only present it to put our contributions in perspective. Our main result is that, under capacity and source conditions, the convergence of SGD is polynomial with exponents determined by the minimum of two parameters: the regularity of the optimum x_* and the regularity of the feature vectors a, where regularities are measured in terms of power norms of the covariance matrix $\Sigma = \mathbb{E}[a \otimes a]$. Our analysis of the convergence is tight as we prove upper and lower bounds on the performance of SGD that almost match. Thus plain-vanilla SGD shows some adaptivity to the complexity of the problem.

Two extensions of our results are studied. First, in Section 2.3.1, we study the application to the interpolation of a real function φ_* on the torus $[0, 1]^d$ from the observation of its value at randomly uniformly sampled points (Example 1.7). In the latter case, we show that the rate of convergence depends on the Sobolev smoothness of the function φ_* and of the interpolating kernel. Second, in Section 2.3.2, we use the parallel of between gossip algorithms and stochastic optimization to obtain

polynomial convergence rates for the simple gossip algorithm depending on the spectral dimension of the graph. Finally, in Section 2.3.3, a toy application instantiates our results in the special case of Gaussian features.

Comparison to the existing literature. There is an extensive research on the performance of different estimators in non-parametric supervised learning, however almost all of them do not consider the special case of the noiseless linear model [Györfi et al., 2006, Caponnetto and De Vito, 2007, Tsybakov, 2008, Fischer and Steinwart, 2017]. The difference is significant; for instance, rates faster than $O(n^{-1})$ for the least-square risk are impossible with additive noise, while in this paper we prove that SGD can converge with arbitrarily fast polynomial rates. Some of these works analyze the performance of SGD [Ying and Pontil, 2008, Bach and Moulines, 2013, Tarrès and Yao, 2014, Rosasco and Villa, 2015, Dieuleveut and Bach, 2016, Dieuleveut et al., 2017, Lin and Cevher, 2018, Pillaud-Vivien et al., 2018, Mücke et al., 2019]. However, because of the additive noise of the data, convergence requires averaging or decaying step sizes. As a notable exception, Jun et al. [2019] study a variant of kernel regularized least-squares and notices that the rate of convergence improves on noiseless data compared to noisy data. However, their rates are not directly comparable to ours as they assume that the optimal predictor is outside of the kernel space while we focus on the attainable case where the optimal predictor is in this space. We make a more precise comparison of this work with our results in Remark 2.2.

While our exposition chooses to study minimization of the test error by single-pass SGD, it is still possible to derive the convergence rates of multi-pass SGD for minimizing the training error; this simply corresponds to the case where the law $\mathcal{P}$ appearing in (2.1) is the uniform law on a dataset $(a_1, b_1), \dots, (a_N, b_N)$. This is, again, the subtle difference between Example 1.4 and Example 1.5 or between the interpolation regime and the noiseless model (Section 1.4.3). While the former is included in the latter, the converse is not true.

However, a recent trend studies the ability of SGD to reach zero training error in this interpolation regime, that is in overparameterized models where a perfect fit on the training data is possible [Schmidt and Le Roux, 2013, Ma et al., 2018, Vaswani et al., 2019, Cevher and Vũ, 2019]. Even with a fixed step size, SGD is shown to achieve zero-training error. However, these results are significantly different from ours: zero training error does not give any information on the generalization ability of the learned models, and the "interpolation regime" does not imply the noiseless model. Moreover, none of these studies give non-parametric convergence rates for SGD.

Setting. We assume the feature variable a to be uniformly bounded, namely that there exists a constant $R_0 < \infty$ such that

$$\|a\|^2 \leqslant R_0 \qquad \text{a.s.} \tag{2.3}$$

We can then define the covariance operator $\Sigma = \mathbb{E}\left[a \otimes a\right]$ of a. It has a finite operator norm that we denote $\|\Sigma\|_{\mathcal{H}\to\mathcal{H}}$. Recall that

$$f(x) = \frac{1}{2}\left\langle x - x_*, \Sigma\left(x - x_*\right)\right\rangle + f(x_*)\,.$$

We do not assume that the linear operator Σ is invertible as this is incompatible in infinite dimension with the boundedness assumption in Eq. (2.3). Finally, we recall the following notation: if α is a positive real and x a vector, $\left\|\Sigma^{-\alpha/2}x\right\|^2 = \langle x, \Sigma^{-\alpha}x\rangle := \inf\left\{\|x'\|^2 \,\middle|\, x' \text{ such that } x = \Sigma^{\alpha/2}x'\right\}$, with the convention that it is equal to ∞ when $x \notin \Sigma^{\alpha/2}(\mathcal{H})$.

2.1. Noiseless model

In this section, we assume that the noiseless linear model holds, i.e., that there exists an optimal predictor $x_* \in \mathcal{H}$ such that $b = \langle x_*, a\rangle$ a.s., or equivalently, $f(x_*) = 0$.

We first give a known result assuming strong convexity (Definition 1.5). For instance, it corresponds to [Bottou et al., 2018, Theorem 4.6] in the least-squares case and with $M = 0$.

Theorem 2.1 (noiseless, parametric)**.** Assume that f is μ-strongly convex for some $\mu > 0$. Assume further $0 < \gamma \leqslant 1/R_0$. The iterates x_n of SGD with step-size γ satisfy for all $n \geqslant 0$,

$$\mathbb{E}\|x_n - x_*\|^2 \leqslant (1 - \mu\gamma)^n \|x_0 - x_*\|^2 .$$

From the theorem above, an exponential bound on the expected population risk $\mathbb{E}f(x_n)$ can also be deduced from the inequality $f(x_n) \leqslant \frac{\|\Sigma\|_{\mathcal{H}\to\mathcal{H}}}{2}\|x_n - x_*\|^2$.

In this section, we adapt this result to the case where we no longer assume strong convexity, but we assume source and capacity conditions (Definitions 1.7 and 1.8). In this case, the reconstruction error $\|x_n - x_*\|^2$ and the population risk $f(x_n)$ converge at different rates.

Theorem 2.2 (noiseless, non-parametric, upper-bound)**.** Assume that there exists a non-negative real number $\underline{\alpha}$ such that

(a) the source condition is satisfied with regularity $\underline{\alpha}$, i.e., $x_* - x_0 \in \Sigma^{\underline{\alpha}/2}(\mathcal{H})$, and
(b) the capacity condition is satisfied with the same regularity $\underline{\alpha}$, i.e., $a \in \Sigma^{\underline{\alpha}/2}(\mathcal{H})$ a.s., and there exists a constant $R_{\underline{\alpha}} < \infty$ such that $\|\Sigma^{-\underline{\alpha}/2}a\|^2 \leqslant R_{\underline{\alpha}}$ a.s.

Assume further $0 < \gamma \leqslant 1/R_0$. The iterates x_n of SGD with step-size γ satisfy for all $n \geqslant 1$,

(1) (reconstruction error) $$\mathbb{E}\|x_n - x_*\|^2 \leqslant \frac{C}{n^{\underline{\alpha}}} ,$$

(2) (generalization error) $$\min_{k=0,\dots,n} \mathbb{E}f(x_k) \leqslant \frac{C'}{n^{\underline{\alpha}+1}} ,$$

where

$$C = \frac{\underline{\alpha}^{\underline{\alpha}}}{\gamma^{\underline{\alpha}}}\left(\|\Sigma^{-\underline{\alpha}/2}(x_* - x_0)\|^2 + \frac{R_{\underline{\alpha}}}{R_0}\|x_* - x_0\|^2\right) ,$$
$$C' = 2^{\underline{\alpha}}\frac{\underline{\alpha}^{\underline{\alpha}}}{\gamma^{\underline{\alpha}+1}}\left(\|\Sigma^{-\underline{\alpha}/2}(x_* - x_0)\|^2 + \frac{R_{\underline{\alpha}}}{R_0}\|x_* - x_0\|^2\right) .$$

Further, the tail-averaged iterate $\overline{x}_n = \dfrac{1}{\lfloor n/2\rfloor + 1}\displaystyle\sum_{k=\lceil n/2\rceil}^{n} x_k$ satisfies the similar bounds

(1) (reconstruction error) $$\mathbb{E}\|\overline{x}_n - x_*\|^2 \leqslant \frac{2^{\underline{\alpha}}C}{n^{\underline{\alpha}}} ,$$

(2) (generalization error) $$\mathbb{E}f(\overline{x}_n) \leqslant \frac{C'}{n^{\underline{\alpha}+1}} .$$

In the following theorem, we show that the non-parametric theorem is tight in the exponents.

Theorem 2.3 (noiseless, non-parametric, lower bound)**.** Assume that there exists a positive real number $\overline{\alpha}$ such that one of the two following conditions holds:

(a) $x_* - x_0 \notin \Sigma^{\overline{\alpha}/2}(\mathcal{H})$, or
(b) with positive probability, $a \notin \Sigma^{\overline{\alpha}/2}(\mathcal{H})$ and $\langle a, x_* - x_0\rangle \neq 0$.

Assume further $0 < \gamma \leqslant 1/R_0$. The iterates x_n of SGD with step-size γ satisfy for all $\varepsilon > 0$,

(1) (reconstruction error) $\mathbb{E}\|x_n - x_*\|^2$ is not asymptotically dominated by $1/n^{\overline{\alpha}+\varepsilon}$,
(2) (generalization error) $\mathbb{E}f(x_n)$ is not asymptotically dominated by $1/n^{\overline{\alpha}+1+\varepsilon}$.

The take-home message of Theorems 2.2, 2.3 is that the convergence rate of SGD is governed by two real numbers: the regularity α_1 of the optimum, that is the supremum of all $\underline{\alpha}$ such that $x_* - x_0 \in \Sigma^{\underline{\alpha}/2}(\mathcal{H})$, and the regularity α_2 of the features, that is the supremum of all $\underline{\alpha}$ such that $a \in \Sigma^{\underline{\alpha}/2}(\mathcal{H})$ almost surely. The polynomial convergence rate of SGD is roughly of the order of $n^{-\alpha}$ for the reconstruction error and $n^{-\alpha-1}$ for the generalization error with $\alpha = \min(\alpha_1, \alpha_2)$: one of the two regularities is a bottleneck for fast convergence. See Section 2.3.1 for an application to the optimal choice of a reproducing kernel Hilbert space. The exponent α_1 corresponds to the decay of the errors of the gradient descent on the population risk f. However, due to the multiplicative noise, the convergence of SGD is slowed down by the irregularity of the feature vectors if $\alpha_2 < \alpha_1$.

In the theorems, the constraint on the step-size $0 < \gamma \leqslant 1/R_0$ is independent of the time horizon n and of the regularities α_1, α_2. Thus fixed step-size SGD shows some adaptivity to the regularity of the problem.

In Section 2.3, we give extensive numerical evidence that the polynomial rates $n^{-\alpha}$ and $n^{-(\alpha+1)}$ in the bounds are indeed sharp in describing convergence rate of SGD.

We now make a few remarks on Theorems 2.2, 2.3. They articulate the significance of the results, but may be skipped.

Remark 2.1. Our upper bound and lower bound on the generalization errors of SGD do not match exactly. Indeed, we prove an upper bound on the *minimum* risk of the past iterates, where we prove a lower bound on a larger quantity, the risk of the *last* iterate. Note that as f is not exactly observable, it is not trivial to determine which past iterate of SGD satisfies the bound. However, Theorem 2.2 also shows that tail-averaging is sufficient to obtain an iterate with the claimed generalization bound.

The risk of the last iterate of SGD was bounded with the same rates in a subsequent paper by Varre et al. [2021]; their assumptions differ from those of this chapter in two ways. First, their analysis is more general as it also covers the non-attainable case where the infimum of (2.1) is not realized by a point in $\mathcal{H}$ (but is still 0). Second, their results requires the step-size γ to be significantly smaller that $1/R_0$; for this reason, it can not be applied to gossip algorithms as below. We do not know whether it is possible to obtain the same rates for the risk of the last iterate for step-sizes as large as $\gamma = 1/R_0$.

Remark 2.2 (related literature)**.** In the case $\underline{\alpha} = 0$, where no regularity assumption is made on the optimum or the features (apart from being bounded), we upper-bound $\min_{k=1,\dots,n} \mathbb{E} f(x_k)$ by $O(n^{-1})$. A similar result was shown by Bach and Moulines [2013]: the excess risk for averaged constant-step size SGD is asymptotically dominated by n^{-1} on any least-squares problem–not necessarily a noiseless one. It is remarkable that under the noiseless linear setting, no averaging or decay of the step-size is needed to obtain the same convergence rate.

Jun et al. [2019] also study the performance of an algorithm, a variant of kernel regularized least-squares, in the noiseless non-parametric setting. However, they do not consider the case where the function is more regular than being in the kernel space, i.e., when $\alpha_1 > 0$ with our notation, $\beta > 1/2$ with theirs. In fact, they leave this case as an open problem in their Section 6. Thus, a fair comparison can only be made when $\alpha_1 = 0, \beta = 1/2$. In this case, SGD and the algorithm of Jun et al. [2019] both achieve the same rate $O(n^{-1})$.

Remark 2.3 (relation to Hölderian error bounds and uniform convexity)**.** Our non-parametric study bears similarities with studies of optimization under the assumption that the function

satisfies a Hölderian error bound

$$f(y) - f(x_*) \geqslant \frac{\mu}{2}\|y - x_*\|^\rho\,, \qquad y \in \mathcal{H}\,,$$

for some $\rho > 2$ and $\mu > 0$ [Juditsky and Nesterov, 2014, Roulet and d'Aspremont, 2020]. This condition is implied by uniform convexity:

$$f(y) \geqslant f(x) + \langle f'(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^\rho\,, \qquad x, y \in \mathcal{H}\,.$$

The case $\rho = 2$ corresponds to strong convexity.

Our quadratic function $f(x)$ is uniformly convex on the ellipsoids $\{x \,|\, \langle x - x_*, \Sigma^{-\alpha}(x - x_*)\rangle \leqslant C\}$, $C > 0$ (with $\rho = 2 + 2/\alpha$). Our proof strategy can be seen as follows: we use the capacity and source conditions to show that the iterates stay in an ellipsoid of this form (Equation (2.16)). We then use the implied Hölderian error bound (Equation (2.17)) and conclude.

Remark 2.4. The theorems stated above stay true if one weakens the assumptions in the following way:

- assume $\mathbb{E}\left[\|a\|^2 a \otimes a\right] \preccurlyeq R_0\Sigma$ instead of $\|a\|^2 \leqslant R_0$ a.s., and
- assume $\mathbb{E}\left[\langle a, \Sigma^{-\underline{\alpha}}a\rangle\, a \otimes a\right] \preccurlyeq R_{\underline{\alpha}}\Sigma$ instead of $\langle a, \Sigma^{-\underline{\alpha}}a\rangle \leqslant R_{\underline{\alpha}}$ a.s.

This weaker set of assumptions is useful in the case of non-bounded features, like the Gaussian features of Section 2.3.3. We thus take special care in using only these weaker assumptions in the proofs of Theorems 2.2-2.5. However we prefer stating results with the stronger assumptions for the sake of clarity.

Remark 2.5 (Articulation between Theorems 2.1 and 2.2)**.** If f is strongly convex, for instance because $\mathcal{H}$ is finite-dimensional and Σ is of full rank, then the assumptions of Theorem 2.2 hold for any $\underline{\alpha} \geqslant 0$. Thus SGD converges asymptotically faster than any polynomial; this is coherent with the exponential convergence given by Theorem 2.1. Although the latter bound is asymptotically better than polynomial rates, for moderate time scales the polynomial rates may describe best the observed behavior, see Section 1.5.3 for an illustration.

Regularity functions and general results. We now generalize Theorems 2.2 and 2.3 by showing the convergence in norms associated to different powers of the covariance Σ. Indeed, the main difficulty in the proof of these theorems is that deriving closed recurrence relations for the expected reconstruction and generalization errors is not straightforward. Instead, we define the regularity function $d_n^2, \overline{d}_n^2$ at iteration n:

$$\begin{aligned} d_n^2(\beta) &= \mathbb{E}\left[\left\langle x_n - x_*, \Sigma^{-\beta}\left(x_n - x_*\right)\right\rangle\right] \in [0, \infty]\,, \\ \overline{d}_n^2(\beta) &= \mathbb{E}\left[\left\langle \overline{x}_n - x_*, \Sigma^{-\beta}\left(\overline{x}_n - x_*\right)\right\rangle\right] \in [0, \infty]\,, \qquad \beta \in \mathbb{R}\,, \end{aligned}$$

and we exhibit a closed recurrence inequality (Property 2 in Appendix 2.A) for the regularity functions d_n^2, $n \geqslant 1$. In particular, one can recover the expected squared norm and the expected risk of the iterates from the regularity function:

$$d_n^2(0) = \mathbb{E}\|x_n - x_*\|^2 \qquad \text{and} \qquad d_n^2(-1) = 2\mathbb{E}f(x_n)\,.$$

Theorems 2.2 and 2.3 can be extended to the following estimates on the regularity functions $d_n^2(\beta), \overline{d}_n^2(\beta)$ on the full interval $\beta \in [-1, \underline{\alpha}]$ (see proofs in Appendices 2.A and 2.C respectively).

Theorem 2.4 (noiseless, non-parametric, upper bound)**.** Under the assumptions of Theorem 2.2, we have for all $n \geqslant 1$,

(1) for all $\beta \in [0, \underline{\alpha}]$, $\quad d_n^2(\beta) \leqslant \dfrac{C}{n^{\underline{\alpha}-\beta}}$,

(2) for all $\beta \in [-1, 0)$, $\quad \min_{k=0,\dots,n} d_k^2(\beta) \leqslant \dfrac{C'}{n^{\underline{\alpha}-\beta}}$,

where

$$C = \frac{\underline{\alpha}^{\underline{\alpha}-\beta}}{\gamma^{\underline{\alpha}-\beta}} \left(\|\Sigma^{-\underline{\alpha}/2}(x_* - x_0)\|^2 + \frac{R_{\underline{\alpha}}}{R_0}\|x_* - x_0\|^2 \right),$$
$$C' = 2^{\underline{\alpha}-\beta}\frac{\underline{\alpha}^{\underline{\alpha}}}{\gamma^{\underline{\alpha}-\beta}} \left(\|\Sigma^{-\underline{\alpha}/2}(x_* - x_0)\|^2 + \frac{R_{\underline{\alpha}}}{R_0}\|x_* - x_0\|^2 \right).$$

Further, the tail-averaged iterate $\overline{x}_n = \dfrac{1}{\lfloor n/2 \rfloor + 1} \displaystyle\sum_{k=\lceil n/2 \rceil}^{n} x_k$ satisfies the similar bounds

$$\text{for all } \beta \in [-1, \underline{\alpha}], \qquad \overline{d}_n^2(\beta) \leqslant \frac{C''}{n^{\underline{\alpha}-\beta}},$$

where $C'' = 2^{\underline{\alpha}-\beta}C$ if $\beta \in [0, \underline{\alpha}]$ and $C'' = C'$ if $\beta \in [-1, 0)$.

Theorem 2.5 (noiseless, non-parametric, lower bound)**.** Under the assumptions of Theorem 2.3, for all $\beta \in [-1, \overline{\alpha}]$, for all $\varepsilon > 0$, $d_n^2(\beta)$ is not asymptotically dominated by $1/n^{\overline{\alpha}-\beta+\varepsilon}$.

2.2. Noisy model and robustness to model perturbation

In this section, we describe how the results of Section 2.1 are perturbed in the case where a linear relation $b = \langle x_*, a \rangle$ a.s. does not hold. Following the statistical learning framework, we assume a joint law $\mathcal{P}$ on (a, b). We further assume that there exists a minimizer $x_* \in \mathcal{H}$ of the population risk $f(x)$:

$$x_* \in \operatorname*{argmin}_{x \in \mathcal{H}} \left\{ f(x) = \frac{1}{2}\mathbb{E}\left(b - \langle x, a \rangle\right)^2 \right\}.$$

This general framework encapsulates two types of perturbations of the noiseless linear model:

- (noisy model) The output b can be uncertain given a. For instance, under the noisy linear model, $b = \langle x_*, a \rangle + z$, where z is centered and independent of a. In this case, $f(x_*) = \frac{1}{2}\mathbb{E}[z^2] = \frac{1}{2}\mathbb{E}[\mathrm{var}\,(b|a)]$.
- (non-linear model) Even if b is deterministic given a, this dependence can be non-linear: $b = \varphi_*(a)$ for some non-linear function φ_*. Then $f(x_*)$ is the squared L^2 distance of the best linear approximation to φ_*: $f(x_*) = \frac{1}{2}\mathbb{E}\left[\left(\varphi_*(a) - \langle x_*, a \rangle\right)^2\right]$.

In the general framework, the optimal population risk is a combination of both sources

$$f(x_*) = \frac{1}{2}\mathbb{E}\left[\mathrm{var}\,(b|a)\right] + \frac{1}{2}\mathbb{E}\left[\left(\mathbb{E}[b|a] - \langle x_*, a \rangle\right)^2\right].$$

Apart from the new definition of x_*, the assumptions and definitions are the same as above. In particular, $d_n^2(\beta) = \mathbb{E}\left\langle x_n - x_*, \Sigma^{-\beta}\left(x_n - x_*\right)\right\rangle$.

As in the first section, we first introduce a known result assuming strong convexity. For instance, it corresponds to [Bottou et al., 2018, Theorem 4.6] in the least-squares case.

Theorem 2.6 (noisy, parametric)**.** We make the same assumptions as in Theorem 2.1. The iterates x_n of SGD satisfy

$$\mathbb{E}\|x_n - x_*\|^2 \leqslant 2\left(1-\mu\gamma\right)^n \|x_0 - x_*\|^2 + \frac{4R_0}{\mu}\gamma f(x_*)\,.$$

The take-home message is that we get an upper bound of the form $2\left(1-\mu\gamma\right)^n \|x_0 - x_*\|^2$, analog to Theorem 2.1, but with an additional constant term $\frac{4R_0}{\mu}\gamma f(x_*)$. This term can be small if $f(x_*)$ is small, that is if the problem is close to the noiseless linear model, or if the step-size γ is small. We now present a similar result under capacity and source conditions.

Theorem 2.7 (noisy, non-parametric)**.** We make the same assumptions as in Theorem 2.2. The iterates x_n of SGD satisfy

$$\min_{k=0,\dots,n} \mathbb{E}\left[f(x_k) - f(x_*)\right] \leqslant 2\frac{C'}{n^{\underline{\alpha}+1}} + 2R_0\gamma f(x_*)\,,$$

where C' is the same constant as in Theorem 2.2.

Further, the tail-averaged iterate $\overline{x}_n = \dfrac{1}{\lfloor n/2\rfloor + 1}\displaystyle\sum_{k=\lceil n/2\rceil}^{n} x_k$ satisfies

$$\mathbb{E}\left[f(\overline{x}_n) - f(x_*)\right] \leqslant 2\frac{C'}{n^{\underline{\alpha}+1}} + 2R_0\gamma f(x_*)\,.$$

Here, we consider the excess risk $f(x_k) - f(x_*)$ instead of the risk $f(x_k)$ as the optimal risk $f(x_*)$ is no longer equal to 0. In the finite horizon setting setting, one can optimize γ as a function of the scheduled number of steps n in order to balance both terms in the upper bound. As $C' \propto \gamma^{-(\underline{\alpha}+1)}$, the optimal choice is $\gamma \propto n^{-(\underline{\alpha}+1)/(\underline{\alpha}+2)}$ which gives a rate $\min_{k=0,\dots,n} \mathbb{E}\left[f(x_k) - f(x_*)\right] = O\left(n^{-(\underline{\alpha}+1)/(\underline{\alpha}+2)}\right)$.

In the theorem below, we study the SGD iterates x_n in terms of the power norms $d_n^2(\beta)$, $\beta \in [-1, \underline{\alpha} - 1]$, in particular in term of the reconstruction error $d_n^2(0) = \mathbb{E}\|x_n - x_*\|^2$ if $\underline{\alpha} \geqslant 1$. Note that the population risk $f(x)$ is a quadratic with Hessian Σ, minimized at x_*, thus

$$\mathbb{E}\left[f(x_n) - f(x_*)\right] = \frac{1}{2}\mathbb{E}\left\langle x_n - x_*, \Sigma(x_n - x_*)\right\rangle = \frac{1}{2}d_n^2(-1)\,.$$

Thus the theorem below extends Theorem 2.7.

Theorem 2.8 (noisy, non-parametric)**.** We make the same assumptions as in Theorem 2.2. The iterates x_n of SGD satisfy

(1) for all $\beta \geqslant 0$, $\beta \leqslant \underline{\alpha} - 1$,

$$d_n^2(\beta) \leqslant 2\frac{C}{n^{\underline{\alpha}-\beta}} + 4R_0^{1-(\beta+1)/\underline{\alpha}} R_{\underline{\alpha}}^{(\beta+1)/\underline{\alpha}} \gamma f(x_*)\,,$$

(2) for all $\beta \in [-1, 0)$, $\beta \leqslant \underline{\alpha} - 1$,

$$\min_{k=1,\dots,n} d_k^2(\beta) \leqslant 2\frac{C'}{n^{\underline{\alpha}-\beta}} + 4R_0^{1-(\beta+1)/\underline{\alpha}} R_{\underline{\alpha}}^{(\beta+1)/\underline{\alpha}} \gamma f(x_*)\,,$$

where C, C' are the same constants as in Theorem 2.4.

Further, the tail-averaged iterates $\overline{x}_n$ satisfy

$$\text{for all } \beta \in [-1, \underline{\alpha} - 1]\,, \qquad \overline{d}_n^2(\beta) \leqslant 2\frac{C''}{n^{\underline{\alpha}-\beta}} + 4R_0^{1-(\beta+1)/\underline{\alpha}} R_{\underline{\alpha}}^{(\beta+1)/\underline{\alpha}} \gamma f(x_*)\,,$$

where C'' is the same constant as in Theorem 2.4.

This theorem is proved in Appendix 2.B. We expect the condition $\beta \leqslant \underline{\alpha} - 1$ to be necessary. More precisely, when $f(x_*)$ is positive, we expect the error $x_n - x_*$ to diverge under the norm $\|\Sigma^{-\beta/2} . \|$ if $\beta > \underline{\alpha} - 1$. In particular, this would imply that the reconstruction error diverges when $\underline{\alpha} < 1$. In Section 2.3.3, we provide simulations in a noisy setting for which the results above accurately predict the qualitative behavior of stochastic gradient descent.

2.3. Applications

2.3.1. Function interpolation on $[0,1]^d$. We consider Example 1.7: let $(u_1, \varphi_*(u_1))$, $(u_2, \varphi_*(u_2))$, ... be the observations of an unknown function φ_* at random points $u_1, u_2, \ldots$ i.i.d. uniform in $[0,1]^d$. We saw that the kernel stochastic gradient descent on functions $\varphi_n : [0,1]^d \to \mathbb{R}$,

$$\varphi_{n+1} = \varphi_n + \gamma(\varphi_*(u_{n+1}) - \varphi_n(u_{n+1}))k(u_{n+1}, .)\,, \tag{2.4}$$

could be seen as a least-squares stochastic gradient descent in the RKHS $\mathcal{H}$, associated to the kernel k, on the noiseless stochastic optimization problem

$$\min_{\varphi \in \mathcal{H}} \quad f(\varphi) = \frac{1}{2}\|\varphi_* - \varphi\|^2_{L^2([0,1]^d)} = \frac{1}{2}\mathbb{E}_{u \sim \mathrm{Unif}([0,1]^d)} \left(\varphi_*(u) - \varphi(u)\right)^2 . \tag{2.5}$$

Moreover, we set ourselves in the Sobolev case of Example 1.8: we assume that $k(u,u') = t(u-u')$ is translation invariant and that the Fourier series of t satisfies a power-law decay:

$$c\left(1 + |v|^2\right)^{-s/2-d/4} \leqslant \widehat{t}(v) \leqslant C\left(1 + |v|^2\right)^{-s/2-d/4}\,, \qquad v \in \mathbb{Z}^d\,,$$

for some constants $c, C > 0$. In this case, the RKHS $\mathcal{H}$ is equivalent to $H^{s/2+d/4}_{\mathrm{per}}$. We also assume $\varphi_* \in H^r_{\mathrm{per}}$.

In this case, from Example 1.12, the problem (2.5) satisfies the source condition for $\alpha_1 = \frac{2r}{s+d/2} - 1$ and the capacity condition for all $\alpha_2 < 1 - \frac{d}{s+d/2}$.

From Theorems 2.2-2.3, $\mathbb{E}f(\varphi_n) = \frac{1}{2}\|\varphi_* - \varphi_n\|^2_{L^2([0,1]^d)}$ decays to zero at a polynomial rate with exponent

$$\alpha_* + 1 = \min\left(\frac{2r}{s+d/2}, 2 - \frac{d}{s+d/2}\right) . \tag{2.6}$$

Note that, given a function φ_*, this rate is maximal when $s = r$, i.e., the smoothness of the kernel coincides with the smoothness of the function, in which case $\alpha_* = 1 - \frac{d}{r+d/2}$. Theorems 2.2, 2.3 also give the convergence rates in terms of the RKHS norm, which happens to be a Sobolev norm. The more general Theorems 2.4 and 2.5 gives convergence rates in terms of a continuity of fractional Sobolev norms, some weaker and some stronger than the RKHS norm.

In Figure 2.1, we show the decay of the L^2 norm in the interpolation of a function φ_* on $[0,1]$ of smoothness 2 using kernels of smaller, matching and larger smoothness. In each case, the rate predicted by (2.6) is sharp, and the convergence is indeed fastest when the smoothnesses match.

Comparison to the literature on function interpolation. The field of scattered data approximation [Wendland, 2004] studies the estimation of a function from the observation of its values at (possibly random) points. Again, most of the work focuses on the case where the observation of the values is noisy, with notable exceptions. Delyon and Juditsky [1997] obtain the optimal rate $n^{-r/d}$ (up to logarithmic factors) in estimating a function in the Sobolev space H^r $(r > d/2)$ from n i.i.d. observations; this outperforms the rate $n^{-r/(r+d/2)}$ obtained above for SGD. Their algorithm is based on estimating the wavelet coefficients of the function and its complexity is $O(n)$. In general, this suggests that SGD might not achieve the non-parametric minimax rates under the noiseless linear model.

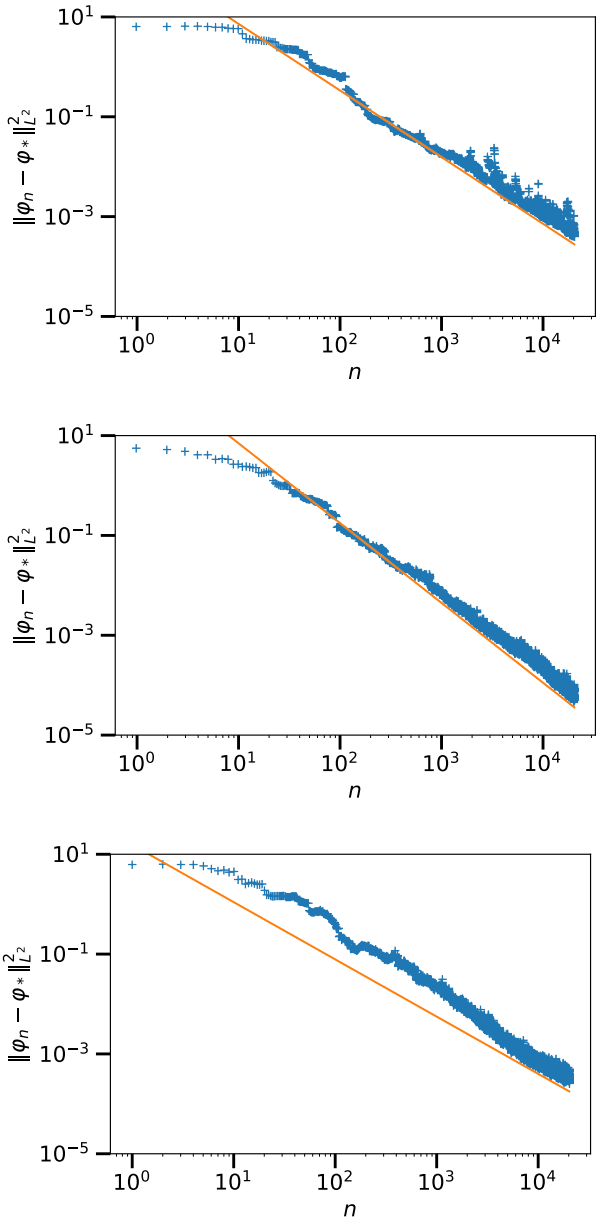


FIGURE 2.1. Interpolation of a function of smoothness $r = 2$ using SGD with kernels of smoothness $s = 1$ (up), $s = 2$ (middle) and $s = 3$ (down). Each plot represents one realization of the algorithm (2.4). The blue crosses represent the square L^2-norms $\|\varphi_n - \varphi_*\|^2_{L^2}$ as a function of the number of iterations n and the orange lines represent the predicted polynomials rates C/n^{α_*+1}, where C is chosen to match best the empirical observations for each plot.

Other works are not directly comparable but are similar in spirit. Bauer et al. [2017] show a minimax rate of $\Omega((\log n/n)^{p/d})$ for estimating a p-smooth function on $[0,1]^d$ in L^∞ norm using n independent uniformly distributed points; the minimax rate is reached with a spline estimate. Kohler and Krzyżak [2013] show a minimax rate of $\Omega(1/n^p)$ for the same problem, but in the special

case of $d = 1$ and estimation in L^1 norm; the minimax rate is reached with some nearest neighbor polynomial interpolation.

2.3.2. The simple gossip algorithm. In this section, we analyze the behavior of the simple gossip algorithm on a graph G depending on the spectral dimension d of the graph. We still denote m the number of vertices and N the number of edges. As explained in Section 1.4.2, considering continuous or discrete time does not matter for the study of this unaccelerated method; for the simplicity of the parallel with stochastic gradient descent, we consider discrete time. We define $x_n := x_{T_n}$, where $(x_t)_{t\geqslant 0}$ is the continuous-time simple gossip algorithm started from x_0 and T_n is the activation time of edge $\{v_n, w_n\}$. Recall that the differences $T_{n+1} - T_n$ are exponential random variables with expectation $1/N$ where N is the number of edges in the graph; it is thus natural to define the rescaled iterate number $s = n/N$ to have $s \approx t$.

In Section 1.4.1, we formulated the simple gossip algorithm as a stochastic gradient descent on a noiseless least-squares problem. This suggests to apply the results of this chapter to gossip algorithms. Moreover, Proposition 1.2 states that the associated least-squares problem satisfies the source and capacity conditions for all $\alpha_1, \alpha_2 < d/2$, where d is the spectral dimension of the graph. This sharp inequality prevents us from obtaining rates of the form $n^{-d/2}$; it causes technicalities leading to additional logarithmic terms. We obtain the following result.

Corollary 2.1 (of Theorem 2.2)**.** Assume that G is of spectral dimension d with constant V, and denote $\delta_{\max}$ the maximal degree of the nodes in the graph. Assume further that the initial observation $x_0 : \mathcal{V} \to \mathbb{R}$ is the indicator of some distinguished vertex $v_* \in \mathcal{V}$: $x_0(v_*) = 1$ and $x_0(v) = 0$ if $v \neq v_*$. Then, for all $s = n/N \geqslant 2$,

$$\mathbb{E}\left[\sum_{v\in\mathcal{V}}\left(x_{Ns}(v) - \frac{1}{m}\right)^2\right] \leqslant D(d, V, \delta_{\max})\frac{\log s}{s^{d/2}}\,, \tag{1}$$

$$\min_{0\leqslant s'\leqslant s} \mathbb{E}\left[\frac{1}{2}\sum_{\{v,w\}\in\mathcal{E}}\left(x_{Ns'}(v) - x_{Ns'}(w)\right)^2\right] \leqslant D'(d, V, \delta_{\max})\frac{\log s}{s^{d/2+1}}\,, \tag{2}$$

where $D(d, V, \delta_{\max}) = \dfrac{2}{\log 2} d^{d/2+1} V^{-1} \delta_{\max}$ and $D'(d, V, \delta_{\max}) = \dfrac{2^{d/2+2}}{\log 2} d^{d/2+1} V^{-1} \delta_{\max}$.

See Appendix 2.D for the proof. Note that as G is a finite graph, G can be of any spectral dimension d for some potentially large constant V. However, for many families of graphs of increasing size, such as the toruses $\mathbb{T}^d_\Lambda$, $\Lambda \geqslant 1$, the spectral dimension constant V and the maximum degree $\delta_{\max}$ remain bounded independently of the size of the graph, see Example 1.13. In that case, the bounds of Corollary 2.1 are independent of the size of the graph.

Indeed, in Figure 2.2, simulations on a large circle $\mathbb{T}^1_{300}$ and on a large torus $\mathbb{T}^2_{40}$ display polynomial decay rates, with polynomial exponents coinciding with those of the corresponding bounds of Corollary 2.1. Note that, if pushed on a longer time scale, the simulations would have shown the exponential convergence due to finite graph effects. While the polynomial exponents are sharp, we expect the logarithmic factors to be an artifact of the method of proof.

In the case $d = 0$ and $V = 1$, where no assumption on the structure of the graph is made, the fact that the minimal past energy is $O(n^{-1})$ (neglecting the logarithmic factor) has been noticed Aldous and Lanoue [2012, Proposition 4]. Aldous leaves as an open problem whether one can prove a bound without taking a minimum; this is a special case of our Remark 2.1.

2.3.3. Linear regression with Gaussian features. In the noiseless setting of Section 2.1, we assume a to be centered Gaussian process of covariance Σ where Σ is a bounded symmetric semidefinite operator. As a is not bounded a.s., we need to use the weaker set of assumptions given

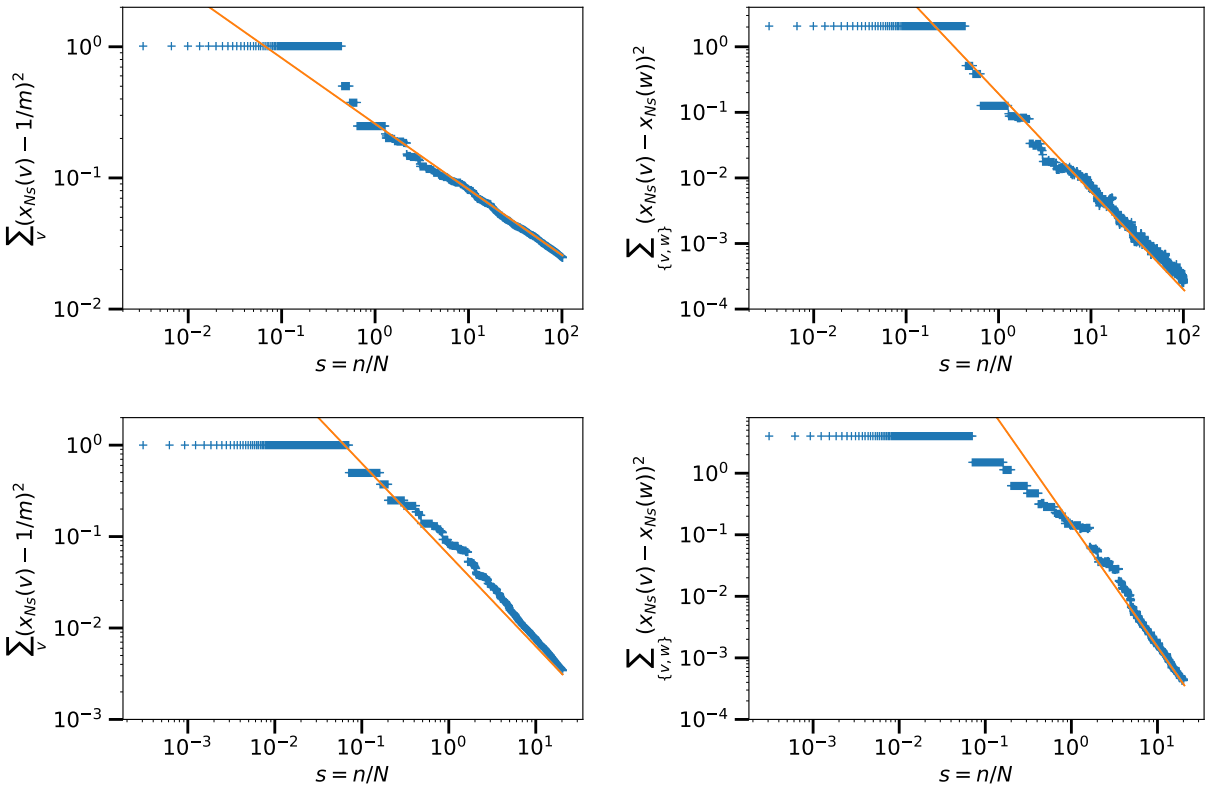


FIGURE 2.2. Convergence rates on the circle $\mathbb{T}^1_{300}$ (up) and on the two dimensional torus $\mathbb{T}^2_{40}$ (bottom). The convergence is measured in terms of squared ℓ^2-distance to $\frac{1}{m}\mathbb{1}$ (left) and sum of the squared differences along the edges (right). In orange are the curves of the form $C/s^{d/2}$ and $C'/s^{d/2+1}$ where C and C' are constants chosen to match best the empirical observations for each plot.

in Remark 2.4. We thus need to compute R_0 such that $\mathbb{E}\left[\|a\|^2 a \otimes a\right] \preccurlyeq R_0\Sigma$ and α, R_α such that $\mathbb{E}\left[\langle a, \Sigma^{-\alpha}a\rangle\, a \otimes a\right] \preccurlyeq R_\alpha\Sigma$. We show here that these conditions are in fact simple trace conditions on Σ, sometimes also called capacity conditions [Pillaud-Vivien et al., 2018].

Lemma 2.1. Assume $a \sim \mathcal{N}(0, \Sigma)$. If M is a bounded symmetric operator such that $\mathrm{Tr}(\Sigma M) < \infty$,

$$\mathbb{E}\left[\langle a, Ma\rangle\, a \otimes a\right] = 2\Sigma M\Sigma + \mathrm{Tr}(\Sigma M)\Sigma \preccurlyeq \left(2\|\Sigma^{1/2}M\Sigma^{1/2}\|_{\mathcal{H}\to\mathcal{H}} + \mathrm{Tr}(\Sigma M)\right)\Sigma\,.$$

PROOF. Diagonalize $\Sigma = \sum_{i\geqslant 1}\lambda_i e_i \otimes e_i$. Then there exists independent standard Gaussian random variables $a_i, i \geqslant 0$ such that $a = \sum_i \lambda_i^{1/2} a_i e_i$.

Let $i, j \geqslant 1$.

$$\begin{aligned}\langle e_i, \mathbb{E}\left[\langle a, Ma\rangle\, a \otimes a\right] e_j\rangle &= \mathbb{E}\left[\langle a, Ma\rangle\, \langle e_i, a \otimes a e_j\rangle\right] = \mathbb{E}\left[\langle a, Ma\rangle\, \lambda_i^{1/2} a_i \lambda_j^{1/2} a_j\right] \\ &= \lambda_i^{1/2}\lambda_j^{1/2}\sum_{k,l} M_{k,l}\lambda_k^{1/2}\lambda_l^{1/2}\mathbb{E}\left[a_i a_j a_k a_l\right]\,.\end{aligned}$$

As $a_i, i \geqslant 1$ are centered independent random variables, the quantity $\mathbb{E}\left[a_i a_j a_k a_l\right]$ is 0 in many cases. More precisely,

- if $i \neq j$, the general term of the sum in non-zero only when $k = i$ and $l = j$ or $k = j$ and $l = i$. This gives
$$\langle e_i, \mathbb{E}\left[\langle a, Ma\rangle\, a \otimes a\right] e_j\rangle = 2M_{i,j}\lambda_i\lambda_j\,.$$
- if $i = j$, the general term of the sum is non-zero only when $k = l$. This gives
$$\begin{aligned}\langle e_i, \mathbb{E}\left[\langle a, Ma\rangle\, a \otimes a\right] e_i\rangle &= \lambda_i \sum_k M_{k,k}\lambda_k \mathbb{E}\left[a_i^2 a_k^2\right] = \lambda_i \sum_{k\neq i} M_{k,k}\lambda_k + 3\lambda_i^2 M_{i,i} \\ &= \lambda_i \sum_k M_{k,k}\lambda_k + 2\lambda_i^2 M_{i,i}\,.\end{aligned}$$

In both cases,
$$\langle e_i, \mathbb{E}\left[\langle a, Ma\rangle\, a \otimes a\right] e_j\rangle = 2\lambda_i\lambda_j M_{i,j} + \left(\sum_k M_{k,k}\lambda_k\right)\lambda_i \mathbb{1}_{i=j}\,.$$

Note that
$$\mathrm{Tr}(M\Sigma) = \sum_k \langle e_k, \Sigma M e_k\rangle = \sum_k \lambda_k M_{k,k}\,.$$

Thus we get
$$\begin{aligned}\langle e_i, \mathbb{E}\left[\langle a, Ma\rangle\, a \otimes a\right] e_j\rangle &= 2\lambda_i\lambda_j M_{i,j} + \mathrm{Tr}(M\Sigma)\lambda_i \mathbb{1}_{i=j} \\ &= 2\,\langle e_i, \Sigma M \Sigma e_j\rangle + \mathrm{Tr}(M\Sigma)\,\langle e_i, \Sigma e_j\rangle \\ &= \langle e_i, \left[2\Sigma M\Sigma + \mathrm{Tr}(\Sigma M)\Sigma\right] e_j\rangle\ .\end{aligned}$$

□

From this lemma with $M = \mathrm{Id}$, we compute $R_0 = 2\|\Sigma\|_{\mathcal{H}\to\mathcal{H}} + \mathrm{Tr}(\Sigma)$, and with $M = \Sigma^{-\alpha}$, we compute $R_\alpha = 2\|\Sigma\|_{\mathcal{H}\to\mathcal{H}}^{1-\alpha} + \mathrm{Tr}(\Sigma^{1-\alpha})$. Thus in the Gaussian case, the condition of (weak) regularity of the features is given by $\mathrm{Tr}(\Sigma^{1-\alpha}) < \infty$.

> **Remark 2.6** (Beyond Gaussian distributions)**.** As suggested by Juditsky et al. [2020, p.17], note that if a satisfies the assumptions of Remark 2.4 with constants R_0 and $R_{\underline{\alpha}}$, and η is an independent random scalar with bounded first and second moment, then the scale mixture $\tilde{a} = \eta a$ also satisfies the assumptions of Remark 2.4 with the constants $\tilde{R}_0 = \frac{\mathbb{E}\eta^4}{\mathbb{E}\eta^2} R_0$ and $\tilde{R}_{\underline{\alpha}} = \frac{\mathbb{E}\eta^4}{(\mathbb{E}\eta^2)^{1+\underline{\alpha}}} R_{\underline{\alpha}}$. In particular, one can obtain multivariate Student distributions by taking the scale mixture of a multivariate Gaussian distribution a and of a scalar η of the form $\sqrt{q/\zeta}$ where ζ follows a χ^2-distribution with q degrees of freedom (see [Juditsky et al., 2020] for details). In particular, this shows that there are some heavy-tailed distributions satisfying the assumptions of Remark 2.4.

Simulations. We present simulations in finite but large dimension $d = 10^5$, and we check that dimension-independent bounds describe the observed behavior. We artificially generate regression problems with different regularities by varying the decay of the eigenvalues of the covariance Σ and varying the decay of the coefficients of x_*.

Choose an orthonormal basis $e_1, \ldots, e_d$ of $\mathcal{H}$. We define $\Sigma = \sum_{i=1}^d i^{-\beta} e_i \otimes e_i$ for some $\beta \geqslant 1$ and $x_* = \sum_{i=1}^d i^{-\delta} e_i$ for some $\delta \geqslant 1/2$. We now check the condition on α such that the assumptions (a) and (b) are satisfied.

(a) $\langle x_*, \Sigma^{-\alpha} x_*\rangle = \sum_{i=1}^d \langle x_*, e_i\rangle^2 i^{\beta\alpha} = \sum_{i=1}^d i^{-2\delta+\alpha\beta}$, which is bounded independently of the dimension d if and only if $\sum_{i=1}^\infty i^{-2\delta+\alpha\beta} < \infty \Leftrightarrow -2\delta + \alpha\beta < -1 \Leftrightarrow \alpha < \frac{2\delta-1}{\beta}$.

(b) $\mathrm{Tr}(\Sigma^{1-\alpha}) = \sum_{i=1}^d i^{-\beta(1-\alpha)}$, which is bounded independently of the dimension d if and only if $\sum_{i=1}^\infty i^{-\beta(1-\alpha)} < \infty \Leftrightarrow -\beta(1-\alpha) < -1 \Leftrightarrow \alpha < 1 - 1/\beta$.

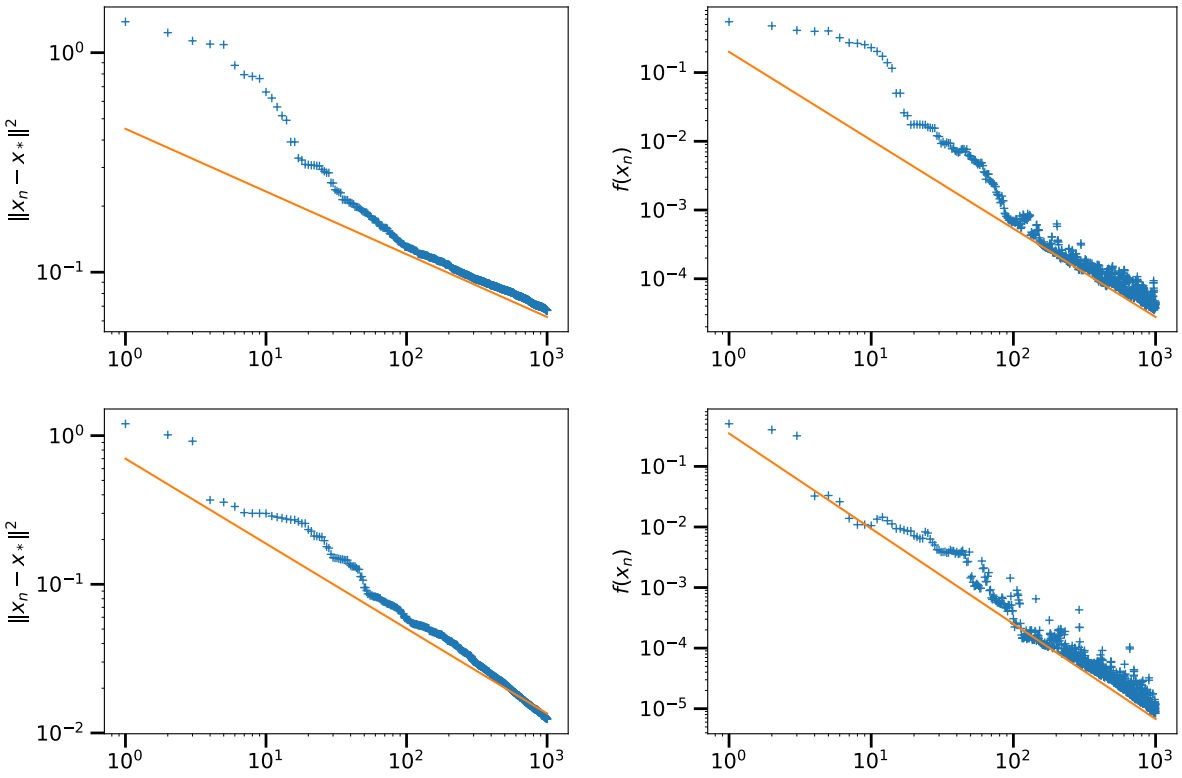


FIGURE 2.3. In blue +, evolution of $\|x_n - x_*\|^2$ (left) and $f(x_n)$ (right) as functions of n, for the problems with parameters $\beta = 1.4, \delta = 1.2$ (up) and $\beta = 3.5, \delta = 1.5$. The orange lines represent the curves D/n^{α_*} (left) and D'/n^{α_*+1} (right).

Thus the corollary gives dimension-independent convergence rates for all $\alpha < \alpha_* = \min\left(1 - \frac{1}{\beta}, \frac{2\delta-1}{\beta}\right)$.

In Figure 2.3, we show the evolution of $\|x_n - x_*\|^2$ and $f(x_n)$ for two realizations of SGD. We chose the step-size $\gamma = 1/R_0 = 1/(2\|\Sigma\|_{\mathcal{H}\to\mathcal{H}} + \mathrm{Tr}(\Sigma))$. The two realizations represent two possible different regimes:

- In the two upper plots, $\beta = 1.4, \delta = 1.2$. The irregularity of the feature vectors is the bottleneck for fast convergence. We have $\alpha_* = \min\left(1 - \frac{1}{\beta}, \frac{2\delta-1}{\beta}\right) \approx \min(0.29, 1) = 0.29$.
- In the two lower plots, $\beta = 3.5, \delta = 1.5$. The irregularity of the optimum is the bottleneck for fast convergence. We have $\alpha_* = \min\left(1 - \frac{1}{\beta}, \frac{2\delta-1}{\beta}\right) \approx \min(0.71, 0.57) = 0.57$.

We compare with the curves D/n^{α_*} and D'/n^{α_*+1} with hand-tuned constants D and D' to fit best the data for each plot. In both regimes, our theory is sharp in predicting the exponents in the polynomial rates of convergence of $\|x_n - x_*\|^2$ and $f(x_n)$.

In Figure 2.4, we show how these simulations are perturbed in the presence of additive noise. We consider the noisy linear model $b = \langle x_*, a\rangle + z$, where $a \sim \mathcal{N}(0, \Sigma)$ and $z \sim \mathcal{N}(0, \sigma^2)$ are independent. Here, we consider the case $d = 10^5$, $\beta = 1.4$, $\delta = 1.2$. In the noiseless case $\sigma^2 = 0$, we have shown that the rate of convergence was given by the polynomial exponent $\alpha_* = \min\left(1 - \frac{1}{\beta}, \frac{2\delta-1}{\beta}\right)$. These predicted rates are represented by the orange lines in the plots. In blue, we show the results of our simulations with some additive noise with variance $\sigma^2 = 2 \times 10^{-4}$. The exponent α_* still describes the behavior of SGD in the initial phase, but in the large n asymptotic the population risk $f(x_n)$

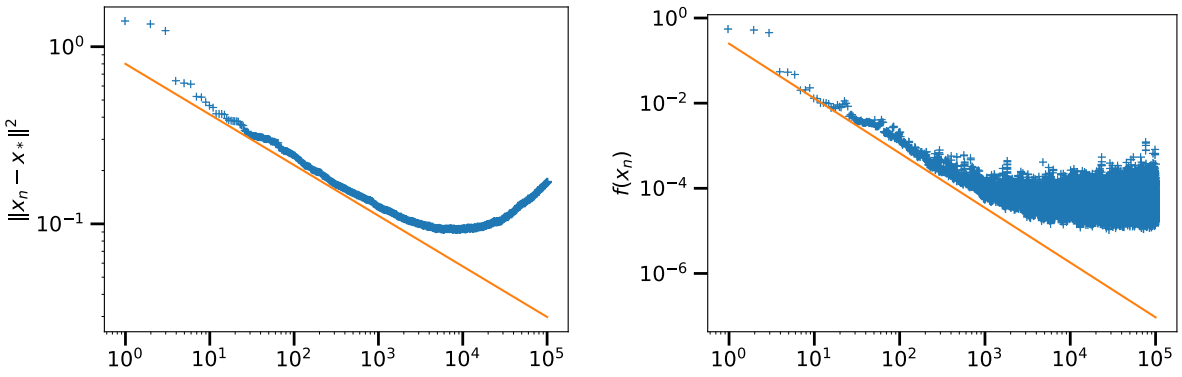


FIGURE 2.4. In blue +, evolution of $\|x_n - x_*\|^2$ (left) and $f(x_n)$ (right) as functions of n, for the problems with parameters $d = 10^5, \beta = 1.4, \delta = 1.2$. The orange lines represent the curves D/n^{α_*} (left) and D'/n^{α_*+1} (right).

stagnates around the order of σ^2. Both of these qualitative behaviors are predicted by Theorem 2.7. Moreover, the reconstruction error $\|x_n - x_*\|^2$ diverges for large n.

Appendix of Chapter 2

This appendix contains the proofs of the results of the chapter. We start by proving with the upper bounds. In Appendix 2.A, we prove the upper bounds in the noiseless case (Theorems 2.1, 2.2 and 2.4); in Appendix 2.B, we continue with the upper bounds in the noisy case (Theorems 2.6, 2.7 and 2.8). We then continue with the lower bounds (Theorems 2.3 and 2.5) in Appendix 2.C. Finally, we finish with the proof of Corollary 2.1 in Appendix 2.D.

2.A. Proof of Theorems 2.1, 2.2 and 2.4

We recall here the definition of the regularity functions

$$d_n^2(\beta) = \mathbb{E}\left[\left\langle x_n - x_*, \Sigma^{-\beta}\left(x_n - x_*\right)\right\rangle\right] \in [0,\infty]\,, \qquad \beta \in \mathbb{R}\,.$$

2.A.1. Properties of the regularity functions. We derive here two properties of the sequence of regularity functions $d_n^2, n \geqslant 1$ that are useful for the proof of Theorem 2.4. The first one is a simple consequence of the above definition of the regularity function. The second property is the closed recurrence relation of the regularity functions d_n^2, $n \geqslant 0$ associated to the iterates of SGD.

Property 1. For all n, the function d_n^2 is log-convex, i.e., for all $\beta_1, \beta_2 \in \mathbb{R}$, for all $\lambda \in [0,1]$,

$$d_n^2\left((1-\lambda)\beta_1 + \lambda\beta_2\right) \leqslant d_n^2(\beta_1)^{1-\lambda} d_n^2(\beta_2)^{\lambda}\,.$$

Proof. The proof is based on the following lemma, that we state clearly for another use below.

Lemma 2.2. Let $x \in \mathcal{H}$. Then for all $\beta_1, \beta_2 \in \mathbb{R}$, $\lambda \in [0,1]$,

$$\left\langle x, \Sigma^{-[(1-\lambda)\beta_1 + \lambda\beta_2]} x\right\rangle \leqslant \left\langle x, \Sigma^{-\beta_1} x\right\rangle^{1-\lambda} \left\langle x, \Sigma^{-\beta_2} x\right\rangle^{\lambda}\,.$$

This lemma follows from Hölder's inequality with $p = (1-\lambda)^{-1}$ and $q = \lambda^{-1}$. Indeed, diagonalize $\Sigma = \sum_i \mu_i e_i \otimes e_i$. Then

$$\begin{aligned}
\left\langle x, \Sigma^{-[(1-\lambda)\beta_1 + \lambda\beta_2]} x\right\rangle &= \sum_i \mu_i^{-[(1-\lambda)\beta_1 + \lambda\beta_2]} \langle x, e_i\rangle^2 \\
&= \sum_i \left(\mu_i^{-\beta_1}\langle x, e_i\rangle^2\right)^{1-\lambda} \left(\mu_i^{-\beta_2}\langle x, e_i\rangle^2\right)^{\lambda} \\
&\leqslant \left(\sum_i \mu_i^{-\beta_1}\langle x, e_i\rangle^2\right)^{1-\lambda} \left(\sum_i \mu_i^{-\beta_2}\langle x, e_i\rangle^2\right)^{\lambda} \\
&= \left\langle x, \Sigma^{-\beta_1} x\right\rangle^{1-\lambda} \left\langle x, \Sigma^{-\beta_2} x\right\rangle^{\lambda}\,.
\end{aligned}$$

We now apply this lemma to prove Property 1.

$$d_n^2((1-\lambda)\beta_1 + \lambda\beta_2) = \mathbb{E}\left[\left\langle x_n - x_*, \Sigma^{-[(1-\lambda)\beta_1 + \lambda\beta_2]}\left(x_n - x_*\right)\right\rangle\right]$$

$$\leqslant \mathbb{E}\left[\left\langle x_n - x_*, \Sigma^{-\beta_1}(x_n - x_*)\right\rangle^{1-\lambda}\left\langle x_n - x_*, \Sigma^{-\beta_2}(x_n - x_*)\right\rangle^{\lambda}\right].$$

Using again Hölder's inequality, we get

$$\begin{aligned} d_n^2((1-\lambda)\beta_1 + \lambda\beta_2) &\leqslant \mathbb{E}\left[\left\langle x_n - x_*, \Sigma^{-\beta_1}(x_n - x_*)\right\rangle\right]^{1-\lambda}\mathbb{E}\left[\left\langle x_n - x_*, \Sigma^{-\beta_2}(x_n - x_*)\right\rangle\right]^{\lambda} \\ &= d_n^2(\beta_1)^{1-\lambda} d_n^2(\beta_2)^{\lambda}. \end{aligned}$$

□

Property 2. Under the assumptions of Theorem 2.4, for all n, the function d_n^2 is finite on $(-\infty, \underline{\alpha}]$, and if $0 \leqslant \beta \leqslant \underline{\alpha}$,

$$d_n^2(\beta) \leqslant d_{n-1}^2(\beta) - 2\gamma d_{n-1}^2(\beta-1) + \gamma^2 R_0^{1-\beta/\underline{\alpha}} R_{\underline{\alpha}}^{\beta/\underline{\alpha}} d_{n-1}^2(-1).$$

Proof. By assumption (a), $d_0^2(\underline{\alpha}) = \|\Sigma^{-\underline{\alpha}/2}(x_0 - x_*)\|^2$ is finite, i.e., there exists $x \in \mathcal{H}$ such that $x_0 - x_* = \Sigma^{\underline{\alpha}/2}x$. Then for any $\beta \leqslant \underline{\alpha}$, $x_0 - x_* = \Sigma^{\beta/2}\left(\Sigma^{(\underline{\alpha}-\beta)/2}(x_0 - x_*)\right)$ thus $d_0^2(\beta) = \|\Sigma^{-\beta/2}(x_0 - x_*)\|^2$ is finite.

Further, assume that for some n, the function d_{n-1}^2 is finite on $(\infty, \underline{\alpha}]$. As we are in the noiseless linear case $b = \langle x_*, a\rangle$, we can rewrite the stochastic gradient iteration (2.2) as

$$x_n - x_* = (\mathrm{Id} - a_n \otimes a_n)(x_{n-1} - x_*).$$

Substituting this expression in the definition of d_n^2 and expanding the formula, we get

$$\begin{aligned} d_n^2(\beta) &= \mathbb{E}\left[\left\langle x_n - x_*, \Sigma^{-\beta}(x_n - x_*)\right\rangle\right] \\ &= \mathbb{E}\left[\left\langle (\mathrm{Id} - \gamma a_n \otimes a_n)(x_{n-1} - x_*), \Sigma^{-\beta}(\mathrm{Id} - \gamma a_n \otimes a_n)(x_{n-1} - x_*)\right\rangle\right] \\ &= \mathbb{E}\left[\left\langle x_{n-1} - x_*, \Sigma^{-\beta}(x_{n-1} - x_*)\right\rangle\right] - 2\gamma\mathbb{E}\left[\left\langle x_{n-1} - x_*, \Sigma^{-\beta} a_n \otimes a_n (x_{n-1} - x_*)\right\rangle\right] \quad (2.7) \\ &\qquad + \gamma^2\mathbb{E}\left[\left\langle x_{n-1} - x_*, a_n \otimes a_n \Sigma^{-\beta} a_n \otimes a_n (x_{n-1} - x_*)\right\rangle\right]. \quad (2.8) \end{aligned}$$

Note that the first term of this sum is $d_{n-1}^2(\beta)$. Further, θ_{n-1} is computed using only $(a_1, b_1), \ldots, (a_{n-1}, b_{n-1})$, thus it is independent of a_n. It follows that

$$\begin{aligned} \mathbb{E}\left[\left\langle x_{n-1} - x_*, \Sigma^{-\beta} a_n \otimes a_n (x_{n-1} - x_*)\right\rangle\right] &= \mathbb{E}\left[\left\langle x_{n-1} - x_*, \Sigma^{-\beta}\mathbb{E}\left[a_n \otimes a_n\right](x_{n-1} - x_*)\right\rangle\right] \\ &= \mathbb{E}\left[\left\langle x_{n-1} - x_*, \Sigma^{-\beta+1}(x_{n-1} - x_*)\right\rangle\right] \\ &= d_{n-1}^2(\beta - 1). \quad (2.9) \end{aligned}$$

Finally,

$$\mathbb{E}\left[\left\langle x_{n-1} - x_*, a_n \otimes a_n \Sigma^{-\beta} a_n \otimes a_n (x_{n-1} - x_*)\right\rangle\right] = \mathbb{E}\left[\langle x_{n-1} - x_*, a_n\rangle^2\left\langle a_n, \Sigma^{-\beta} a_n\right\rangle\right] \quad (2.10)$$

We now assume that $0 \leqslant \beta \leqslant \underline{\alpha}$. We apply Lemma 2.2 with $\beta_1 = 0, \beta_2 = \underline{\alpha}, \lambda = \beta/\underline{\alpha}$:

$$\left\langle a_n, \Sigma^{-\beta} a_n\right\rangle \leqslant \|a_n\|^{2(1-\beta/\underline{\alpha})}\langle a_n, \Sigma^{-\underline{\alpha}} a_n\rangle^{\beta/\underline{\alpha}}$$

Let $\mathbb{E}_{a_n}$ denote the expectation with respect to a_n only, while keeping $a_0, \ldots, a_{n-1}$ random. Applying Hölder's inequality, we get

$$\begin{aligned} &\mathbb{E}_{a_n}\left[\left\langle a_n, \Sigma^{-\beta} a_n\right\rangle\langle x_{n-1} - x_*, a_n\rangle^2\right] \\ &\quad \leqslant \mathbb{E}_{a_n}\left[\|a_n\|^{2(1-\beta/\underline{\alpha})}\langle a_n, \Sigma^{-\underline{\alpha}} a_n\rangle^{\beta/\underline{\alpha}}\langle x_{n-1} - x_*, a_n\rangle^2\right] \end{aligned}$$

$$
\begin{aligned}
&\leqslant \mathbb{E}_{a_n}\left[\|a_n\|^2 \langle x_{n-1}-x_*, a_n\rangle^2\right]^{1-\beta/\underline{\alpha}} \mathbb{E}\left[\langle a_n, \Sigma^{-\underline{\alpha}} a_n\rangle \langle x_{n-1}-x_*, a_n\rangle^2\right]^{\beta/\underline{\alpha}} \\
&= \left\langle x_{n-1}-x_*, \mathbb{E}\left[\|a_n\|^2 a_n\otimes a_n\right](x_{n-1}-x_*)\right\rangle^{1-\beta/\underline{\alpha}} \\
&\qquad \times \left\langle x_{n-1}-x_*, \mathbb{E}\left[\langle a_n, \Sigma^{-\underline{\alpha}} a_n\rangle a_n\otimes a_n\right](x_{n-1}-x_*)\right\rangle^{\beta/\underline{\alpha}} \\
&\leqslant R_0^{1-\beta/\underline{\alpha}} R_{\underline{\alpha}}^{\beta/\underline{\alpha}} \langle x_{n-1}-x_*, \Sigma(x_{n-1}-x_*)\rangle \ ,
\end{aligned}
$$

where in this last step, we use the assumptions that the features a are bounded and regular, in their weak formulation of Remark 2.4. Returning to the computation of (2.10), we get

$$
\begin{aligned}
\mathbb{E}\left[\left\langle x_{n-1}-x_*, a_n\otimes a_n\Sigma^{-\beta}a_n\otimes a_n(x_{n-1}-x_*)\right\rangle\right] &= \mathbb{E}\left[\mathbb{E}_{a_n}\left[\langle x_{n-1}-x_*, a_n\rangle^2\left\langle a_n, \Sigma^{-\beta}a_n\right\rangle\right]\right] \\
&\leqslant R_0^{1-\beta/\underline{\alpha}} R_{\underline{\alpha}}^{\beta/\underline{\alpha}} \mathbb{E}\left[\langle x_{n-1}-x_*, \Sigma(x_{n-1}-x_*)\rangle\right] \\
&= R_0^{1-\beta/\underline{\alpha}} R_{\underline{\alpha}}^{\beta/\underline{\alpha}} d_{n-1}^2(-1)\,. \qquad (2.11)
\end{aligned}
$$

The result is obtained by putting together Equations (2.7)-(2.8), (2.9) and (2.11). □

2.A.2. Proof of the theorems. A remarkable feature of the proofs that follow is that only Properties 1 and 2 of the regularity functions are used to derive the theorems. In particular, we do not use the definition of the regularity functions d_n^2 in this section.

We start with a few preliminary remarks. Using the recurrence Property 2 and that $\gamma R_0 \leqslant 1$,

$$
\begin{aligned}
d_k^2(0) &\leqslant d_{k-1}^2(0) - \gamma\,(2-\gamma R_0)\, d_{k-1}^2(-1) \\
&\leqslant d_{k-1}^2(0) - \gamma d_{k-1}^2(-1)\,. \qquad (2.12)
\end{aligned}
$$

Proof of Theorem 2.1. As a quick aside, we prove Theorem 2.1. In this paragraph, we assume that f is μ-strongly convex. Thus

$$
d_{k-1}^2(-1) = 2\mathbb{E}f(x_{k-1}) \geqslant \mu\mathbb{E}\left[\|x_{k-1}-x_*\|^2\right] = \mu d_{k-1}^2(0)\,.
$$

Then from (2.12), we obtain

$$
d_k^2(0) \leqslant (1-\mu\gamma) d_{k-1}^2(0) \leqslant \ldots \leqslant (1-\mu\gamma)^k d_0^2(0)\,.
$$

This is the statement of Theorem 2.1.

Proof of Theorem 2.3. We now return to the proof of the non-parametric theorems. We do not assume strong convexity anymore but source and capacity conditions. From (2.12), the sequence $d_k^2(0)$, $k \geqslant 0$ decreases, and,

$$
\gamma d_{k-1}^2(-1) \leqslant d_{k-1}^2(0) - d_k^2(0)\,. \qquad (2.13)
$$

By summing this inequality over $k \geqslant 1$, we get

$$
\gamma \sum_{k=0}^{\infty} d_k^2(-1) \leqslant d_0^2(0)\,. \qquad (2.14)
$$

Using again the recurrence Property 2,

$$
\begin{aligned}
d_k^2(\underline{\alpha}) &\leqslant d_{k-1}^2(\underline{\alpha}) - 2\gamma d_{k-1}^2(\underline{\alpha}-1) + \gamma^2 R_{\underline{\alpha}} d_{k-1}^2(-1) \qquad (2.15) \\
&\leqslant d_{k-1}^2(\underline{\alpha}) + \gamma^2 R_{\underline{\alpha}} d_{k-1}^2(-1)\,.
\end{aligned}
$$

By summing for $k = 1, \ldots, n$ and using the bound (2.14),

$$
d_n^2(\underline{\alpha}) \leqslant d_0^2(\underline{\alpha}) + \gamma^2 R_{\underline{\alpha}} \sum_{k=0}^{n-1} d_k^2(-1)
$$

$$\leqslant d_0^2(\underline{\alpha}) + \gamma R_{\underline{\alpha}} d_0^2(0)$$
$$\leqslant d_0^2(\underline{\alpha}) + \frac{R_{\underline{\alpha}}}{R_0} d_0^2(0)\,. \tag{2.16}$$

In words, the sequence $d_n^2(\underline{\alpha})$, $n \geqslant 0$ is bounded by $D := d_0^2(\underline{\alpha}) + \frac{R_{\underline{\alpha}}}{R_0} d_0^2(0)$. As a side note, this proves Theorem 2.4 for $\beta = \underline{\alpha}$.

We can now give a closed recurrence relation $d_k^2(0)$, $k \geqslant 0$. Using the log-convexity Property 1,

$$d_{k-1}^2(0) \leqslant d_{k-1}^2(-1)^{\underline{\alpha}/(\underline{\alpha}+1)} d_{k-1}^2(\underline{\alpha})^{1/(\underline{\alpha}+1)} \leqslant d_{k-1}^2(-1)^{\underline{\alpha}/(\underline{\alpha}+1)} D^{1/(\underline{\alpha}+1)}\,. \tag{2.17}$$

Substituting in (2.13), we obtain

$$d_{k-1}^2(0) - d_k^2(0) \geqslant \gamma d_{k-1}^2(-1)$$
$$\geqslant \gamma D^{-1/\underline{\alpha}} d_{k-1}^2(0)^{1+1/\underline{\alpha}}\,.$$

This gives the wanted closed recurrence relation for $d_k^2(0)$, $k \geqslant 0$. It implies a decay of $d_k^2(0)$ as follows: consider the real function $f(\lambda) = \frac{1}{\lambda^{1/\underline{\alpha}}}$. It is a convex function on the positive reals, with derivative $f'(\lambda) = -\frac{1}{\underline{\alpha}} \frac{1}{\lambda^{1+1/\underline{\alpha}}}$. Using that a convex function is above its tangents, we obtain

$$f\left(d_k^2(0)\right) - f\left(d_{k-1}^2(0)\right) \geqslant f'\left(d_{k-1}^2(0)\right)\left(d_k^2(0) - d_{k-1}^2(0)\right)$$
$$= -\frac{1}{\underline{\alpha}} \frac{1}{d_{k-1}^2(0)^{1+1/\underline{\alpha}}} \left(d_k^2(0) - d_{k-1}^2(0)\right)$$
$$\geqslant \frac{1}{\underline{\alpha}} \gamma D^{-1/\underline{\alpha}}\,.$$

By summing this inequality for $k = 1, \dots, n$, we obtain

$$\frac{1}{d_n^2(0)^{1/\underline{\alpha}}} = f\left(d_n^2(0)\right) \geqslant f\left(d_0^2(0)\right) + \frac{1}{\underline{\alpha}} \gamma D^{-1/\underline{\alpha}} n \geqslant \frac{1}{\underline{\alpha}} \gamma D^{-1/\underline{\alpha}} n\,.$$

This implies the bound of the reconstruction error of SGD in Theorem 2.2:

$$\mathbb{E}\left[\|x_n - x_*\|^2\right] = d_n^2(0) \leqslant C \frac{1}{n^{\underline{\alpha}}}\,, \qquad C = \frac{\underline{\alpha}^{\underline{\alpha}}}{\gamma^{\underline{\alpha}}} D\,. \tag{2.18}$$

The corresponding bound for tail-averaged SGD follows easily: using convexity and that $d_n^2(0) = \mathbb{E}\|x_n - x_*\|^2$ is decreasing, we obtain

$$\mathbb{E}\|\overline{x}_n - x_*\|^2 \leqslant \frac{1}{\lfloor n/2 \rfloor + 1} \sum_{k=\lceil n/2 \rceil}^{n} \mathbb{E}\|x_k - x_*\|^2 \leqslant \mathbb{E}\|x_{\lceil n/2 \rceil} - x_*\|^2 \leqslant \frac{C}{\lceil n/2 \rceil^{\underline{\alpha}}} \leqslant \frac{2^{\underline{\alpha}} C}{n^{\underline{\alpha}}}\,.$$

We now turn to the study of the generalization errors. We have

$$\frac{1}{\lfloor n/2 \rfloor + 1} \sum_{k=\lceil n/2 \rceil}^{n} d_k^2(-1) \leqslant \frac{2}{n} \frac{1}{\gamma} \sum_{k=\lceil n/2 \rceil}^{n} \left(d_k^2(0) - d_{k+1}^2(0)\right)\,,$$

where in the last step we used (2.13). Telescoping the sum, we obtain

$$\frac{1}{\lfloor n/2 \rfloor + 1} \sum_{k=\lceil n/2 \rceil}^{n} d_k^2(-1) \leqslant \frac{2}{n} \frac{1}{\gamma} \frac{\underline{\alpha}^{\underline{\alpha}}}{\gamma^{\underline{\alpha}}} D \frac{1}{\lceil n/2 \rceil^{\underline{\alpha}}} \leqslant 2^{\underline{\alpha}+1} \frac{\underline{\alpha}^{\underline{\alpha}}}{\gamma^{\underline{\alpha}+1}} D \frac{1}{n^{\underline{\alpha}+1}}\,.$$

We now use that

$$2 \min_{0 \leqslant k \leqslant n} \mathbb{E} f(x_n) \leqslant \min_{0 \leqslant k \leqslant n} d_k^2(-1) \leqslant \min_{\lceil n/2 \rceil \leqslant k \leqslant n} d_k^2(-1) \leqslant \frac{1}{\lfloor n/2 \rfloor + 1} \sum_{k=\lceil n/2 \rceil}^{n} d_k^2(-1)\,.$$

Combining the two last inequalities, we obtain the claimed generalization bound for SGD. Similarly, using the convexity of f,

$$2\mathbb{E}f(\overline{x}_n) \leqslant \frac{2}{\lfloor n/2 \rfloor + 1} \sum_{k=\lceil n/2 \rceil}^{n} \mathbb{E}f(x_k) \leqslant \frac{1}{\lfloor n/2 \rfloor + 1} \sum_{k=\lceil n/2 \rceil}^{n} d_k^2(-1)\,.$$

We thus obtain the same generalization bound for tail-averaged SGD.

Generalization: proof of Theorem 2.4. We continue the proof of Theorem 2.2 to prove Theorem 2.4. By the log-convexity Property 1, for all $\beta \in [0, \underline{\alpha}]$,

$$d_n^2(\beta) \leqslant d_n^2(0)^{1-\beta/\underline{\alpha}} d_n^2(\underline{\alpha})^{\beta/\underline{\alpha}}\,.$$

Using Equations (2.18) and (2.16), we obtain

$$d_n^2(\beta) \leqslant \frac{\underline{\alpha}^{\underline{\alpha}-\beta}}{\gamma^{\underline{\alpha}-\beta}} D \frac{1}{n^{\underline{\alpha}-\beta}}\,.$$

This proves conclusion (1) of the theorem, for SGD. In the case of tail-average SGD, we have by convexity

$$\begin{aligned}\overline{d}_n^2(\beta) &\leqslant \frac{1}{\lfloor n/2 \rfloor + 1} \sum_{k=\lceil n/2 \rceil}^{n} d_k^2(\beta) \leqslant \frac{1}{\lfloor n/2 \rfloor + 1} \sum_{k=\lceil n/2 \rceil}^{n} \frac{\underline{\alpha}^{\underline{\alpha}-\beta}}{\gamma^{\underline{\alpha}-\beta}} D \frac{1}{k^{\underline{\alpha}-\beta}} \leqslant \frac{\underline{\alpha}^{\underline{\alpha}-\beta}}{\gamma^{\underline{\alpha}-\beta}} D \frac{1}{\lceil n/2 \rceil^{\underline{\alpha}-\beta}} \\ &\leqslant \frac{2^{\underline{\alpha}-\beta}\underline{\alpha}^{\underline{\alpha}-\beta}}{\gamma^{\underline{\alpha}-\beta}} D \frac{1}{n^{\underline{\alpha}-\beta}}\,.\end{aligned}$$

The bound for tail-averaged SGD and $\beta \in [0, \underline{\alpha}]$ follows.

We now consider the case $\beta \in [-1, 0)$. Using the log-convexity property and Hölder inequality, we have

$$\sum_{k=\lceil n/2 \rceil}^{n} d_k(\beta) \leqslant \sum_{k=\lceil n/2 \rceil}^{n} d_k^2(-1)^{-\beta} d_k^2(0)^{1+\beta} \leqslant \left(\sum_{k=\lceil n/2 \rceil}^{n} d_k^2(-1) \right)^{-\beta} \left(\sum_{k=\lceil n/2 \rceil}^{n} d_k^2(0) \right)^{1+\beta}\,.$$

Using, again, the telescoping from bound (2.13), we obtain

$$\sum_{k=\lceil n/2 \rceil}^{n} d_k(\beta) \leqslant \left(\gamma^{-1} d_{\lceil n/2 \rceil}^2(0) \right)^{-\beta} \left((\lfloor n/2 \rfloor + 1) d_{\lceil n/2 \rceil}^2(0) \right)^{1+\beta} = \gamma^{\beta} (\lfloor n/2 \rfloor + 1)^{1+\beta} d_{\lceil n/2 \rceil}^2(0)\,.$$

Thus, combining with Theorem 2.2, we obtain

$$\frac{1}{\lfloor n/2 \rfloor + 1} \sum_{k=\lceil n/2 \rceil}^{n} d_k(\beta) \leqslant \gamma^{\beta} (\lfloor n/2 \rfloor + 1)^{\beta} \frac{C}{\lceil n/2 \rceil^{\underline{\alpha}}} \leqslant \gamma^{\beta} 2^{\underline{\alpha}-\beta} \frac{C}{n^{\underline{\alpha}-\beta}}\,.$$

As before, one can use this bound to derive a bound on $d_k(\beta)$ for the best of the past iterates, or for the tail-averaged quantity $\overline{d}_n(\beta)$. These are the presented bounds.

2.B. Proof of Theorems 2.6, 2.7 and 2.8

Note that in this proof, we use the strong assumptions of regularity of the feature vector a. We do not know whether it is possible to prove the same result under the weak assumptions of Remark 2.4.

Our proof stategy is the following: we decompose the SGD iterates sequence x_n as a sum of sequences $x_n = \nu_n + \sum_{l=1}^{n} \eta_n^{(l)}$, where each of the auxiliary sequences is interpreted as the iterates of some SGD iteration under a noiseless linear model. We thus apply the results of Section 2.1 to control these auxiliary sequences and obtain the presented bound.

Define $z_n = b_n - \langle x_*, a_n \rangle$, the error of the best linear estimator. Then Equation (2.2) can be rewritten as

$$x_n = x_{n-1} - \gamma \langle x_{n-1} - x_*, a_n \rangle a_n + \gamma z_n a_n \,.$$

We see this iteration as an additively perturbed version of the iteration

$$\nu_0 = x_0 \,, \qquad\qquad \nu_n = \nu_{n-1} - \gamma \langle \nu_{n-1} - x_*, a_n \rangle a_n \,,$$

studied in Section 2.1. To understand the effect of the additive noise, define for all $l \geqslant 1$,

$$\eta_l^{(l)} = \gamma z_l a_l \,, \qquad\qquad \eta_n^{(l)} = \eta_{n-1}^{(l)} - \gamma \langle \eta_{n-1}^{(l)}, a_n \rangle a_n \,, \qquad n > l \,.$$

Then

$$x_n = \nu_n + \sum_{l=1}^{n} \eta_n^{(l)} \,. \tag{2.19}$$

Indeed, this last equation is checked by induction: $x_0 = \nu_0$, and if the equation is satisfied for some $n \geqslant 0$,

$$\begin{aligned}
x_{n+1} &= x_n - \gamma \langle x_n - x_*, a_{n+1} \rangle a_{n+1} + \gamma z_{n+1} a_{n+1} \\
&= \nu_n + \sum_{l=1}^{n} \eta_n^{(l)} - \gamma \left\langle \nu_n + \sum_{l=1}^{n} \eta_n^{(l)} - x_*, a_{n+1} \right\rangle a_{n+1} + \eta_{n+1}^{(n+1)} \\
&= [\nu_n - \gamma \langle \nu_n - x_*, a_{n+1} \rangle a_{n+1}] + \sum_{l=1}^{n} \left[\eta_n^{(l)} - \gamma \langle \eta_n^{(l)}, a_{n+1} \rangle a_{n+1} \right] + \eta_{n+1}^{(n+1)} \\
&= \nu_{n+1} + \sum_{l=1}^{n} \eta_{n+1}^{(l)} + \eta_{n+1}^{(n+1)} \,.
\end{aligned}$$

We use the decomposition (2.19) to study $d_n^2(\beta)$. Using the triangle inequality,

$$\begin{aligned}
d_n^2(\beta) &= \mathbb{E} \left[\left\| \Sigma^{-\beta/2} \left(\nu_n + \sum_{l=1}^{n} \eta_n^{(l)} - x_* \right) \right\|^2 \right] \\
&\leqslant \mathbb{E} \left[\left(\left\| \Sigma^{-\beta/2} \left(\nu_n - x_* \right) \right\| + \left\| \Sigma^{-\beta/2} \sum_{l=1}^{n} \eta_n^{(l)} \right\| \right)^2 \right] \\
&\leqslant 2\mathbb{E} \left[\left\| \Sigma^{-\beta/2} (\nu_n - x_*) \right\|^2 \right] + 2\mathbb{E} \left[\left\| \Sigma^{-\beta/2} \sum_{l=1}^{n} \eta_n^{(l)} \right\|^2 \right]
\end{aligned} \tag{2.20}$$

The first term is studied in Section 2.1. We detail the analysis of the second term. Note that

$$\begin{aligned}
\eta_n^{(l)} &= (I - \gamma a_n \otimes a_n) \eta_{n-1}^{(l)} = \cdots = (I - \gamma a_n \otimes a_n) \cdots (I - \gamma a_{l+1} \otimes a_{l+1}) \eta_l^{(l)} \\
&= (I - \gamma a_n \otimes a_n) \cdots (I - \gamma a_{l+1} \otimes a_{l+1}) \gamma z_l a_l \,.
\end{aligned} \tag{2.21}$$

Thus if $l < l'$,

$$\begin{aligned}
\mathbb{E} \left[\left\langle \eta_n^{(l)}, \Sigma^{-\beta} \eta_n^{(l')} \right\rangle \right] &= \mathbb{E} \left[\left\langle \mathbb{E} \left[\eta_n^{(l)} \middle| a_{l+1}, \ldots, a_n \right], \Sigma^{-\beta} \eta_n^{(l')} \right\rangle \right] \\
&= \mathbb{E} \left[\left\langle (I - \gamma a_n \otimes a_n) \cdots (I - \gamma a_{l+1} \otimes a_{l+1}) \gamma \mathbb{E}[z_l a_l], \Sigma^{-\beta} \eta_n^{(l')} \right\rangle \right]
\end{aligned}$$

Note that by definition of x_*, $0 = \nabla f(x_*) = -\mathbb{E}\left[(b_l - \langle x_*, a_l\rangle)a_l\right] = -\mathbb{E}\left[z_l a_l\right]$ thus we obtain that the cross products $\mathbb{E}\left[\left\langle \eta_n^{(l)}, \Sigma^{-\beta}\eta_n^{(l')}\right\rangle\right]$ are zero. This gives

$$\mathbb{E}\left[\left\|\Sigma^{-\beta/2}\sum_{l=1}^{n}\eta_n^{(l)}\right\|^2\right] = \sum_{l=1}^{n}\mathbb{E}\left[\left\|\Sigma^{-\beta/2}\eta_n^{(l)}\right\|^2\right].$$

Note that from Equation (2.21), $\eta_n^{(l)}$ and $\eta_{n-l+1}^{(1)}$ are equal in law. Thus

$$\mathbb{E}\left[\left\|\Sigma^{-\beta/2}\sum_{l=1}^{n}\eta_n^{(l)}\right\|^2\right] = \sum_{l=1}^{n}\mathbb{E}\left[\left\|\Sigma^{-\beta/2}\eta_{n-l+1}^{(1)}\right\|^2\right] = \sum_{l=1}^{n}\mathbb{E}\left[\left\|\Sigma^{-\beta/2}\eta_l^{(1)}\right\|^2\right]. \tag{2.22}$$

This last quantity is the sum of the expected squared power norms

$$\tilde{d}_l^2(\beta) := \mathbb{E}\left[\left\|\Sigma^{-\beta/2}\eta_l^{(1)}\right\|^2\right]$$

of the SGD iterates $\eta_l^{(1)}, l \geqslant 1$ on a noiseless linear model, with initialization $\eta_1^{(1)} = \gamma z_1 a_1$. We now divide the discussion for the parametric and the non-parametric cases.

Proof of Theorem 2.6. We assume that f is μ-strongly convex. We apply Theorem 2.1 to the iterates ν_n and $\eta_l^{(1)}$. We obtain:

$$\mathbb{E}\|\nu_n - x_*\|^2 \leqslant (1-\gamma\mu)^n\|\nu_0 - x_*\|^2,$$
$$\mathbb{E}\|\eta_l^{(1)}\|^2 \leqslant (1-\gamma\mu)^{l-1}\mathbb{E}\|\eta_1^{(1)}\|^2.$$

From the second equation,

$$\begin{aligned}\sum_{l=1}^{n}\mathbb{E}\left[\left\|\eta_l^{(1)}\right\|^2\right] &\leqslant \left(\sum_{l=1}^{n}(1-\gamma\mu)^{l-1}\right)\mathbb{E}\left[\|\eta_1^{(1)}\|^2\right] \leqslant \gamma^2\left(\sum_{l=1}^{\infty}(1-\gamma\mu)^{l-1}\right)\mathbb{E}\left[\|z_1 a_1\|^2\right]\\ &\leqslant \frac{\gamma}{\mu}R_0\mathbb{E}[z_1^2] = \frac{2\gamma}{\mu}R_0 f(x_*).\end{aligned}$$

Combining (2.20), (2.22) (in the case $\beta = 0$) and the last equations, we obtain

$$\begin{aligned}\mathbb{E}\|x_n - x_*\|^2 = d_n^2(0) &\leqslant 2\mathbb{E}\|\nu_n - x_*\|^2 + 2\sum_{l=1}^{n}\mathbb{E}\left[\left\|\eta_l^{(1)}\right\|^2\right]\\ &\leqslant (1-\gamma\mu)^n\|\nu_0 - x_*\|^2 + \frac{4\gamma}{\mu}R_0 f(x_*).\end{aligned}$$

This proves Theorem 2.6.

Proof of Theorems 2.7 and 2.8. We now continue in the non-parametric case. In this case, when $\beta = -1$, the control of (2.22) is given by (2.14): with our notation here, this gives

$$\sum_{l=1}^{n}\tilde{d}_l^2(-1) \leqslant \sum_{l=1}^{\infty}\tilde{d}_l^2(-1) \leqslant \frac{1}{\gamma}\tilde{d}_1^2(0). \tag{2.23}$$

When $\beta = \underline{\alpha} - 1$, a similar control can be obtained from (2.15) which gives:

$$2\gamma\tilde{d}_{l-1}^2(\underline{\alpha}-1) \leqslant \tilde{d}_{l-1}^2(\underline{\alpha}) - \tilde{d}_l^2(\underline{\alpha}) + \gamma^2 R_{\underline{\alpha}}\tilde{d}_{l-1}^2(-1).$$

By summing these inequalities for $l = 2, 3, \ldots$, we obtain,

$$2\gamma\sum_{l=1}^{\infty}\tilde{d}_l^2(\underline{\alpha}-1) \leqslant \tilde{d}_1^2(\underline{\alpha}) + \gamma^2 R_{\underline{\alpha}}\sum_{l=1}^{\infty}\tilde{d}_l^2(-1)$$

$$\leqslant \tilde{d}_1^2(\underline{\alpha}) + \frac{R_{\underline{\alpha}}}{R_0}\tilde{d}_1^2(0) \tag{2.24}$$

Note that using the strong assumption of regularity of the feature vectors,

$$\tilde{d}_1^2(0) = \mathbb{E}\left[\|\gamma z_1 a_1\|^2\right] \leqslant \gamma^2 R_0 \mathbb{E}\left[z_1^2\right] = 2\gamma^2 R_0 f(x_*)\,,$$
$$\tilde{d}_1^2(\underline{\alpha}) = \mathbb{E}\left[\left\|\Sigma^{-\underline{\alpha}/2}\gamma z_1^2 X\right\|^2\right] \leqslant \gamma^2 R_{\underline{\alpha}} \mathbb{E}\left[z_1^2\right] = 2\gamma^2 R_{\underline{\alpha}} f(x_*)\,.$$

We use these expressions to simply further (2.23) and (2.24):

$$\sum_{l=1}^{n} \tilde{d}_l^2(-1) \leqslant 2\gamma R_0 f(x_*)\,,$$
$$\sum_{l=1}^{\infty} \tilde{d}_l^2(\underline{\alpha}-1) \leqslant 2\gamma R_{\underline{\alpha}} f(x_*)\,.$$

If $\beta \in [-1, \underline{\alpha}-1]$, we use the log-convexity Property 1 and Hölder's inequality: decompose $\beta = (1-\lambda)(-1) + \lambda(\underline{\alpha}-1)$ with $\lambda = (\beta+1)/\underline{\alpha}$,

$$\begin{aligned}
\sum_{l=1}^{\infty} \tilde{d}_l^2(\beta) &\leqslant \sum_{l=1}^{\infty} \tilde{d}_l^2(-1)^{1-\lambda} \tilde{d}_l^2(\underline{\alpha}-1)^{\lambda} \\
&\leqslant \left(\sum_{l=1}^{n} \tilde{d}_l^2(-1)\right)^{1-\lambda} \left(\sum_{l=1}^{\infty} \tilde{d}_l^2(\underline{\alpha}-1)\right)^{\lambda} \\
&\leqslant \left(2\gamma R_0 f(x_*)\right)^{1-\lambda} \left(2\gamma R_{\underline{\alpha}} f(x_*)\right)^{\lambda} \\
&= 2\gamma R_0^{1-\lambda} R_{\underline{\alpha}}^{\lambda} f(x_*)\,.
\end{aligned} \tag{2.25}$$

Putting back together Equations (2.20), (2.22) and (2.25), we obtain

$$d_n^2(\beta) \leqslant 2\mathbb{E}\left[\left\|\Sigma^{-\beta/2}\nu_n\right\|^2\right] + 4\gamma R_0^{1-\lambda} R_{\underline{\alpha}}^{\lambda} f(x_*)$$

The bounds for SGD follow from the application of Theorem 2.4 to the sequence ν_n in order to control the first term.

We now turn to the bounds for tail-averaged SGD. The iterate of tail-averaged SGD can be decomposed as

$$\overline{x}_n = \frac{1}{\lfloor n/2 \rfloor + 1} \sum_{k=\lceil n/2 \rceil}^{n} x_k = \overline{\nu}_n + \frac{1}{\lfloor n/2 \rfloor + 1} \sum_{k=\lceil n/2 \rceil}^{n} \sum_{l=1}^{k} \eta_k^{(l)}\,, \qquad \overline{\nu}_n = \frac{1}{\lfloor n/2 \rfloor + 1} \sum_{k=\lceil n/2 \rceil}^{n} \nu_k\,.$$

Thus as in (2.22), we can decompose the error

$$\overline{d}_n^2(\beta) \leqslant 2\mathbb{E}\left[\left\|\Sigma^{-\beta/2}(\overline{\nu}_n - x_*)\right\|^2\right] + 2\mathbb{E}\left[\left\|\Sigma^{-\beta/2} \frac{1}{\lfloor n/2 \rfloor + 1} \sum_{k=\lceil n/2 \rceil}^{n} \sum_{l=1}^{k} \eta_k^{(l)}\right\|^2\right] \tag{2.26}$$

Again, the first term is studied in Section 2.1. Further,

$$\mathbb{E}\left[\left\|\Sigma^{-\beta/2} \frac{1}{\lfloor n/2 \rfloor + 1} \sum_{k=\lceil n/2 \rceil}^{n} \sum_{l=1}^{k} \eta_k^{(l)}\right\|^2\right] = \frac{1}{\left(\lfloor n/2 \rfloor + 1\right)^2} \sum_{k,k'=\lceil n/2 \rceil}^{n} \sum_{l=1}^{k} \sum_{l'=1}^{k'} \left\langle \eta_k^{(l)}, \Sigma^{-\beta} \eta_{k'}^{(l')} \right\rangle$$

Again, the dot product is zero if $l \neq l'$, thus

$$\begin{aligned}
\mathbb{E}\left[\left\|\Sigma^{-\beta/2}\frac{1}{\lfloor n/2\rfloor+1}\sum_{k=\lceil n/2\rceil}^{n}\sum_{l=1}^{k}\eta_k^{(l)}\right\|^2\right] &= \frac{1}{(\lfloor n/2\rfloor+1)^2}\sum_{k,k'=\lceil n/2\rceil}^{n}\sum_{l=1}^{\min(k,k')}\mathbb{E}\left\langle\eta_k^{(l)},\Sigma^{-\beta}\eta_{k'}^{(l)}\right\rangle \\
&\leqslant \frac{1}{(\lfloor n/2\rfloor+1)^2}\sum_{k,k'=\lceil n/2\rceil}^{n}\sum_{l=1}^{\min(k,k')}\frac{1}{2}\left(\mathbb{E}\|\Sigma^{-\beta/2}\eta_k^{(l)}\|^2+\mathbb{E}\|\Sigma^{-\beta/2}\eta_{k'}^{(l)}\|^2\right) \\
&= \frac{1}{(\lfloor n/2\rfloor+1)^2}\sum_{l=1}^{n}\sum_{k,k'=\max(\lceil n/2\rceil,l)}^{n}\frac{1}{2}\left(\mathbb{E}\|\Sigma^{-\beta/2}\eta_k^{(l)}\|^2+\mathbb{E}\|\Sigma^{-\beta/2}\eta_{k'}^{(l)}\|^2\right) \\
&\leqslant \frac{1}{(\lfloor n/2\rfloor+1)^2}\sum_{l=1}^{n}\sum_{k=\max(\lceil n/2\rceil,l)}^{n}(\lfloor n/2\rfloor+1)\mathbb{E}\|\Sigma^{-\beta/2}\eta_k^{(l)}\|^2 \\
&= \frac{1}{\lfloor n/2\rfloor+1}\sum_{l=1}^{n}\sum_{k=\max(\lceil n/2\rceil,l)}^{n}\mathbb{E}\|\Sigma^{-\beta/2}\eta_k^{(l)}\|^2\,.
\end{aligned}$$

Here, we use again that $\eta_k^{(l)}$ and $\eta_{k-l+1}^{(1)}$ have the same law. This gives

$$\mathbb{E}\left[\left\|\Sigma^{-\beta/2}\frac{1}{\lfloor n/2\rfloor+1}\sum_{k=\lceil n/2\rceil}^{n}\sum_{l=1}^{k}\eta_k^{(l)}\right\|^2\right] \leqslant \frac{1}{\lfloor n/2\rfloor+1}\sum_{l=1}^{n}\sum_{k=\max(\lceil n/2\rceil,l)}^{n}\mathbb{E}\|\Sigma^{-\beta/2}\eta_{k-l+1}^{(1)}\|^2\,.$$

Note that $1\leqslant k-l+1\leqslant n$ and that k can take at most $\lfloor n/2\rfloor+1$ different values, so we can bound this last double sum

$$\begin{aligned}
\mathbb{E}\left[\left\|\Sigma^{-\beta/2}\frac{1}{\lfloor n/2\rfloor+1}\sum_{k=\lceil n/2\rceil}^{n}\sum_{l=1}^{k}\eta_k^{(l)}\right\|^2\right] &\leqslant \frac{1}{\lfloor n/2\rfloor+1}\sum_{l=1}^{n}\sum_{k=\max(\lceil n/2\rceil,l)}^{n}\mathbb{E}\|\Sigma^{-\beta/2}\eta_{k-l+1}^{(1)}\|^2 \\
&\leqslant \sum_{j=1}^{n}\mathbb{E}\|\Sigma^{-\beta/2}\eta_j^{(1)}\|^2\,.
\end{aligned}$$

The rest of the proof in the tail-averaged case is similar to the non-averaged case.

2.C. Proof of Theorems 2.3 and 2.5

We start in the case (a) where the optimum is irregular: $x_*-x_0\notin\Sigma^{-\overline{\alpha}/2}(\mathcal{H})$. In that case, we give a lower bound in the convergence rate by studying the expected process $\mathbb{E}[x_n]$. Indeed, by Jensen's inequality,

$$d_n^2(\beta)=\mathbb{E}\left[\left\langle x_n-x_*,\Sigma^{-\beta}\left(x_n-x_*\right)\right\rangle\right]\geqslant\left\langle\mathbb{E}[x_n]-x_*,\Sigma^{-\beta}\left(\mathbb{E}[x_n]-x_*\right)\right\rangle\,. \tag{2.27}$$

The expectation $\mathbb{E}[x_n]$ can be interpreted as the (non-stochastic) gradient descent on the population risk $f(x)$. Indeed, by taking the expectation of (2.2) under the noiseless linear assumption, we obtain

$$\mathbb{E}[x_n]-x_*=(\mathrm{Id}-\gamma\Sigma)(\mathbb{E}[x_{n-1}]-x_*)=-(\mathrm{Id}-\gamma\Sigma)^n(x_*-x_0)\,. \tag{2.28}$$

Note that as $\gamma\leqslant 1/R_0$, $I-\gamma\Sigma$ is a positive definite matrix. Indeed, by the weak definition of R_0 in Remark 2.4,

$$R_0\Sigma\succcurlyeq\mathbb{E}\left[\|a\|^2a\otimes a\right]=\mathbb{E}\left[(a\otimes a)(a\otimes a)\right]\succcurlyeq\mathbb{E}[a\otimes a]^2=\Sigma^2\,,$$

thus R_0 is larger than the operator norm of Σ. Thus $\gamma\Sigma\preccurlyeq\frac{1}{R_0}\Sigma\preccurlyeq\mathrm{Id}$.

In the following, if $\alpha \in \mathbb{R}$ and $k \in \mathbb{N}$, $\binom{\alpha}{k}$ denotes the generalized binomial coefficient: $\binom{\alpha}{k} = \frac{\alpha(\alpha-1)\cdots(\alpha-k+1)}{k!}$. Fix now $\alpha \geqslant 0$. We have the (formal) power series

$$\begin{aligned}(1+\mu)^{-\alpha} &= \sum_{k=0}^{\infty} \binom{-\alpha}{k} \mu^k \\ (1-\mu)^{-\alpha} &= \sum_{k=0}^{\infty} \binom{-\alpha}{k} (-1)^k \mu^k = \sum_{k=0}^{\infty} \binom{\alpha+k-1}{k} \mu^k \\ \lambda^{-\alpha} &= \sum_{k=0}^{\infty} \binom{\alpha+k-1}{k} (1-\lambda)^k \,.\end{aligned}$$

This last equality holds in $[0,\infty]$ for $\lambda \in [0,1]$. In that case, all terms of the series are positive, thus the meaning of the sum is unambiguous.

Note that $0 \preccurlyeq \gamma\Sigma \preccurlyeq \mathrm{Id}$, thus we have, formally,

$$\gamma^{-\alpha}\Sigma^{-\alpha} = \sum_{k=0}^{\infty} \binom{\alpha+k-1}{k} (\mathrm{Id} - \gamma\Sigma)^k \,.$$

The rigorous meaning of this equality is that for all $x \in \mathcal{H}$,

$$\gamma^{-\alpha}\langle x, \Sigma^{-\alpha} x\rangle = \sum_{k=0}^{\infty} \binom{\alpha+k-1}{k} \langle x, (\mathrm{Id} - \gamma\Sigma)^k x\rangle \,.$$

Both terms of the equality can be infinite: again, here we are using the convention that $\langle x, \Sigma^{-\alpha} x\rangle = \infty \Leftrightarrow x \notin \Sigma^{\alpha/2}(\mathcal{H})$. In particular, take $\alpha = \overline{\alpha} - \beta$ and $x = \Sigma^{-\beta/2}(x_* - x_0)$:

$$\begin{aligned}\infty = \gamma^{\beta-\overline{\alpha}} \left\langle x_* - x_0, \Sigma^{-\overline{\alpha}}(x_* - x_0)\right\rangle &= \sum_{k=0}^{\infty} \binom{\overline{\alpha}-\beta+k-1}{k} \left\langle x_* - x_0, \Sigma^{-\beta}(\mathrm{Id} - \gamma\Sigma)^k (x_* - x_0)\right\rangle \\ = \sum_{n=0}^{\infty} \Bigg[&\binom{\overline{\alpha}-\beta+2n-1}{2n} \left\langle x_* - x_0, \Sigma^{-\beta}(\mathrm{Id} - \gamma\Sigma)^{2n} (x_* - x_0)\right\rangle \\ &+ \binom{\overline{\alpha}-\beta+2n}{2n+1} \left\langle x_* - x_0, \Sigma^{-\beta}(\mathrm{Id} - \gamma\Sigma)^{2n+1} (x_* - x_0)\right\rangle \Bigg] \,.\end{aligned}$$

Using that $\binom{\overline{\alpha}-\beta+2n-1}{2n} \leqslant \binom{\overline{\alpha}-\beta+2n}{2n+1}$ and

$$\left\langle x_* - x_0, \Sigma^{-\beta}(\mathrm{Id} - \gamma\Sigma)^{2n} (x_* - x_0)\right\rangle \geqslant \left\langle x_* - x_0, \Sigma^{-\beta}(\mathrm{Id} - \gamma\Sigma)^{2n+1} (x_* - x_0)\right\rangle$$

and then (2.28), (2.27), we obtain

$$\begin{aligned}\infty &\leqslant 2 \sum_{n=0}^{\infty} \binom{\overline{\alpha}-\beta+2n}{2n+1} \left\langle x_* - x_0, \Sigma^{-\beta}(\mathrm{Id} - \gamma\Sigma)^{2n} (x_* - x_0)\right\rangle \\ &= 2 \sum_{n=0}^{\infty} \binom{\overline{\alpha}-\beta+2n}{2n+1} \left\langle \mathbb{E}[x_n] - x_*, \Sigma^{-\beta}(\mathbb{E}[x_n] - x_*)\right\rangle \\ &\leqslant 2 \sum_{n=0}^{\infty} \binom{\overline{\alpha}-\beta+2n}{2n+1} d_n^2(\beta) \,.\end{aligned}$$

From [Olver et al., 2010, Equation 5.8.1], we have the formula $\Gamma(z) = \lim_{k\to\infty} \frac{k!k^z}{z(z+1)\cdots(z+k)}$ where Γ denotes the Gamma function. Thus as $n \to \infty$

$$\binom{\overline{\alpha}-\beta+2n}{2n+1} = \frac{(\overline{\alpha}-\beta)(\overline{\alpha}-\beta+1)\cdots(\overline{\alpha}-\beta+2n)}{(2n+1)(2n)!} \sim \frac{(2n)^{\overline{\alpha}-\beta}}{(2n+1)\Gamma(\overline{\alpha}-\beta)}\,.$$

As a consequence, the series $\sum_n n^{\overline{\alpha}-\beta-1} d_n^2(\beta)$ diverges. The criteria for the convergence of Riemann series implies that $d_n^2(\beta)$ can not be asymptotically dominated by $1/n^{\overline{\alpha}-\beta+\varepsilon}$ for $\varepsilon > 0$.

We now turn to the case (b) where the features are irregular: with positive probability $p > 0$, $a \notin \Sigma^{\overline{\alpha}/2}(\mathcal{H})$ and $\langle a, x_* - x_0\rangle \neq 0$. We can assume that we are not in case (a): $x_* - x_0 \in \Sigma^{-\overline{\alpha}/2}(\mathcal{H})$. Then with probability p, the second iterate $x_1 = x_0 - \gamma\langle a_1, x_* - x_0\rangle a_1$ is irregular, i.e., $x_1 \notin \Sigma^{\overline{\alpha}/2}(\mathcal{H})$. Thus we can apply case (a) to the iteration started from x_1. This shows that $d_n^2(\beta)$ is not asymptotically dominated by $1/n^{\overline{\alpha}-\beta+\varepsilon}$, for $\varepsilon > 0$.

2.D. Proof of Corollary 2.1

Recall from Section 1.4.1 that the simple gossip algorithm can be seen as a stochastic gradient descent, with step-size $\gamma = 1/2$, on the least-squares risk

$$f(x) = \frac{1}{2N} \sum_{\{v,w\}\in\mathcal{E}} \langle e_v - e_w, x\rangle^2\,,$$

which is of the form (2.2) for $b = 0$ and $a = e_v - e_w$, for $\{e_v, e_w\}$ a uniformly sampled edge in the graph. From this, we compute $\|a\|^2 = 2$ thus we can take $R_0 = 2$. The minimum $x_* = \bar{x}\mathbb{1}$ is the unique minimum of $f(x)$ in the set $x_0 + \mathbb{1}^\perp$. Thus the simple gossip algorithm converges to this point. Note that $\gamma = 1/R_0$ and thus Theorem 2.2 applies. This justifies our special care in tolerating step-sizes as large as $1/R_0$ in our study.

Under the assumptions of this corollary, Proposition 1.2 provides the source and capacity conditions of the least-squares problem; we repeat the result here for the convenience of the reader:

(1) (source condition) for any $\alpha < d/2$, the optimum has regularity α, and

$$\|\Sigma^{-\alpha/2}(x_* - x_0)\|^2 \leqslant N^\alpha V^{-1} \delta_{\max}^{d/2-\alpha} \frac{d}{d-2\alpha}\,,$$

where again N denotes then number of edges in the graph, and

(2) (capacity condition) for any $\alpha < d/2$, the optimum has regularity α with associated constant

$$R_\alpha = 2N^\alpha V^{-1} \delta_{\max}^{d/2-\alpha} \frac{d}{d-2\alpha}\,.$$

Theorem 2.2 gives

$$\begin{aligned}\mathbb{E}\left[\|x_n - x_*\|^2\right] &\leqslant \frac{\alpha^\alpha}{\gamma^\alpha}\left(\|\Sigma^{-\alpha/2}(x_* - x_0)\|^2 + \frac{R_\alpha}{R_0}\|x_* - x_0\|^2\right)\frac{1}{n^\alpha} \\ &\leqslant \frac{(d/2)^\alpha}{(1/2)^\alpha}\left(N^\alpha V^{-1}\delta_{\max}^{d/2-\alpha}\frac{d}{d-2\alpha} + N^\alpha V^{-1}\delta_{\max}^{d/2-\alpha}\frac{d}{d-2\alpha}\|x_* - x_0\|^2\right)\frac{1}{n^\alpha}\end{aligned}$$

Note that $\|x_* - x_0\|_2^2 \leqslant 1$ and recall the scaling $s = n/N$:

$$\mathbb{E}\left[\|x_n - x_*\|^2\right] \leqslant d^{d/2+1} V^{-1} \delta_{\max}^{d/2-\alpha} \frac{1}{d/2-\alpha}\frac{1}{s^\alpha}\,.$$

This bound is valid for all $\alpha < \frac{d}{2}$. Choose $\alpha = \frac{d}{2} - \frac{\log 2}{\log s}$.

$$\mathbb{E}\left[\|x_n - x_*\|^2\right] \leqslant d^{d/2+1} V^{-1} \delta_{\max}^{\log 2/\log s} \frac{\log s}{\log 2}\frac{2}{s^{d/2}}$$

As we assume $s \geqslant 2$, $\delta_{\max}^{\log 2/\log s} \leqslant \delta_{\max}$. Thus we obtain conclusion 1.

The proof of 2 is similar. Theorem 2.2 gives

$$\begin{aligned}\min_{0\leqslant k\leqslant n} \mathbb{E}\left[\frac{1}{2}\sum_{\{v,w\}\in\mathcal{E}}(x_k(v)-x_k(w))^2\right] &= N \min_{0\leqslant k\leqslant n} \mathbb{E}\left[\frac{1}{2}\langle x_k - x_*, \Sigma(x_k-x_*)\rangle\right] \\ &\leqslant 2^\alpha \frac{\alpha^\alpha}{\gamma^{\alpha+1}}\left(\|\Sigma^{-\alpha/2}(x_*-x_0)\|^2 + \frac{R_\alpha}{R_0}\|x_*-x_0\|^2\right)\frac{1}{n^\alpha} \\ &\leqslant 2^{\alpha+1} d^\alpha V^{-1}\delta_{\max}^{d/2-\alpha}\frac{d}{d/2-\alpha}\frac{1}{s^{\alpha+1}}\,.\end{aligned}$$

Taking again $\alpha = \frac{d}{2} - \frac{1}{2\log s}$ and $s \geqslant 2$,

$$\min_{0\leqslant k\leqslant n} \mathbb{E}\left[\frac{1}{2}\sum_{\{v,w\}\in\mathcal{E}}(x_k(v)-x_k(w))^2\right] \leqslant 2^{d/2+1} d^{d/2} V^{-1}\delta_{\max}\frac{d\log s}{\log 2}\frac{2}{s^{d/2+1}}$$

This gives conclusion 2 of the corollary.

CHAPTER 3

A Continuized View on Nesterov Acceleration

We remind that the majority of the contents of this chapter are available in the preprint (under review):

> M. Even, R. Berthier, F. Bach, N. Flammarion, P. Gaillard, H. Hendrikx, L. Massoulié, A. Taylor. A Continuized View on Nesterov Acceleration for Stochastic Gradient Descent and Randomized Gossip, 2021, preprint.

In this chapter, we assume that $f : \mathcal{H} \to \mathbb{R}$ is a convex and differentiable function. We assume that f is minimized at a point $x_* \in \mathcal{H}$. We assume throughout the paper that f is L-smooth, and sometimes assume that f is μ-strongly convex for some $\mu > 0$. For the problem of minimizing f, gradient descent is well-known to achieve a rate $f(x_k) - f(x_*) = O(k^{-1})$ in the smooth case (Theorem 1.1), and a rate $f(x_k) - f(x_*) = O((1 - \mu/L)^k)$ in the smooth and strongly convex case (Theorem 1.3). In both cases, Nesterov introduced an alternative method with essentially the same running-time complexity, that achieves faster rates: it converges at the rate $O(k^{-2})$ in the smooth convex case and at the rate $O((1 - \sqrt{\mu/L})^k)$ in the smooth and strongly convex case [Nesterov, 2003]. These rates are then optimal among all methods that access gradients and linearly combine them [Nesterov, 2003, Nemirovskij and Yudin, 1983].

Nesterov acceleration introduces several sequences of iterates—two or three, depending on the formulations—and relies on a clever blend of gradient steps and mixing steps between the iterates. Many works contributed to interpret and motivate the precise structure of the iteration that lead to the success of the method, see for instance [Bubeck et al., 2015, Flammarion and Bach, 2015, Arjevani et al., 2016, Kim and Fessler, 2016, Allen-Zhu and Orecchia, 2017]. A large number of these works found useful to study continuous time equivalents of Nesterov acceleration, obtained by taking the limit when step-sizes vanish, or from a variational framework. The continuous time index t of the limit allowed to use differential calculus to study the convergence of these equivalents. For examples of studies that use continuous time, see [Su et al., 2014, Krichene et al., 2015, Wilson et al., 2016, Wibisono et al., 2016, Betancourt et al., 2018, Diakonikolas and Orecchia, 2019, Shi et al., 2018, 2019, Attouch et al., 2018, 2019, Zhang et al., 2018, Muehlebach and Jordan, 2019].

In this paper, we propose another way to obtain a continuous time equivalent of Nesterov acceleration, that we call the *continuized* version of Nesterov acceleration, that does not require vanishing step-sizes. It is built by considering two variables $x_t, z_t \in \mathcal{H}$, $t \in \mathbb{R}_{\geqslant 0}$, that continuously mix following a linear ordinary differential equation (ODE), and that take gradient steps at random times $T_1, T_2, T_3, \ldots$. Thus, in this modeling, mixing and gradient steps alternate randomly.

Thanks to the continuous index t and some stochastic calculus, one can differentiate averaged quantities (expectations) with respect to t. In particular, this leads to simple analytical expressions for the optimal parameters as functions of t.

The discretization $\tilde{x}_k = x_{T_k}, \tilde{z}_k = z_{T_k}, k \in \mathbb{N}$, of the continuized process can be computed directly and exactly: the result is a recursion of the same form as Nesterov iteration, but with randomized parameters, that performs similarly to Nesterov original deterministic version both in theory and in simulations.

The continuized framework can be adapted to various settings and extensions of Nesterov acceleration: we study how the continuized acceleration behaves in the presence of *additive* and *multiplicative* noise in the gradients. In the multiplicative noise setting, our acceleration satisfies a convergence rate similar to the acceleration of Jain et al. [2018]; it depends on the *statistical condition number* of the problem. The two accelerations are not directly comparable as we work in a continuized setting and only deal with pure multiplicative noise, but our analysis is much simpler, as it closely mimics that of Nesterov acceleration.

The continuized modeling is natural in asynchronous parallel computing where gradient steps arrive at random times. But more importantly, there are situations where the continuized version of Nesterov acceleration can be practically implemented while the original acceleration can not. In distributed settings for instance, the total number k for gradient steps taken in the network may not be known to a particular node; the advantage of the continuized acceleration is that is requires to know only the time t and not k. As an illustration, we use the parallel of Section 1.4 to derive the acceleration of asynchronous gossip algorithms of Even et al. [2020] from the continuized framework. Other acceleration schemes [Hendrikx et al., 2019, Loizou et al., 2019] were practically limited by the requirement of additional synchronizations between nodes, such as the knowledge of a global iteration counter. Their accelerated gossip algorithm recovers the same accelerated rates, and only requires the knowledge of a common continuous time.

To sum up, the continuized acceleration should be seen as a close approximation to Nesterov acceleration, that features both an insightful and convenient expression as a continuous time process and a direct implementation as a discrete iteration. We thus hope to contribute to the understanding of Nesterov acceleration. In practice, the continuized framework is relevant for handling asynchrony in decentralized optimization, where agents of a network can not share a global iteration counter, preventing accelerated decentralized and asynchronous methods.

Notations. The index k always denotes a non-negative integer, while indices t, s always denote non-negative reals.

Outline of this chapter. In Section 3.1, we recall standard results on gradient descent and Nesterov acceleration. In Section 3.2, we introduce a continuized variant of Nesterov acceleration. In Section 3.3, we show that discretizing the continuized acceleration yields an iterative method similar to that of Nesterov but with random parameters. In Section 3.4, we study continuized Nesterov acceleration under pure-multiplicative noise. We finally present accelerated asynchronous algorithms for the gossip problem in Section 3.5.

3.1. Reminders on Nesterov acceleration

For the sake of comparison, let us first recall the classical Nesterov acceleration. To improve the convergence rate of gradient descent, Nesterov introduced iterations of three sequences, parameterized by $\tau_k, \tau'_k, \gamma_k, \gamma'_k, k \geqslant 0$, of the form

$$y_k = x_k + \tau_k(z_k - x_k)\,, \tag{3.1}$$

$$x_{k+1} = y_k - \gamma_k \nabla f(y_k)\,, \tag{3.2}$$

$$z_{k+1} = z_k + \tau'_k(y_k - z_k) - \gamma'_k \nabla f(y_k)\,. \tag{3.3}$$

Depending on whether the function f is known to be (1) convex, or (2) strongly convex with a known strong convexity parameter, Nesterov provided a set of parameter choices for achieving acceleration.

Theorem 3.1 (Convergence of accelerated gradient descent)**.** Nesterov accelerated scheme satisfies:

(1) Choose the parameters $\tau_k = 1 - \frac{A_k}{A_{k+1}}, \tau'_k = 0, \gamma_k = \frac{1}{L}, \gamma'_k = \frac{A_{k+1}-A_k}{L}, k \geqslant 0$, where the sequence $A_k, k \geqslant 0$, is defined by the recurrence relation

$$A_0 = 0\,, \qquad\qquad A_{k+1} = A_k + \frac{1}{2}(1 + \sqrt{4A_k + 1})\,.$$

Then

$$f(x_k) - f(x_*) \leqslant \frac{2L\|x_0 - x_*\|^2}{k^2}\,.$$

(2) Assume further that f is μ-strongly convex, $\mu > 0$. Choose the constant parameters $\tau_k \equiv \frac{\sqrt{\mu/L}}{1+\sqrt{\mu/L}}$, $\tau'_k \equiv \sqrt{\frac{\mu}{L}}$, $\gamma_k \equiv \frac{1}{L}$, $\gamma'_k \equiv \frac{1}{\sqrt{\mu L}}$, $k \geqslant 0$. Then

$$f(x_k) - f(x_*) \leqslant \left(f(x_0) - f(x_*) + \frac{\mu}{2}\|z_0 - x_*\|^2\right)\left(1 - \sqrt{\frac{\mu}{L}}\right)^k\,.$$

This result can be found as is in d'Aspremont et al. [2021, Sections 4.4.1 and 4.5.3]. From a high-level perspective, Nesterov acceleration iterates over several variables, alternating between gradient steps (always with respect to the gradient at y_k) and mixing steps, where the running value of a variable is replaced by a linear combination of the other variables. However, the precise way gradient and mixing steps are coupled is rather mysterious, and the success of the proof of Theorem 3.1 relies heavily on the detailed structure of the iterations. In the next section, we try to gain perspective on this structure by developing a continuized version of the acceleration.

3.2. Continuized version of Nesterov acceleration

This paper uses several mathematical notions related to random processes. The following sections expose the results from heuristic considerations of those notions, rigorously defined in Appendix 3.A.

We argue that the accelerated iteration becomes more natural when considering two variables x_t, z_t indexed by a continuous time $t \geqslant 0$, that are continuously mixing and that take gradient steps at random times. More precisely, let $T_1, T_2, T_3, \ldots \geqslant 0$ be random times such that $T_1, T_2 - T_1, T_3 - T_2, \ldots$ are independent identically distributed (i.i.d.), of exponential law with rate 1 (any constant rate would do, we choose 1 to make the comparison with discrete time k straightforward). By convention, we choose that our stochastic processes $t \mapsto x_t, t \mapsto z_t$ are càdlàg almost surely, i.e., right continuous with well-defined left-limits x_{t-}, z_{t-} (Definition 3.5 in Appendix 3.A). Our dynamics are parameterized by functions $\gamma_t, \gamma'_t, \eta_t, \eta'_t$, $t \geqslant 0$. At random times $T_1, T_2, \ldots$, our sequences take gradient steps

$$x_{T_k} = x_{T_k-} - \gamma_{T_k}\nabla f(x_{T_k-})\,, \tag{3.4}$$

$$z_{T_k} = z_{T_k-} - \gamma'_{T_k}\nabla f(x_{T_k-})\,. \tag{3.5}$$

Because of the memoryless property of the exponential distribution, in a infinitesimal time interval $[t, t+\mathrm{d}t]$, the variables take gradients steps with probability $\mathrm{d}t$, independently of the past. Between these random times, the variables mix through a linear, translation-invariant, ordinary differential equation (ODE)

$$\mathrm{d}x_t = \eta_t(z_t - x_t)\mathrm{d}t\,, \tag{3.6}$$

$$\mathrm{d}z_t = \eta'_t(x_t - z_t)\mathrm{d}t\,. \tag{3.7}$$

Following the notation of stochastic calculus, we can write the process more compactly in terms of the Poisson point measure $\mathrm{d}N(t) = \sum_{k \geqslant 1} \delta_{T_k}(\mathrm{d}t)$, which has as intensity the Lebesgue measure $\mathrm{d}t$,

$$\mathrm{d}x_t = \eta_t(z_t - x_t)\mathrm{d}t - \gamma_t \nabla f(x_t)\mathrm{d}N(t)\,, \tag{3.8}$$

$$\mathrm{d}z_t = \eta'_t(x_t - z_t)\mathrm{d}t - \gamma'_t \nabla f(x_t)\mathrm{d}N(t)\,. \tag{3.9}$$

Before giving convergence guarantees for such processes, let us digress quickly on why we can expect an iteration of this form to be mathematically appealing.

First, from a Markov chain indexed by a discrete time index k, one can associate the so-called *continuized* Markov chain, indexed by a continuous time t, that makes transition with the same Markov kernel, but at random times, with independent exponential time intervals [Aldous and Fill, 2002]. Following this terminology, we refer to our acceleration (3.8)-(3.9) as the continuized acceleration. The continuized Markov chain is appreciated for its continuous time parameter t, while keeping many properties of the original Markov chain; similarly the continuized acceleration is arguably simpler to analyze, while performing similarly to Nesterov acceleration.

Second, it can also be compared with coordinate gradient descent methods, that are easier to analyze when coordinates are selected randomly rather than in an ordered way [Wright, 2015]. Similarly, the continuized acceleration is simpler to analyze because the gradient steps (3.4)-(3.5) and the mixing steps (3.6)-(3.7) alternate randomly, due to the randomness of T_k, $k \geqslant 0$.

In analogy with Theorem 3.1, we give choices of parameters that lead to accelerated convergence rates, in the convex case (1) and in the strongly convex case (2). Convergence is analyzed as a function of t. As $\mathrm{d}N(t)$ is a Poisson point process with rate 1, t is the expected number of gradient steps done by the algorithm. Thus t is analoguous to k in Theorem 3.1. In the theorem below, $\mathbb{E}$ denotes the expectation with respect to the Poisson point process $\mathrm{d}N(t)$, the only source of randomness.

Theorem 3.2 (Convergence of continuized Nesterov acceleration)**.** The continuized Nesterov acceleration satisfies the following two points.

(1) Choose the parameters $\eta_t = \frac{2}{t}, \eta'_t = 0, \gamma_t = \frac{1}{L}, \gamma'_t = \frac{t}{2L}$. Then

$$\mathbb{E}f(x_t) - f(x_*) \leqslant \frac{2L\|z_0 - x_*\|^2}{t^2}\,.$$

(2) Assume further that f is μ-strongly convex, $\mu > 0$. Choose the constant parameters $\eta_t = \eta'_t \equiv \sqrt{\frac{\mu}{L}}$, $\gamma_t \equiv \frac{1}{L}$, $\gamma'_t \equiv \frac{1}{\sqrt{\mu L}}$. Then

$$\mathbb{E}f(x_t) - f(x_*) \leqslant \left(f(x_0) - f(x_*) + \frac{\mu}{2}\|z_0 - x_*\|^2\right) \exp\left(-\sqrt{\frac{\mu}{L}}t\right)\,.$$

We give an elementary sketch of proof in Appendix 3.B.1 and a complete proof in Appendix 3.B.2. Many authors have proposed continuous-time versions of Nesterov acceleration using differential calculus, see the numerous references in the introduction. For instance, in Su et al. [2014], an ODE is obtained from Nesterov acceleration by taking the joint asymptotic where the step-sizes vanish and the number of iterates is rescaled. The resulting ODE must be discretized to be implemented; choosing the right discretization is not straightforward as it introduces stability and approximation errors that must be controlled [Zhang et al., 2018, Shi et al., 2019, Sanz-Serna and Zygalakis, 2020].

On the contrary, our continuous time process (3.8)-(3.9) does not correspond to a limit where the step-sizes vanish. However, in Appendix 3.D, we check that the random continuized acceleration

has the same deterministic ODE scaling limit as Nesterov acceleration. This sanity check emphasizes that the continuized acceleration is fundamentally different from previous continuous-time equivalents.

3.3. Discrete implementation of the continuized acceleration with random parameters

In this section, we show that the continuized acceleration can be implemented exactly as a discrete algorithm. This contrasts with the discretization of ODEs that introduces discretization errors; here, we compute exactly

$$\tilde{x}_k := x_{T_k}\,, \qquad \tilde{y}_k := x_{T_{k+1}-}\,, \qquad \tilde{z}_k := z_{T_k}\,,$$

with the convention that $T_0 = 0$. The three sequences $\tilde{x}_k, \tilde{y}_k, \tilde{z}_k$, $k \geqslant 0$, satisfy a recurrence relation of the same structure as Nesterov acceleration, but with random weights. The resulting randomized discrete algorithm satisfies performance guarantees similar to those of Nesterov acceleration.

Theorem 3.3 (Discrete version of continuized acceleration)**.** For any stochastic process of the form (3.8)-(3.9), we have

$$\tilde{y}_k = \tilde{x}_k + \tau_k(\tilde{z}_k - \tilde{x}_k)\,, \tag{3.10}$$
$$\tilde{x}_{k+1} = \tilde{y}_k - \tilde{\gamma}_k \nabla f(\tilde{y}_k)\,, \tag{3.11}$$
$$\tilde{z}_{k+1} = \tilde{z}_k + \tau'_k(\tilde{y}_k - \tilde{z}_k) - \tilde{\gamma}'_k \nabla f(\tilde{y}_k)\,, \tag{3.12}$$

for some random parameters $\tau_k, \tau'_k, \tilde{\gamma}_k, \tilde{\gamma}'_k$ (that are functions of $T_k, T_{k+1}, \eta_t, \eta'_t, \gamma_t, \gamma'_t$).

(1) For the parameters of Theorem 3.2.(1), $\tau_k = 1 - \left(\frac{T_k}{T_{k+1}}\right)^2$, $\tau'_k = 0$, $\tilde{\gamma}_k = \frac{1}{L}$, and $\tilde{\gamma}'_k = \frac{T_k}{2L}$. Then

$$\mathbb{E}\left[T_k^2\left(f(\tilde{x}_k) - f(x_*)\right)\right] \leqslant 2L\|z_0 - x_*\|^2\,.$$

(2) For the parameters of Theorem 3.2.(2), $\tau_k = \frac{1}{2}\left(1 - \exp\left(-2\sqrt{\frac{\mu}{L}}(T_{k+1} - T_k)\right)\right)$, $\tau'_k = \tanh\left(\sqrt{\frac{\mu}{L}}(T_{k+1} - T_k)\right)$, $\tilde{\gamma}_k = \frac{1}{L}$, and $\tilde{\gamma}'_k = \frac{1}{\sqrt{\mu L}}$. Then

$$\mathbb{E}\left[\exp\left(\sqrt{\frac{\mu}{L}}T_k\right)\left(f(\tilde{x}_k) - f(x_*)\right)\right] \leqslant f(x_0) - f(x_*) + \frac{\mu}{2}\|z_0 - x_*\|^2\,.$$

The law of T_k is well known: it is the sum of k i.i.d. random variables of exponential law with rate 1; this is called an Erlang or Gamma distribution with shape parameter k and rate 1. Alternatively, $T_k/2$ follows a chi-square distribution with $2k$ degrees of freedom. One can use well-known properties of this law, such as its concentration around its expectation $\mathbb{E}T_k = k$, to derive corollaries of the bounds above. The performance guarantees are proved in Appendix 3.B.2, and the formula for the discretization is studied in Appendix 3.C.

In Figure 3.1, we compare this continuized Nesterov acceleration (3.10)-(3.12) with the classical Nesterov acceleration (3.1)-(3.3) and gradient descent. In the strongly convex case (right), we run the algorithms with the parameters of Theorem 3.1.(2) and 3.3.(2) on the function

$$f(x_1, x_2, x_3) = \frac{\mu}{2}(x_1 - 1)^2 + \frac{3\mu}{2}(x_2 - 1)^2 + \frac{L}{2}(x_3 - 1)^2\,,$$

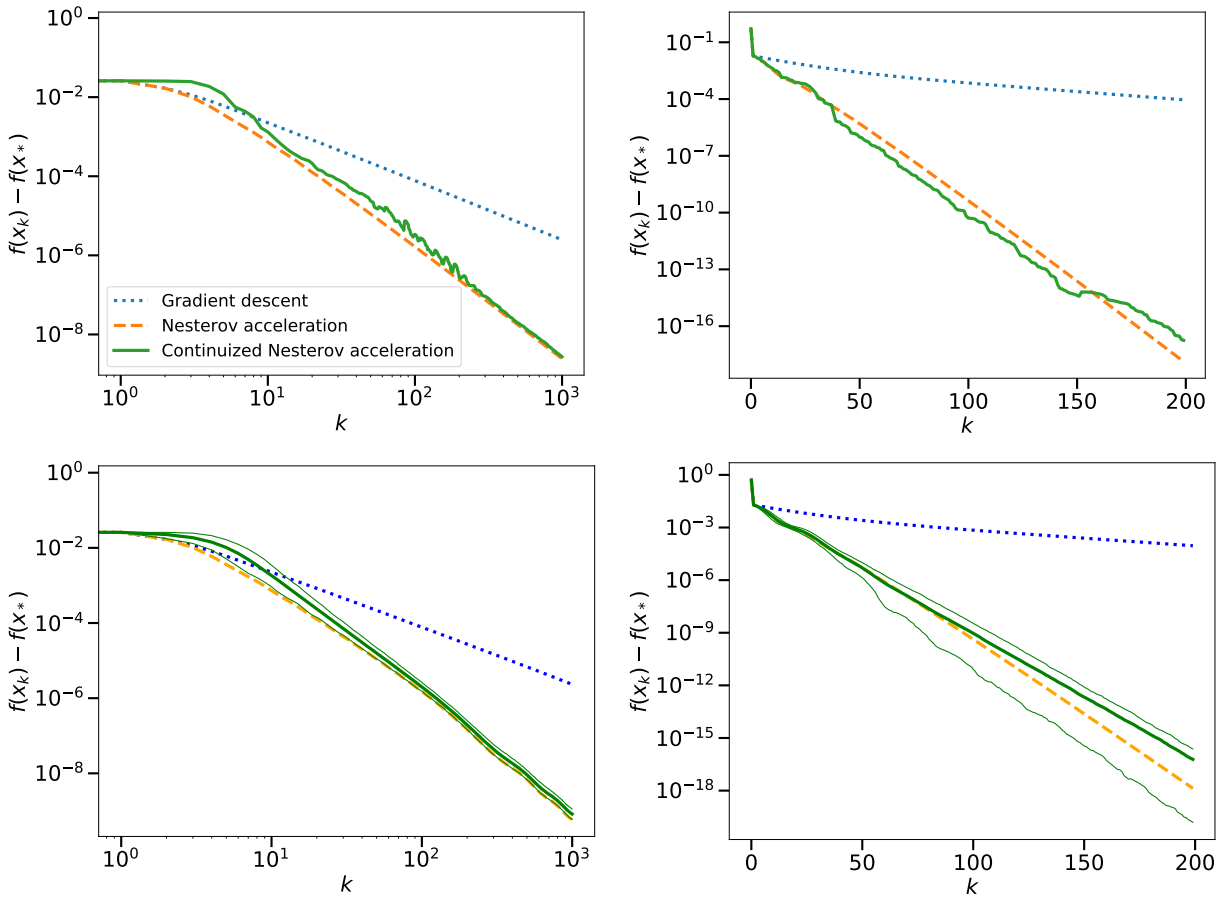


FIGURE 3.1. Comparison between gradient descent, Nesterov acceleration, and the continuized version of Nesterov acceleration, on a convex function (left plots) and a strongly convex function (right plots). For the continuized acceleration, which is randomized, the results shown in the above plots correspond to a single run. In the plots below, the thick line represents the average performance over $N = 1000$ runs of the continuized acceleration, while the thin lines represent the 5% and 95% quantiles.

with $\mu = 10^{-2}$ and $L = 1$. In the convex case, we run the algorithms with the parameters of Theorem 3.1.(1) and 3.3.(1) on the function

$$f(x_1, \dots, x_{100}) = \frac{1}{2} \sum_{i=1}^{100} \frac{1}{i^2} \left(x_i - \frac{1}{i} \right)^2 ,$$

which has negligible strong convexity parameter. All iterations were initialized from $x_0 = z_0 = 0$.

3.4. Continuized Nesterov acceleration of stochastic gradient descent

We now investigate the design of continuized accelerations of stochastic gradient descent. We assume that we do not have direct access to the gradient $\nabla f(x)$ but to a random estimate $g(x, \xi)$, where $\xi \in \Xi$ is random of law $\mathcal{P}$. In the continuized framework, the randomness of the stochastic gradient and its time mix in a particularly convenient way. For similar reasons, Latz studied

stochastic gradient descent as a gradient flow on a random function that is regenerated at a Poisson rate [Latz, 2021]. However, this approach has the same shortcomings as the other approaches based on gradient flows: the subsequent discretization introduces non-trivial errors. We avoid this problem here.

We keep the algorithms of the same form, replacing gradients by stochastic gradients. Let $\xi_1, \xi_2, \dots$ be i.i.d. random variables of law $\mathcal{P}$. We take stochastic gradient steps at the random times $T_1, T_2, \dots,$

$$x_{T_k} = x_{T_k-} - \gamma_{T_k} g(x_{T_k-}, \xi_k)\,,$$
$$z_{T_k} = z_{T_k-} - \gamma'_{T_k} g(x_{T_k-}, \xi_k)\,.$$

Between these random times, the variables mix through the same ODE

$$\mathrm{d}x_t = \eta_t(z_t - x_t)\mathrm{d}t\,,$$
$$\mathrm{d}z_t = \eta'_t(x_t - z_t)\mathrm{d}t\,.$$

This can be written more compactly in terms of the Poisson point measure $\mathrm{d}N(t,\xi) = \sum_{k\geqslant 1} \delta_{(T_k,\xi_k)}(\mathrm{d}t, \mathrm{d}\xi)$ on $\mathbb{R}_{\geqslant 0} \times \Xi$, which has as intensity $\mathrm{d}t \otimes \mathcal{P}$,

$$\mathrm{d}x_t = \eta_t(z_t - x_t)\mathrm{d}t - \gamma_t \int_\Xi g(x_t, \xi)\mathrm{d}N(t,\xi)\,, \tag{3.13}$$
$$\mathrm{d}z_t = \eta'_t(x_t - z_t)\mathrm{d}t - \gamma'_t \int_\Xi g(x_t, \xi)\mathrm{d}N(t,\xi)\,. \tag{3.14}$$

Here, the discussion depends on the properties satisfied by the stochastic gradients $g(x,\xi)$. First, in Section 3.4.1, we study the so-called *additive noise* case. We show that the continuized acceleration satisfies perturbed convergence rates with the same choices of parameters as in Theorem 3.2. We thus show some robustness of the above acceleration to additive noise. Second, in Section 3.4.2, we focus on the noiseless case, or pure multiplicative noise case, as it is crucial for the study of asynchronous gossip that follows. In this setting, parameters need to be chosen differently for our proof technique to work. A continuized acceleration is still possible, depending on the statistical condition number.

3.4.1. Robustness of the continuized Nesterov acceleration to additive noise. In this section, we study the continuized acceleration (3.13)-(3.14) under stochastic gradients. We assume that our gradient estimates are unbiased, i.e.,

$$\forall x \in \mathcal{H}\,, \qquad \mathbb{E}_\xi g(x,\xi) = \nabla f(x)\,, \tag{3.15}$$

and has a uniformly bounded variance, i.e., there exists $\sigma^2 \geqslant 0$ such that

$$\forall x \in \mathcal{H}\,, \qquad \mathbb{E}_\xi \left\|g(x,\xi) - \nabla f(x)\right\|^2 \leqslant \sigma^2\,. \tag{3.16}$$

These assumptions typically hold in the additive noise model of Example 1.1, where $g(x,\xi) = \nabla f(x) + \xi$, and $\xi \in \mathcal{H}$ satisfies $\mathbb{E}\xi = 0$, $\mathbb{E}\|\xi\|^2 \leqslant \sigma^2$.

We should emphasize that similar studies of Nesterov acceleration under additive noise has been done [Lan, 2012, Hu et al., 2009, Xiao, 2010, Devolder, 2011, Cohen et al., 2018, Aybat et al., 2020].

> **Theorem 3.4** (Continuized acceleration with additive noise)**.** Assume that the stochastic gradients are unbiased (3.15) and have a variance uniformly bounded by σ^2 (3.16). Then the continuized acceleration (3.13)-(3.14) satisfies the following.

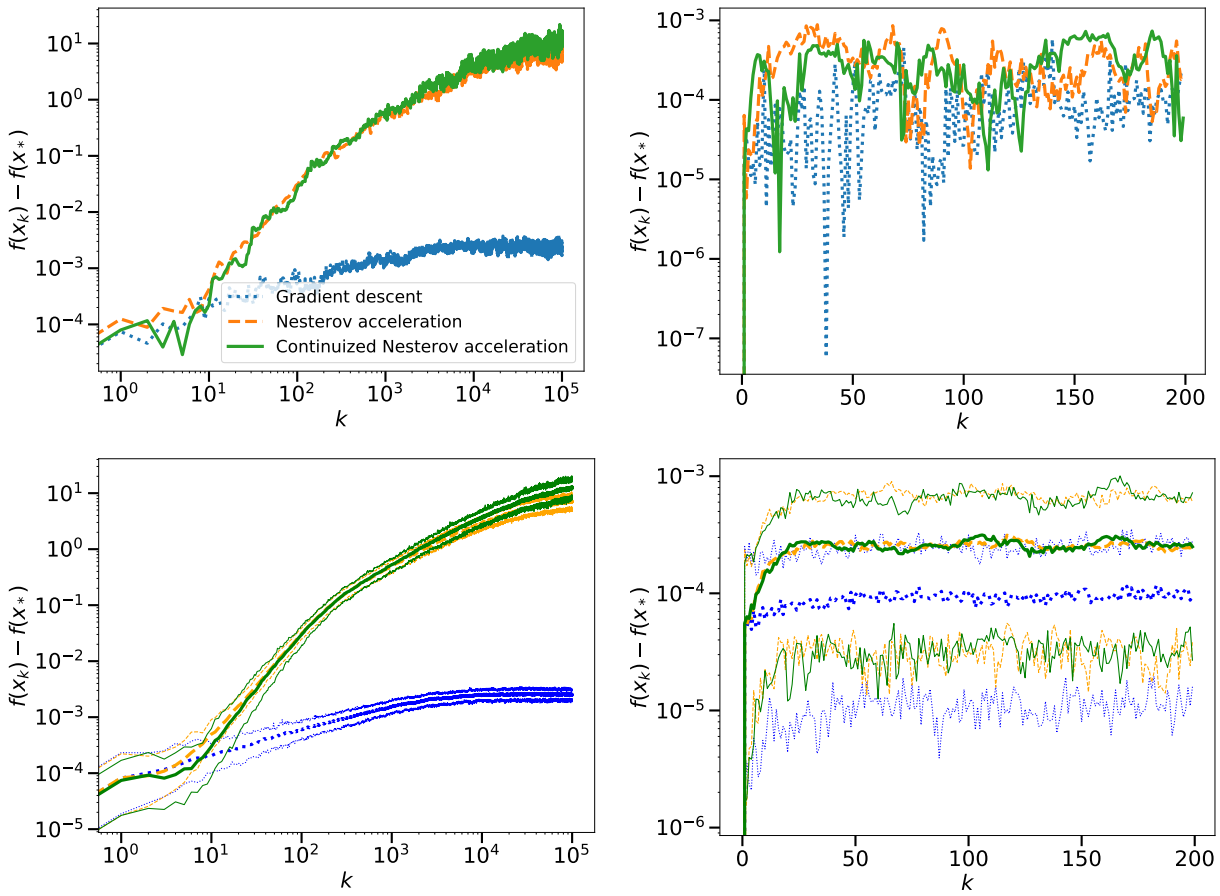


FIGURE 3.2. Effect of additive noise on gradient descent, Nesterov acceleration, and the continuized version of Nesterov acceleration, on a convex function (left) and a strongly convex function (right). All algorithms are started from the optimum x_*. The results shown in the above plots correspond to a single run. In the plots below, the thick line represents the average performance over $N = 100$ runs of each algorithm, while the thin lines represent the 5% and 95% quantiles.

(1) For the parameters of Theorem 3.2.(1),

$$\mathbb{E} f(x_t) - f(x_*) \leqslant \frac{2L\|z_0 - x_*\|^2}{t^2} + \sigma^2 \frac{t}{3L}\,.$$

(2) Assume further that f is μ-strongly convex, $\mu > 0$. For the parameters of Theorem 3.2.(2),

$$\mathbb{E} f(x_t) - f(x_*) \leqslant \left(f(x_0) - f(x_*) + \frac{\mu}{2}\|z_0 - x_*\|^2\right) \exp\left(-\sqrt{\frac{\mu}{L}}t\right) + \sigma^2 \frac{1}{\sqrt{\mu L}}\,.$$

This theorem is proved in Appendix 3.B.3.

In the above bounds, L is a parameter of the algorithm, that can be taken greater than the best known smoothness constant of the function f. Increasing L reduces the step-sizes of the algorithm and performs some variance reduction. If the bound σ^2 on the variance is known, one can choose L optimizing the above bounds in order to obtain algorithms that adapt to additive noise.

In Figure 3.2, we run the same simulations as in Figure 3.1, with two differences: (1) we add isotropic Gaussian noise on the gradients, with covariance $10^{-4}\,\mathrm{Id}$, and (2) we initialized algorithms at the optimum, i.e., $x_0 = z_0 = x_*$. Initializing at the optimum enables to isolate the effect of the additive noise only. These simulations confirm Theorem 3.4: the noise term is (sub-)linearly increasing in the convex case and constant in the strongly convex case.

Note that similarly to Theorem 3.3, one could obtain convergence bounds for the discrete implementation under the presence of additive noise.

3.4.2. Continuized acceleration for noiseless stochastic optimization. We now set ourselves in the least-squares supervised learning setting of Example 1.5:

$$\forall x \in \mathcal{H},\ f(x) = \mathbb{E}_{(a,b)\sim\mathcal{P}}\left[\frac{1}{2}\left(b - \langle x, a\rangle\right)^2\right], \tag{3.17}$$

where $\xi = (a, b) \in \mathcal{H} \times \mathbb{R}$ is random of law $\mathcal{P}$. Our *stochastic first order oracle* is the gradient of one realization of the expectation, namely,

$$g(x, \xi) = -(b - \langle x, a\rangle)a\,, \qquad \xi = (a, b)\,.$$

As motivated by Section 1.4.3, we investigate *noiseless*—or purely multiplicative—stochastic gradients, in the sense that almost surely, for $\xi = (a, b) \sim \mathcal{P}$:

$$b = \langle x_*, a\rangle, \text{ so that } g(x_*, \xi) = 0 \text{ a.s.} \tag{3.18}$$

Let $\Sigma = \mathbb{E}[aa^\top]$ be the Hessian of f. For $x \in \mathcal{H}$, denote $\|x\|^2_{\Sigma^{-1}} = \langle x, \Sigma^{-1}x\rangle$. Let R_0 be the smallest positive real number such that:

$$\mathbb{E}\left[\|a\|^2 aa^\top\right] \preccurlyeq R_0\Sigma\,. \tag{3.19}$$

Further, similarly to Jain et al. [2018], we define the statistical condition number of the problem as the smallest $\tilde{\kappa} > 0$ such that:

$$\mathbb{E}\left[\|a\|^2_{\Sigma^{-1}} aa^\top\right] \preccurlyeq \tilde{\kappa}\Sigma\,. \tag{3.20}$$

Note that this is also the constant R_α for $\alpha = 1$ of the previous chapter, in its weak definition of Remark 2.4.

Theorem 3.5 (Continuized acceleration with pure multiplicative noise)**.** Assume that (3.18), (3.19) and (3.20) hold true. Then the continuized acceleration satisfies the following.

(1) Choose the parameters $\eta_t = \frac{2}{t}, \eta'_t = 0, \gamma_t = \frac{1}{R_0}, \gamma'_t = \frac{t}{2R_0\tilde{\kappa}}$. Then

$$\mathbb{E}\|x_t - x_*\|^2 \leqslant \frac{2R_0\tilde{\kappa}\|z_0 - x_*\|^2_{\Sigma^{-1}}}{t^2}\,.$$

(2) Assume further that f is μ-strongly convex, i.e., all eigenvalues of Σ are greater or equal to μ, where $\mu > 0$. The condition number of f is then defined as $\kappa = R_0/\mu$. For the parameters $\eta_t = \eta'_t = \frac{1}{\sqrt{\kappa\tilde{\kappa}}}$, $\gamma_t = \frac{1}{R_0}$ and $\gamma'_t = \frac{1}{R_0}\sqrt{\frac{\kappa}{\tilde{\kappa}}}$, we have:

$$\mathbb{E}\|x_t - x_*\|^2 \leqslant \left(\|x_0 - x_*\|^2 + \mu\|z_0 - x_*\|^2_{\Sigma^{-1}}\right)\exp\left(-\frac{t}{\sqrt{\kappa\tilde{\kappa}}}\right).$$

This theorem is proved in Appendix 3.B.4. In the strongly convex case, acceleration brings benefits similar to those of Jain et al. [2018] with classical discrete iterates: while stochastic gradient descent with step-size $1/R_0$ is easily shown to achieve an exponential rate of convergence $1/\kappa$, the acceleration enjoys a rate of convergence of $1/\sqrt{\kappa\tilde{\kappa}}$. Note that from the definitions, $\tilde{\kappa} \leqslant \kappa$, thus the acceleration performs as least as well as the naive algorithm. However, depending on the distribution of a, the improvement can be significant, or null. We refer the reader to the rich discussion of Jain et al. [2018] that provides insights on the interpretation of $\tilde{\kappa}$ and on the possibility to accelerate.

Below, we give a complementary perspective on the statistical condition number by translating it in terms of effective resistances in the case of gossip algorithms.

Compared to Jain et al. [2018], even though our assumptions are more restrictive, our acceleration analysis is much simpler as it relies on a standard Lyapunov function, similar to that of the continuized acceleration (Theorem 3.2).

3.5. Accelerating asynchronous gossip

The continuized framework is useful to design accelerated decentralized algorithms requiring synchronized clocks, but no synchronization of the communications. In this section, we illustrate this statement in the simple case of gossip algorithms. From Section 1.4, the gossip problem corresponds to a noiseless stochastic optimization problem, this section is thus a special case of Section 3.4.2. However, it was directly derived earlier by Even et al. [2020].

Recall that the gossip problem corresponds to a least-squares problem on the function

$$f(x) = \frac{1}{2N} \sum_{\{v,w\}\in\mathcal{E}} \langle e_v - e_w, x\rangle^2 = \frac{1}{2N} \langle x - x_*, \mathcal{L}(x - x_*)\rangle \ ,$$

where $x_* = \bar{x}\mathbb{1}$ and $\mathcal{L} = \sum_{\{v,w\}\in\mathcal{E}}(e_v - e_w)(e_v - e_w)^\top$ is the Laplacian of the graph. This is a special case of (3.17), with $a_{\{v,w\}} = e_v - e_w$ and $b_{\{v,w\}} = 0$. In this parallel, communications correspond to stochastic gradient steps, that are generated by a Poisson point measure $\mathrm{d}N(t,e) = \sum_{k\geq 1}\delta_{(T_k,\{v_k,w_k\})}$ with intensity measure $\mathrm{d}t \otimes \mathrm{Unif}(\mathcal{E})$. Note that in Section 1.3, we took the counting measure instead of the uniform measure on $\mathcal{E}$. This only has the effect of a time rescaling; we prefer to have a probability measure on $\mathcal{E}$ to fit the framework of the previous section.

The parameters of the previous section can naturally be interpreted as graph quantities here. As $\|a\|^2 = \|e_v - e_w\|^2 = 2$ a.s., we have $R_0 = 2$. For simplicity, we assume that f is μ-strongly convex for some $\mu > 0$, which corresponds to a spectral gap assumption on the Laplacian $\mathcal{L}$ (as in Section 1.5.1). Finally, $R_{\mathrm{eff}}(v,w) = (e_v - e_w)^\top\mathcal{L}^{-1}(e_v - e_w)$ is a natural quantity called the effective resistance of the graph between v and w [Ellens et al., 2011]. It corresponds to the following intuition: imagine that the graph G is an electrical circuit, with unit resistance on each of the edges. Then the effective resistance $R_{\mathrm{eff}}(v,w)$ is the potential difference between v and w, when a unit current is pushed in the circuit at v and pulled out of the circuit at w. Roughly speaking, it measures if it is easy to go from v to w in the graph G. Finally, we take $\tilde{\kappa} = N\max_{\{v,w\}\in\mathcal{E}} R_{\mathrm{eff}}(v,w)$ to be the maximal effective resistance along edges of the graph (up to a N factor).

The algorithm of Theorem 3.5.(2) gives the following gossip algorithm. Take $z_0 = x_0$. Upon activation of edge $\{v_k, w_k\}$ at time T_k,

$$\begin{aligned}
x_{T_k}(v_k) = x_{T_k}(w_k) &= \frac{x_{T_k-}(v_k) + x_{T_k-}(w_k)}{2} \ , \\
z_{T_k}(v_k) &= z_{T_k-}(v_k) + \frac{1}{\sqrt{2\mu\tilde{\kappa}}}(x_{T_k-}(w_k) - x_{T_k-}(v_k)) \ , \\
z_{T_k}(w_k) &= z_{T_k-}(w_k) + \frac{1}{\sqrt{2\mu\tilde{\kappa}}}(x_{T_k-}(v_k) - x_{T_k-}(w_k)) \ .
\end{aligned}$$

Between these updates, $x_t(v)$ and $z_t(v)$ locally mix at all nodes $v \in \mathcal{V}$, according to the coupled ODE:

$$\begin{aligned}
\mathrm{d}x_t(v) &= \sqrt{\frac{\mu}{2\tilde{\kappa}}}(z_t(v) - x_t(v))\mathrm{d}t, \\
\mathrm{d}z_t(v) &= \sqrt{\frac{\mu}{2\tilde{\kappa}}}(x_t(v) - z_t(v))\mathrm{d}t.
\end{aligned}$$

This algorithm is *asynchronous* in the sense that it does not require global synchronous operations: the mixing of local variables does not require any synchronization since parameter $t \in \mathbb{R}_{\geqslant 0}$ is available at all nodes independently from the number of past updates, while a local pairwise update between adjacent nodes v and w only requires a local synchronization.

Theorem 3.6 (Accelerated Randomized Gossip)**.** Let $(x_t(v))_{v\in\mathcal{V},t\geq 0}$ be generated with accelerated randomized gossip. For any $t \in \mathbb{R}_{\geq 0}$:

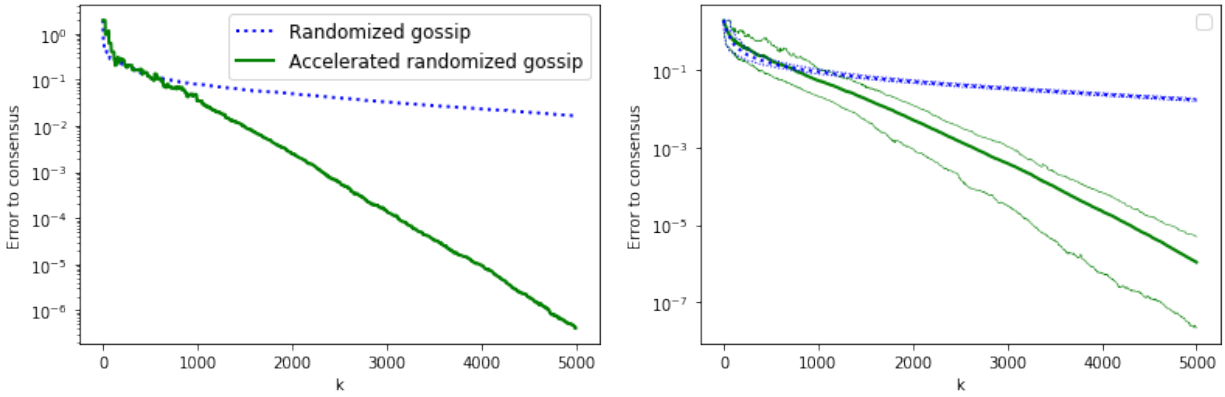
$$\sum_{v\in\mathcal{V}} \mathbb{E}\left[\left(x_t(v) - \bar{x}\right)^2\right] \leqslant 2\left(\sum_{v\in\mathcal{V}} \left(x_0(v) - \bar{x}\right)^2\right) \exp\left(-\sqrt{\frac{\mu}{2\tilde{\kappa}}}t\right).$$

Let us quickly illustrate why this is an acceleration over the simple gossip algorithm. From the intuition given above, the effective resistance only increases when we remove edges in the graph. Indeed, when removing an edge, we restrict the possibilities for the current to flow, thus increasing the effective resistance. As a consequence, we can upper bound the effective resistance $R_{\text{eff}}(v, w)$, when $\{v, w\}$ is an edge of the graph. We remove all edges in the graph but $\{v, w\}$, in which case the effective resistance between v and w is 1: there is only one possible flow. Thus $R_{\text{eff}}(v, w) \leqslant 1$ and $\tilde{\kappa} \leqslant N$.

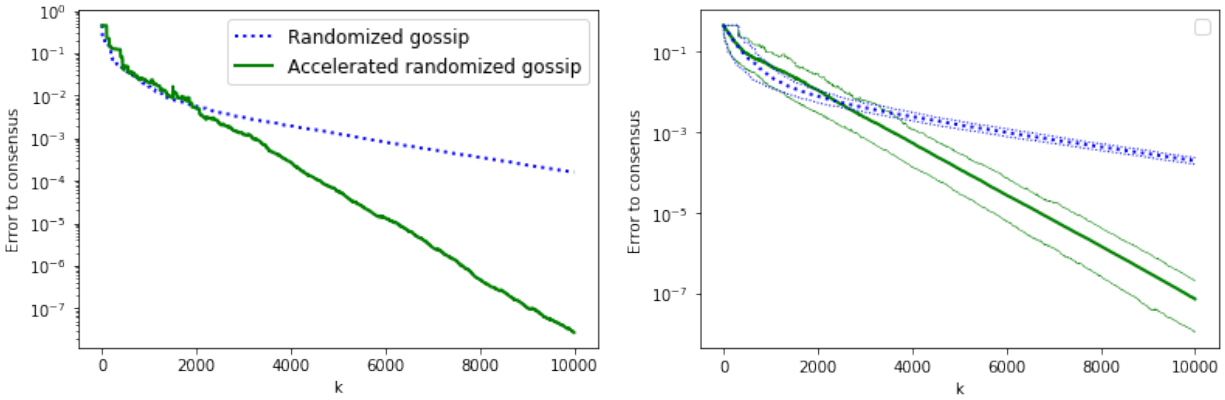
It turns out that this crude bound is sufficient to obtain a significant acceleration on the torus $\mathbb{T}^d_\Lambda$. Indeed, from Example 1.10, the spectral gap of the Laplacian of the torus scales like Λ^{-2}, thus the strong convexity parameter μ of the function f scales like $\Lambda^{-2}N^{-1}$. From Theorem 2.1, the typical time of the exponential convergence of simple gossip is thus $\Lambda^2 N$. As a comparison, the typical time of the accelerated method is $\sqrt{\frac{2\tilde{\kappa}}{\mu}} = \Theta(\Lambda N)$. Acceleration thus improves the dependence of the rate in Λ. We thus recover the same rates as Dimakis et al. [2008] for the graphs they study, but generalized to any network.

Let us also note that there is no acceleration on the complete graph K_m. Indeed, from Example 1.9, the spectral gap of the Laplacian of the torus is $\Theta(m)$ as $m \to \infty$. Thus the strong convexity parameter μ of the function f scales like $m/N = \Theta(m^{-1})$; simple gossip converges in a typical time $\Theta(m)$. Moreover, as there are $m - 1$ parallel paths of resistance 1 or 2 between any pair of vertices in the complete graph, the effective resistance between the vertices is $\Theta(m^{-1})$. Thus $\tilde{\kappa} = \Theta(Nm^{-1}) = \Theta(m)$. Thus typical times of the naive and the accelerated methods have a similar order of magnitude.

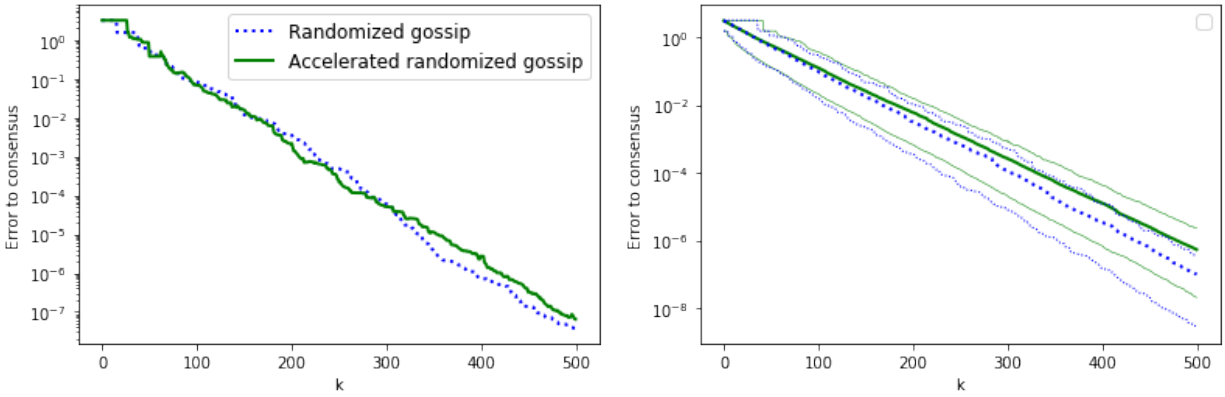
In Figure 3.3, we compare simulations of the accelerated method with the simple gossip algorithm. We observe the expected asymptotic improvement on the line graph and the 2D grid. On the complete graph, the two methods perform similarly.



(A) Line graph, 30 nodes



(B) 2D-Grid, 225 nodes



(C) Complete graph, 10 nodes

FIGURE 3.3. Comparison between the simple gossip algorithm (or randomized gossip) and the accelerated gossip method of this section, on 3 different graphs: a line graph with 30 nodes, 2D grid with 225 nodes and complete graph with 30 nodes. In all simulations, initialization was taken with a vector x_0 such that $x_0(v) = 0$ at all nodes, except one where $x_0(v) = 1$. Figures on the left represent one run of the algorithms. Figures on the right represent the average performance (thick line) for 1000 runs with the same settings, and the 5% and 95% quantiles (thin lines).

Appendix of Chapter 3

In Appendix 3.A, we start by giving a few technical tools that ground rigorously the chapter and that are used in the proofs that follow. In Appendix 3.B, we analyze the continuized Nesterov acceleration. In Appendix 3.B.1, we give a first sketch of proof of the deterministic case, without using any technical tool. We continue with the complete technical proofs: of the deterministic case in Appendix 3.B.2, of the case with additive noise in Appendix 3.B.3 and of the case with pure multiplicative noise in Appendix 3.B.4. We finish with the proof of Theorem 3.3 in Appendix 3.C and with a sanity check of the ODE scaling limit of the continuized acceleration in Appendix 3.D.

3.A. Stochastic calculus toolbox

In this appendix, we give a short introduction to the mathematical tools that we use in this paper. For more details, the reader can consult the more rigorous monographs of Jacod and Shiryaev [2013], Ikeda and Watanabe [2014], Le Gall [2016].

3.A.1. Poisson point measures. We fix $\mathcal{P}$ a probability law on some space Ξ.

Definition 3.1. A *(homogenous) Poisson point measure* on $\mathbb{R}_{\geqslant 0}\times\Xi$, with intensity $\nu(\mathrm{d}t,\mathrm{d}\xi) = \mathrm{d}t\otimes\mathrm{d}\mathcal{P}(\xi)$, is a random measure N on $\mathbb{R}_{\geqslant 0}\times\Xi$ such that

- For any disjoint measurable subsets A and B of $\mathbb{R}_{\geqslant 0}\times\Xi$, $N(A)$ and $N(B)$ are independent.
- For any measurable subset A of $\mathbb{R}_{\geqslant 0}\times\Xi$, $N(A)$ is a Poisson random variable with parameter $\nu(A)$. (If $\nu(A)=\infty$, $N(A)$ is equal to ∞ almost surely.)

Proposition 3.1. Let N be a Poisson point measure on $\mathbb{R}_{\geqslant 0}\times\Xi$ with intensity $\mathrm{d}t\otimes\mathrm{d}\mathcal{P}(\xi)$.

There exists a decomposition $\mathrm{d}N(t,\xi) = \sum_{k\geqslant 1}\delta_{(T_k,\xi_k)}(\mathrm{d}t,\mathrm{d}\xi)$ on $\mathbb{R}_{\geqslant 0}\times\Xi$ where $0<T_1<T_2<T_3<\dots$ and $\xi_1,\xi_2,\xi_3,\dots\in\Xi$ satisfy:

- $T_1, T_2-T_1, T_3-T_2,\dots$ are i.i.d. of exponential law with rate 1,
- $\xi_1,\xi_2,\xi_3,\dots$ are i.i.d. of law $\mathcal{P}$ and independent of the $T_1,T_2,T_3,\dots$.

Definition 3.2. Let N be a Poisson point measure on $\mathbb{R}_{\geqslant 0}\times\Xi$ with intensity $\mathrm{d}t\otimes\mathrm{d}\mathcal{P}(\xi)$. The *filtration $\mathcal{F}_t$, $t\geqslant 0$, generated by N* is defined by the formula

$$\mathcal{F}_t = \sigma\left(N([0,s]\times A)\,,\ s\leqslant t, A\subset\Xi \text{ measurable}\right)\,.$$

3.A.2. Martingales and supermartingales. Let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space and $\mathcal{F}_t$, $t\geqslant 0$, a filtration on this probability space.

Definition 3.3. A random process $x_t\in\mathcal{H}$, $t\geqslant 0$, is *adapted* if for all $t\geqslant 0$, x_t is $\mathcal{F}_t$-measurable. An adapted process $x_t\in\mathbb{R}$, $t\geqslant 0$ is a *martingale* (resp. *supermartingale*) if for all $0\leqslant s\leqslant t$, $\mathbb{E}[x_t|\ \mathcal{F}_s] = x_s$ (resp. $\mathbb{E}[x_t|\ \mathcal{F}_s]\leqslant x_s$).

Definition 3.4. A random variable $T \in [0, \infty]$ is a *stopping time* if for all $t \geqslant 0$, $\{T \leqslant t\} \in \mathcal{F}_t$.

Definition 3.5. A function $x_t, t \geqslant 0$, is said to be *càdlàg* if it is right continuous and for every $t > 0$, the limit $x_{t-} := \lim_{s \to t, s<t} x_s$ exists and is finite.

Theorem 3.7 (Martingale stopping theorem)**.** Let x_t, $t \geqslant 0$, be a martingale (resp. supermartingale) with càdlàg trajectories and uniformly integrable. Let T be a stopping time. Then $\mathbb{E}X_T = X_0$ (resp. $\mathbb{E}X_T \leqslant X_0$).

3.A.3. Stochastic ordinary differential equation with Poisson jumps. The continuized processes are the composition of an ordinary differential equation and stochastic Poisson jumps. It is thus a piecewise-deterministic Markov process [Davis, 1984, 2018], a special case of stochastic models that do not include any diffusion term. The stochastic calculus of this class of processes is particularly intuitive: there is no Ito correction term as with diffusive processes.

We fix $\mathcal{P}$ a probability law on some space Ξ, N a Poisson point measure on $\mathbb{R}_{\geqslant 0} \times \Xi$ with intensity $\mathrm{d}t \otimes \mathrm{d}\mathcal{P}(\xi)$, and denote $\mathcal{F}_t$, $t \geqslant 0$, the filtration generated by N.

Definition 3.6. Let $b : \mathcal{H} \to \mathcal{H}$ and $G : \mathcal{H} \times \Xi \to \mathcal{H}$ be two functions. An random process $x_t \in \mathcal{H}$, $t \geqslant 0$, is said to be a solution of the equation

$$\mathrm{d}x_t = b(x_t)\mathrm{d}t + \int_\Xi G(x_t, \xi)\mathrm{d}N(t, \xi)$$

if it is adapted, càdlàg, and for all $t \geqslant 0$,

$$x_t = x_0 + \int_0^t b(x_s)\mathrm{d}s + \int_{[0,t]\times\Xi} G(x_{s-}, \xi)\mathrm{d}N(s, \xi)\,.$$

If we consider the decomposition $\mathrm{d}N(t, \xi) = \sum_{k\geqslant 1} \delta_{(T_k,\xi_k)}(\mathrm{d}t, \mathrm{d}\xi)$ given by Proposition 3.1, then

$$\int_{[0,t]\times\Xi} G(x_{s-}, \xi)\mathrm{d}N(s, \xi) = \sum_{k\geqslant 1} \mathbb{1}_{\{T_k \leqslant t\}} G(x_{T_k-}, \xi_k)\,.$$

Here, we consider only autonomous equations as b and G are a function of x_t, but not of t. However, there is no loss of generality, one can study time-dependent systems by studying the equation in the variable (t, x_t). This trick is used in Appendix 3.B.

Proposition 3.2. Let $x_t \in \mathcal{H}$ be a solution of

$$\mathrm{d}x_t = b(x_t)\mathrm{d}t + \int_\Xi G(x_t, \xi)\mathrm{d}N(t, \xi)$$

and $\varphi : \mathcal{H} \to \mathbb{R}$ be a smooth function. Then

$$\varphi(x_t) = \varphi(x_0) + \int_0^t \langle \nabla\varphi(x_s), b(x_s)\rangle \mathrm{d}s + \int_{[0,t]\times\Xi} \left(\varphi(x_{s-} + G(x_{s-}, \xi)) - \varphi(x_{s-})\right) \mathrm{d}N(s, \xi)\,.$$

Moreover, we have the decomposition

$$\begin{aligned}\int_{[0,t]\times\Xi} &\left(\varphi(x_{s-} + G(x_{s-}, \xi)) - \varphi(x_{s-})\right) \mathrm{d}N(s, \xi) \\ &= \int_0^t \int_\Xi \left(\varphi(x_s + G(x_s, \xi)) - \varphi(x_s)\right) \mathrm{d}t\mathrm{d}\mathcal{P}(\xi) + M_t\,,\end{aligned}$$

where $M_t = \int_{[0,t]\times\Xi} \left(\varphi(x_{s-} + G(x_{s-}, \xi)) - \varphi(x_{s-})\right) \left(\mathrm{d}N(s, \xi) - \mathrm{d}t\mathrm{d}\mathcal{P}(\xi)\right)$ is a martingale.

This proposition is an elementary calculus of variations formula: to compute the value of the observable $\varphi(x_t)$, one must sum the effects of the continuous part and of the Poisson jumps. Moreover, the integral with respect to the Poisson measure N becomes a martingale if the same integral with respect to its intensity measure $\mathrm{d}t \otimes \mathrm{d}\mathcal{P}(\xi)$ is removed.

3.B. Analysis of the continuized Nesterov acceleration

To encompass the proofs in the convex and in the strongly convex cases in a unified way, we assume f is μ-strongly convex, $\mu \geqslant 0$. If $\mu > 0$, this corresponds to assuming the μ-strong convexity in the usual sense; if $\mu = 0$, it means that we only assume the function to be convex. In other words, the proofs in the convex case can be obtained by taking $\mu = 0$ below.

In this section, $\mathcal{F}_t$, $t \geqslant 0$, is the filtration associated to the Poisson point measure N.

3.B.1. Sketch of proof for Theorem 3.2. A complete and rigorous proof is given in Appendix 3.B.2. Here, we only provide the heuristic of the main lines of the proof.

The proof is similar to the one of Nesterov acceleration: we prove that for some choices of parameters $\eta_t, \eta'_t, \gamma_t, \gamma'_t$, $t \geqslant 0$, and for some functions A_t, B_t, $t \geqslant 0$,

$$\phi_t = A_t \left(f(x_t) - f(x_*)\right) + \frac{B_t}{2}\|z_t - x_*\|^2$$

is a supermartingale. In particular, this implies that $\mathbb{E}\phi_t$ is a Lyapunov function, i.e., a non-increasing function of t.

To prove that ϕ_t is a supermartingale, it is sufficient to prove that for all infinitesimal time intervals $[t, t+\mathrm{d}t]$, $\mathbb{E}_t\phi_{t+\mathrm{d}t} \leqslant \phi_t$, where $\mathbb{E}_t$ denotes the conditional expectation knowing all the past of the Poisson process up to time t. Thus we would like to compute the first order variation of $\mathbb{E}_t\phi_{t+\mathrm{d}t}$. This implies computing the first order variation of $\mathbb{E}_t f(x_{t+\mathrm{d}t})$.

From (3.8), we see that $f(x_t)$ evolves for two reasons between t and $t + \mathrm{d}t$:

- x_t follows the linear ODE (3.6), which results in the infinitesimal variation $f(x_t) \to f(x_t) + \eta_t\langle\nabla f(x_t), z_t - x_t\rangle\mathrm{d}t$, and
- with probability $\mathrm{d}t$, x_t takes a gradient step, which results in a macroscopic variation $f(x_t) \to f\left(x_t - \gamma_t\nabla f(x_t)\right)$.

Combining both variations, we obtain that

$$\mathbb{E}_t f(x_{t+\mathrm{d}t}) \approx f(x_t) + \eta_t\langle\nabla f(x_t), z_t - x_t\rangle\mathrm{d}t + \mathrm{d}t\left(f\left(x_t - \gamma_t\nabla f(x_t)\right) - f(x_t)\right),$$

where the $\mathrm{d}t$ in the second term corresponds to the probability that a gradient step happens; note that the latter event is independent of the past up to time t.

A similar computation can be done for $\mathbb{E}_t\|z_t - x_*\|^2$. Putting things together, we obtain

$$\begin{aligned}\mathbb{E}_t\phi_{t+\mathrm{d}t} - \phi_t \approx \mathrm{d}t\Bigg(&\frac{\mathrm{d}A_t}{\mathrm{d}t}(f(x_t) - f(x_*)) + A_t\eta_t\langle\nabla f(x_t), z_t - x_t\rangle \\ &- A_t\left(f(x_t - \gamma_t\nabla f(x_t)) - f(x_t)\right) + \frac{\mathrm{d}B_t}{\mathrm{d}t}\frac{1}{2}\|z_t - x_*\|^2 \\ &+ B_t\eta'_t\langle z_t - x_*, x_t - z_t\rangle + \frac{B_t}{2}\left(\|z_t - \gamma'_t\nabla f(x_t) - x_*\|^2 - \|z_t - x_*\|^2\right)\Bigg).\end{aligned}$$

Using convexity and strong convexity inequalities, and a few computations, we obtain the following upper bound:

$$\begin{aligned}\mathbb{E}_t\phi_{t+\mathrm{d}t} - \phi_t \lesssim \mathrm{d}t\Bigg(&\left(\frac{\mathrm{d}A_t}{\mathrm{d}t} - A_t\eta_t\right)\langle\nabla f(x_t), x_t - x_*\rangle + \left(\frac{\mathrm{d}B_t}{\mathrm{d}t} - B_t\eta'_t\right)\frac{1}{2}\|z_t - x_*\|^2 \\ &+ (A_t\eta_t - B_t\gamma'_t)\langle\nabla f(x_t), z_t - x_*\rangle + \left(B_t\eta'_t - \frac{\mathrm{d}A_t}{\mathrm{d}t}\mu\right)\frac{1}{2}\|x_t - x_*\|^2\end{aligned}$$

$$+\left(B_t\gamma_t'^2 - A_t\gamma_t\left(2 - L\gamma_t\right)\right)\frac{1}{2}\|\nabla f(x_t)\|^2\Big)\,.$$

We want this infinitesimal variation to be non-positive. Here, we choose the parameters so that $\gamma_t = 1/L$, and all prefactors in the above expression are zero. This gives some constraints on the choices of parameters. We show that only one degree of freedom is left: the choice of the function A_t, that must satisfy the ODE

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\left(\sqrt{A_t}\right) = \frac{\mu}{4L}\sqrt{A_t}\,,$$

but whose initialization remains free. Once the initialization of the function A_t is chosen, this determines the full function A_t and, through the constraints, all parameters of the algorithm. As ϕ_t is a supermartingale (by design), a bound on the performance of the algorithm is given by

$$\mathbb{E}f(x_t) - f(x_*) \leqslant \frac{\mathbb{E}\phi_t}{A_t} \leqslant \frac{\phi_0}{A_t}\,.$$

The results presented in Theorem 3.2 correspond to one special choice of initialization for the function A_t.

In this sketch of proof, our derivation of the infinitesimal variation is intuitive and elementary; however it can be made more rigorous and concise—albeit more technical—using classical results from stochastic calculus, namely Proposition 3.2. This is our approach in Appendix 3.B.2.

3.B.2. Deterministic case: proofs of Theorem 3.2 and of the bounds of Theorem 3.3. In this section, we analyze the convergence of the continuized iteration (3.8)-(3.9), that we recall for the reader's convenience:

$$\begin{aligned}\mathrm{d}x_t &= \eta_t(z_t - x_t)\mathrm{d}t - \gamma_t\nabla f(x_t)\mathrm{d}N(t)\,,\\ \mathrm{d}z_t &= \eta_t'(x_t - z_t)\mathrm{d}t - \gamma_t'\nabla f(x_t)\mathrm{d}N(t)\,.\end{aligned}$$

The choices of parameters $\eta_t, \eta_t', \gamma_t, \gamma_t',\ t \geqslant 0$, and the corresponding convergence bounds follow naturally from the analysis. We seek sufficient conditions under which the function

$$\phi_t = A_t\left(f(x_t) - f_*\right) + \frac{B_t}{2}\|z_t - x_*\|^2$$

is a supermartingale.

The process $\bar{x}_t = (t, x_t, z_t)$ satisfies the equation

$$\mathrm{d}\bar{x}_t = b(\bar{x}_t)\mathrm{d}t + G(\bar{x}_t)\mathrm{d}N(t)\,, \qquad b(\bar{x}_t) = \begin{pmatrix}1\\ \eta_t(z_t - x_t)\\ \eta_t'(x_t - z_t)\end{pmatrix}\,, \qquad G(\bar{x}_t) = \begin{pmatrix}0\\ -\gamma_t\nabla f(x_t)\\ -\gamma_t'\nabla f(x_t)\end{pmatrix}\,.$$

We thus apply Proposition 3.2 to $\phi_t = \varphi(\bar{x}_t) = \varphi(t, x_t, z_t)$ where

$$\varphi(t, x, z) = A_t\left(f(x) - f(x_*)\right) + \frac{B_t}{2}\|z - x_*\|^2\,,$$

we obtain:

$$\phi_t = \phi_0 + \int_0^t \langle\nabla\varphi(\bar{x}_s), b(\bar{x}_s)\rangle\mathrm{d}s + \int_0^t \left(\varphi(\bar{x}_s + G(\bar{x}_s)) - \varphi(\bar{x}_s)\right)\mathrm{d}s + M_t\,,$$

where M_t is a martingale. Thus, to show that φ_t is a supermartingale, it is sufficient to show that the map $t \mapsto \int_0^t\langle\nabla\varphi(\bar{x}_s), b(\bar{x}_s)\rangle\mathrm{d}s + \int_0^t\left(\varphi(\bar{x}_s + G(\bar{x}_s)) - \varphi(\bar{x}_s)\right)\mathrm{d}s$ is non-increasing almost surely, i.e.,

$$I_t := \langle\nabla\varphi(\bar{x}_t), b(\bar{x}_t)\rangle + \varphi(\bar{x}_t + G(\bar{x}_t)) - \varphi(\bar{x}_t) \leqslant 0\,.$$

We now compute

$$\begin{aligned}\langle \nabla\varphi(\bar{x}_t), b(\bar{x}_t)\rangle &= \partial_t\varphi(\bar{x}_t) + \langle \partial_x\varphi(\bar{x}_t), \eta_t(z_t - x_t)\rangle + \langle \partial_z\varphi(\bar{x}_t), \eta_t'(x_t - z_t)\rangle \\ &= \frac{\mathrm{d}A_t}{\mathrm{d}t}\left(f(x_t) - f(x_*)\right) + \frac{\mathrm{d}B_t}{\mathrm{d}t}\frac{1}{2}\|z_t - x_*\|^2 + A_t\eta_t\langle \nabla f(x_t), z_t - x_t\rangle \\ &\quad + B_t\eta_t'\langle z_t - x_*, x_t - z_t\rangle\,.\end{aligned}$$

Here, we use that as f is μ-strongly convex,

$$f(x_t) - f(x_*) \leqslant \langle \nabla f(x_t), x_t - x_*\rangle - \frac{\mu}{2}\|x_t - x_*\|^2\,,$$

and the simple bound

$$\begin{aligned}\langle z_t - x_*, x_t - z_t\rangle &= \langle z_t - x_*, x_t - x_*\rangle - \|z_t - x_*\|^2 \leqslant \|z_t - x_*\|\|x_t - x_*\| - \|z_t - x_*\|^2 \\ &\leqslant \frac{1}{2}\left(\|z_t - x_*\|^2 + \|x_t - x_*\|^2\right) - \|z_t - x_*\|^2 = \frac{1}{2}\left(\|x_t - x_*\|^2 - \|z_t - x_*\|^2\right)\,.\end{aligned}$$

This gives

$$\langle \nabla\varphi(\bar{x}_t), b(\bar{x}_t)\rangle \leqslant \left(\frac{\mathrm{d}A_t}{\mathrm{d}t} - A_t\eta_t\right)\langle \nabla f(x_t), x_t - x_*\rangle + \left(B_t\eta_t' - \frac{\mathrm{d}A_t}{\mathrm{d}t}\mu\right)\frac{1}{2}\|x_t - x_*\|^2 \tag{3.21}$$

$$+ \left(\frac{\mathrm{d}B_t}{\mathrm{d}t} - B_t\eta_t'\right)\frac{1}{2}\|z_t - x_*\|^2 + A_t\eta_t\langle \nabla f(x_t), z_t - x_*\rangle\,. \tag{3.22}$$

Further,

$$\begin{aligned}\varphi(\bar{x}_t + G(\bar{x}_t)) - \varphi(\bar{x}_t) &= A_t\left(f(x_t - \gamma_t\nabla f(x_t)) - f(x_t)\right) \\ &\quad + \frac{B_t}{2}\left(\|(z_t - x_*) - \gamma_t'\nabla f(x_t)\|^2 - \|z_t - x_*\|^2\right)\,.\end{aligned}$$

As f is L-smooth,

$$\begin{aligned}f(x_t - \gamma_t\nabla f(x_t)) - f(x_t) &\leqslant \langle \nabla f(x_t), -\gamma_t\nabla f(x_t)\rangle + \frac{L}{2}\|\gamma_t\nabla f(x_t)\|^2 \\ &= -\gamma_t\left(2 - L\gamma_t\right)\frac{1}{2}\|\nabla f(x_t)\|^2\,.\end{aligned}$$

This gives

$$\varphi(\bar{x}_t + G(\bar{x}_t)) - \varphi(\bar{x}_t) \leqslant \left(B_t\gamma_t'^2 - A_t\gamma_t\left(2 - L\gamma_t\right)\right)\frac{1}{2}\|\nabla f(x_t)\|^2 - B_t\gamma_t'\langle \nabla f(x_t), z_t - x_*\rangle\,. \tag{3.23}$$

Finally, combining (3.21)-(3.22) with (3.23), we obtain

$$I_t \leqslant \left(\frac{\mathrm{d}A_t}{\mathrm{d}t} - A_t\eta_t\right)\langle \nabla f(x_t), x_t - x_*\rangle + \left(\frac{\mathrm{d}B_t}{\mathrm{d}t} - B_t\eta_t'\right)\frac{1}{2}\|z_t - x_*\|^2 \tag{3.24}$$

$$+ (A_t\eta_t - B_t\gamma_t')\langle \nabla f(x_t), z_t - x_*\rangle + \left(B_t\eta_t' - \frac{\mathrm{d}A_t}{\mathrm{d}t}\mu\right)\frac{1}{2}\|x_t - x_*\|^2 \tag{3.25}$$

$$+ \left(B_t\gamma_t'^2 - A_t\gamma_t\left(2 - L\gamma_t\right)\right)\frac{1}{2}\|\nabla f(x_t)\|^2\,. \tag{3.26}$$

Remember that $I_t \leqslant 0$ is a sufficient condition for ϕ_t to be a supermartingale. Here, we choose the parameters $\eta_t, \eta_t', \gamma_t, \gamma_t', t \geqslant 0$, so that all prefactors are 0. We start by taking $\gamma_t \equiv \frac{1}{L}$ (other choices $\gamma_t < \frac{2}{L}$ could be possible but would give similar results) and we want to satisfy

$$\frac{\mathrm{d}A_t}{\mathrm{d}t} = A_t\eta_t\,, \qquad \frac{\mathrm{d}B_t}{\mathrm{d}t} = B_t\eta_t' \qquad A_t\eta_t = B_t\gamma_t'\,, \qquad B_t\eta_t' = \frac{\mathrm{d}A_t}{\mathrm{d}t}\mu\,, \qquad B_t\gamma_t'^2 = \frac{A_t}{L}\,.$$

To satisfy the last equation, we choose

$$\gamma'_t = \sqrt{\frac{A_t}{LB_t}}\,. \tag{3.27}$$

To satisfy the third equation, we choose

$$\eta_t = \frac{B_t\gamma'_t}{A_t} = \sqrt{\frac{2B_t}{LA_t}}\,. \tag{3.28}$$

To satisfy the fourth equation, we choose

$$\eta'_t = \frac{\mathrm{d}A_t}{\mathrm{d}t}\frac{\mu}{B_t} = \frac{A_t\eta_t\mu}{B_t} = \mu\sqrt{\frac{A_t}{LB_t}}\,. \tag{3.29}$$

Having now all parameters $\eta_t, \eta'_t, \gamma_t, \gamma'_t$ constrained, we now have that ϕ_t is Lyapunov if

$$\frac{\mathrm{d}A_t}{\mathrm{d}t} = A_t\eta_t = \sqrt{\frac{A_tB_t}{L}}\,, \qquad\qquad \frac{\mathrm{d}B_t}{\mathrm{d}t} = B_t\eta'_t = \mu\sqrt{\frac{A_tB_t}{L}}\,.$$

This only leaves the choice of the initialization (A_0, B_0) as free: both the algorithm and the Lyapunov depend on it. (Actually, only the relative value A_0/B_0 matters.) Instead of solving the above system of two coupled non-linear ODEs, it is convenient to turn them into a single second-order linear ODE:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\sqrt{A_t}\right) = \frac{1}{2\sqrt{A_t}}\frac{\mathrm{d}A_t}{\mathrm{d}t} = \frac{1}{2}\sqrt{\frac{B_t}{L}}\,, \qquad \frac{\mathrm{d}}{\mathrm{d}t}\left(\sqrt{B_t}\right) = \frac{1}{2\sqrt{B_t}}\frac{\mathrm{d}B_t}{\mathrm{d}t} = \frac{\mu}{2}\sqrt{\frac{A_t}{L}}\,. \tag{3.30}$$

This can also be restated as

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\left(\sqrt{A_t}\right) = \frac{\mu}{4L}\sqrt{A_t}\,, \qquad\qquad \sqrt{B_t} = 2\sqrt{L}\frac{\mathrm{d}}{\mathrm{d}t}\left(\sqrt{A_t}\right)\,. \tag{3.31}$$

Proof of the first part (convex case). We now assume $\mu = 0$, and we choose the solution such that $A_0 = 0$ and $B_0 = 1$. From (3.30), we have $\frac{\mathrm{d}}{\mathrm{d}t}\left(\sqrt{B_t}\right) = 0$, thus $B_t \equiv 1$, and $\frac{\mathrm{d}}{\mathrm{d}t}\left(\sqrt{A_t}\right) = \frac{1}{2\sqrt{L}}$, thus $\sqrt{A_t} = \frac{t}{2\sqrt{L}}$. The parameters of the algorithm are given by (3.27)-(3.29): $\eta_t = \frac{2}{t}$, $\eta'_t = 0$, $\gamma'_t = \frac{t}{2L}$ (and we had chosen $\gamma_t = \frac{1}{L}$).

From the fact that ϕ_t is a supermartingale, we obtain that the associated algorithm satisfies

$$\mathbb{E}f(x_t) - f(x_*) \leqslant \frac{\mathbb{E}\phi_t}{A_t} \leqslant \frac{\phi_0}{A_t} = \frac{2L\|z_0 - x_*\|^2}{t^2}\,.$$

This proves the first part of Theorem 3.2.

Further, one can apply martingale stopping Theorem 3.7 to the supermartingale ϕ_t with the stopping time T_k to obtain

$$\mathbb{E}\left[A_{T_k}\left(f(\tilde{x}_k) - f(x_*)\right)\right] = \mathbb{E}\left[A_{T_k}\left(f(x_{T_k}) - f(x_*)\right)\right] \leqslant \mathbb{E}\phi_{T_k} \leqslant \phi_0 = \|z_0 - x_*\|^2\,.$$

This proves the formula of Theorem 3.3.1.

Proof of the second part (strongly convex case). We now assume $\mu > 0$. We consider the solution of (3.31) that is exponential:

$$\sqrt{A_t} = \sqrt{A_0}\exp\left(\frac{1}{2}\sqrt{\frac{\mu}{L}}t\right)\,, \qquad\qquad \sqrt{B_t} = \sqrt{A_0}\sqrt{\mu}\exp\left(\frac{1}{2}\sqrt{\frac{\mu}{L}}t\right)\,.$$

The parameters of the algorithm are given by (3.27)-(3.29): $\eta_t = \eta'_t = \sqrt{\frac{\mu}{L}}$, $\gamma'_t = \frac{1}{\sqrt{\mu L}}$ (and we had chosen $\gamma_t = \frac{1}{L}$).

From the fact that ϕ_t is a supermartingale, we obtain that the associated algorithm satisfies

$$\mathbb{E}f(x_t) - f(x_*) \leqslant \frac{\mathbb{E}\phi_t}{A_t} \leqslant \frac{\phi_0}{A_t} = \frac{A_0(f(x_0) - f(x_*)) + A_0\frac{\mu}{2}\|z_0 - x_*\|^2}{A_t}$$
$$= \left(f(x_0) - f(x_*) + \frac{\mu}{2}\|z_0 - x_*\|^2\right) \exp\left(-\sqrt{\frac{\mu}{L}}t\right) .$$

This proves the second part of Theorem 3.2. Similarly to above, one can also apply the martingale stopping theorem to prove the formula of Theorem 3.3.2.

Remark 3.1. In the above derivation, in both the convex and strongly convex cases, we choose a particular solution of (3.31), while several solutions are possible. In the convex case, we make the choice $A_0 = 0$ to have a succinct bound that does not depend on $f(x_0) - f(x_*)$. More importantly, in the strongly convex case, we choose the solution that satisfies the relation $\sqrt{\mu}\sqrt{A_t} = \sqrt{B_t}$, which implies that $\eta_t, \eta_t', \gamma_t'$, are constant functions of t, and $\eta_t = \eta_t'$. These conditions help solving in closed form the continuous part of the process

$$\mathrm{d}x_t = \eta_t(z_t - x_t)\mathrm{d}t\,,$$
$$\mathrm{d}z_t = \eta_t'(x_t - z_t)\mathrm{d}t\,,$$

which is crucial if we want to have a discrete implementation of our method (for more details, see Theorem 3.3 and its proof). However, in the strongly convex case, considering other solutions would be interesting, for instance to have an algorithm converging to the convex one as $\mu \to 0$.

3.B.3. With additive noise: proof of Theorem 3.4. The proof of this theorem is along the same lines as the proof of Theorem 3.2 above. Here, we only give the major differences.

We analyze the convergence of the continuized stochastic iteration (3.13)-(3.14), that we recall for the reader's convenience:

$$\mathrm{d}x_t = \eta_t(z_t - x_t)\mathrm{d}t - \gamma_t \int_\Xi g(x_t, \xi)\mathrm{d}N(t, \xi)\,,$$
$$\mathrm{d}z_t = \eta_t'(x_t - z_t)\mathrm{d}t - \gamma_t' \int_\Xi g(x_t, \xi)\mathrm{d}N(t, \xi)\,.$$

In this setting, we loose the property that

$$\phi_t = A_t\left(f(x_t) - f_*\right) + \frac{B_t}{2}\|z_t - x_*\|^2$$

is a supermartingale. However, we bound the increase of ϕ_t.

The process $\bar{x}_t = (t, x_t, z_t)$ satisfies the equation

$$\mathrm{d}\bar{x}_t = b(\bar{x}_t)\mathrm{d}t + \int_\Xi G(\bar{x}_t, \xi)\mathrm{d}N(t, \xi), \qquad b(\bar{x}_t) = \begin{pmatrix} 1 \\ \eta_t(z_t - x_t) \\ \eta_t'(x_t - z_t) \end{pmatrix}, \qquad G(\bar{x}_t, \xi) = \begin{pmatrix} 0 \\ -\gamma_t g(x_t, \xi) \\ -\gamma_t' g(x_t, \xi) \end{pmatrix}.$$

We apply Proposition 3.2 to $\phi_t = \varphi(\bar{x}_t) = \varphi(t, x_t, z_t)$ and obtain

$$\phi_t = \phi_0 + \int_0^t I_s\mathrm{d}s + M_t\,, \tag{3.32}$$

where M_t is a martingale and

$$I_t = \langle \nabla\varphi(\bar{x}_t), b(\bar{x}_t)\rangle + \mathbb{E}_\xi\varphi(\bar{x}_t + G(\bar{x}_t, \xi)) - \varphi(\bar{x}_t)\,.$$

The computation of the first term remains the same: the inequality (3.21)-(3.22) holds. The computation of the second term becomes

$$\mathbb{E}_\xi \varphi(\bar{x}_t + G(\bar{x}_t, \xi)) - \varphi(\bar{x}_t) = A_t \left(\mathbb{E}_\xi f(x_t - \gamma_t g(x_t, \xi)) - f(x_t)\right) + \frac{B_t}{2}\left(\mathbb{E}_\xi \|(z_t - x_*) - \gamma_t' g(x_t, \xi)\|^2 - \|z_t - x_*\|^2\right).$$

As f is L-smooth,

$$f(x_t - \gamma_t g(x_t, \xi)) - f(x_t) \leqslant \langle \nabla f(x_t), -\gamma_t g(x_t, \xi)\rangle + \frac{L}{2}\|\gamma_t g(x_t, \xi)\|^2,$$
$$\mathbb{E}_\xi f(x_t - \gamma_t g(x_t, \xi)) - f(x_t) \leqslant \langle \nabla f(x_t), -\gamma_t \mathbb{E}_\xi g(x_t, \xi)\rangle + \frac{L}{2}\mathbb{E}_\xi\|\gamma_t g(x_t, \xi)\|^2.$$

By assumptions (3.15) and (3.16), the stochastic gradient $g(x, \xi)$ is unbiased and has a variance bounded by σ^2, which implies $\mathbb{E}_\xi\|g(x_t, \xi)\|^2 \leqslant \|\nabla f(x_t)\|^2 + \sigma^2$. Thus

$$\mathbb{E}_\xi f(x_t - \gamma_t g(x_t, \xi)) - f(x_t) \leqslant -\gamma_t\left(2 - L\gamma_t\right)\frac{1}{2}\|\nabla f(x_t)\|^2 + \sigma^2\frac{L\gamma_t^2}{2}.$$

Similarly,

$$\begin{aligned}\mathbb{E}_\xi\|(z_t - x_*) - \gamma_t' g(x_t, \xi)\|^2 - \|z_t - x_*\|^2 &= -2\gamma_t'\langle \mathbb{E}_\xi g(x_t, \xi), z_t - x_*\rangle + \gamma_t'^2\mathbb{E}_\xi\|g(x_t, \xi)\|^2 \\ &\leqslant -2\gamma_t'\langle \nabla f(x_t), z_t - x_*\rangle + \gamma_t'^2\|\nabla f(x_t)\|^2 + \sigma^2\gamma_t'^2.\end{aligned}$$

This gives

$$\varphi(\bar{x}_t + G(\bar{x}_t)) - \varphi(\bar{x}_t) \leqslant \left(B_t\gamma_t'^2 - A_t\gamma_t\left(2 - L\gamma_t\right)\right)\frac{1}{2}\|\nabla f(x_t)\|^2 - B_t\gamma_t'\langle \nabla f(x_t), z_t - x_*\rangle + \frac{\sigma^2}{2}\left(A_t L\gamma_t^2 + B_t\gamma_t'^2\right).$$

Combining the bounds, we obtain

$$\begin{aligned}I_t \leqslant{}& \left(\frac{\mathrm{d}A_t}{\mathrm{d}t} - A_t\eta_t\right)\langle \nabla f(x_t), x_t - x_*\rangle + \left(\frac{\mathrm{d}B_t}{\mathrm{d}t} - B_t\eta_t'\right)\frac{1}{2}\|z_t - x_*\|^2 \\ &+ (A_t\eta_t - B_t\gamma_t')\langle \nabla f(x_t), z_t - x_*\rangle + \left(B_t\eta_t' - \frac{\mathrm{d}A_t}{\mathrm{d}t}\mu\right)\frac{1}{2}\|x_t - x_*\|^2 \\ &+ \left(B_t\gamma_t'^2 - A_t\gamma_t\left(2 - L\gamma_t\right)\right)\frac{1}{2}\|\nabla f(x_t)\|^2 + \frac{\sigma^2}{2}\left(A_t L\gamma_t^2 + B_t\gamma_t'^2\right),\end{aligned}$$

which is an additive perturbation of the bound (3.24)-(3.26) in the noiseless case, with a perturbation proportional to σ^2. The choices of parameters of Theorem 3.2 cancel all first five prefactors, and satisfy $\gamma_t = \frac{1}{L}$, $A_t L\gamma_t^2 = B_t\gamma_t'^2$. We thus obtain

$$I_t \leqslant \sigma^2\frac{A_t}{L}.$$

This bound controls the increase of ϕ_t. Using the decomposition (3.32), we obtain

$$\begin{aligned}\mathbb{E}f(x_t) - f(x_*) \leqslant \frac{\mathbb{E}\phi_t}{A_t} &\leqslant \frac{\phi_0}{A_t} + \frac{\int_0^t \mathbb{E}I_s \mathrm{d}s}{A_t} \\ &\leqslant \frac{A_0(f(x_0) - f(x_*)) + B_0\|z_0 - x_*\|^2}{A_t} + \frac{\sigma^2}{L}\frac{\int_0^t A_s \mathrm{d}s}{A_t}.\end{aligned}$$

Proof of the first part (convex case). In this case, $A_t = \frac{t^2}{2L}$ and $B_0 = 1$. Thus $\int_0^t A_s \mathrm{d}s = \frac{1}{2L}\frac{t^3}{3}$. Thus

$$\mathbb{E}f(x_t) - f(x_*) \leqslant \frac{2L\|z_0 - x_*\|^2}{t^2} + \sigma^2 \frac{t}{3L}\,.$$

Proof of the second part (strongly convex case). In this case, $A_t = A_0 \exp\left(\sqrt{\frac{\mu}{L}}t\right)$ and $B_0 = A_0\frac{\mu}{2}$. Thus $\int_0^t A_s \mathrm{d}s \leqslant A_0\sqrt{\frac{\mu}{L}}^{-1} \exp\left(\sqrt{\frac{\mu}{L}}t\right) = \sqrt{\frac{L}{\mu}}A_t$. Thus

$$\mathbb{E}f(x_t) - f(x_*) \leqslant \left(f(x_0) - f(x_*) + \frac{\mu}{2}\|z_0 - x_*\|^2\right) \exp\left(-\sqrt{\frac{\mu}{L}}t\right) + \sigma^2 \frac{1}{\sqrt{\mu L}}\,.$$

3.B.4. With pure multiplicative: Proof of Theorem 3.5. The proof of this theorem mimics the proof of Theorem 3.2, with a slightly different Lyapunov function.

We recall that in Section 3.4, the function f is of the form:

$$f(x) = \mathbb{E}\left[\frac{1}{2}(\langle a, x\rangle - b)^2\right] = \frac{1}{2}\|x - x_*\|_\Sigma^2\,.$$

The Lyapunov function studied in the proof of Theorem 3.2 would then write as, for $t \in \mathbb{R}_{\geqslant 0}$:

$$\phi_t = \frac{A_t}{2}\|x_t - x_*\|_\Sigma^2 + \frac{B_t}{2}\|z_t - x_*\|^2.$$

An acceleration of stochastic gradient descent using this Lyapunov function has been done by Vaswani et al. [2019]. In order to have an analysis similar to Nesterov acceleration, the authors make a strong growth condition, which is too strong for many stochastic gradient problems and for our application to gossip algorithms. Instead, our analysis requires a bounded statistical condition number $\tilde{\kappa}$, and performs a shift in terms of dependency over Σ: $\|x - x_*\|_\Sigma^2$ becomes $\|x - x_*\|^2$, and $\|z_t - x_*\|^2$ becomes $\|z_t - x_*\|_{\Sigma^{-1}}^2$. The new Lyapunov function writes:

$$\phi_t = \frac{A_t}{2}\|x_t - x_*\|^2 + \frac{B_t}{2}\|z_t - x_*\|_{\Sigma^{-1}}^2.$$

As in Theorem 3.2, the proof consists in proving that for carefully chosen parameters, ϕ_t is a supermatingale. The process $\bar{x}_t = (t, x_t, z_t)$ satisfies the equation

$$\mathrm{d}\bar{x}_t = b(\bar{x}_t)\mathrm{d}t + \int_\Xi G(\bar{x}_t, \xi)\mathrm{d}N(t, \xi), \qquad b(\bar{x}_t) = \begin{pmatrix} 1 \\ \eta_t(z_t - x_t) \\ \eta_t'(x_t - z_t) \end{pmatrix}, \qquad G(\bar{x}_t, \xi) = \begin{pmatrix} 0 \\ -\gamma_t g(x_t, \xi) \\ -\gamma_t' g(x_t, \xi) \end{pmatrix}.$$

We apply Proposition 3.2 to $\phi_t = \varphi(\bar{x}_t) = \varphi(t, x_t, z_t)$ and obtain:

$$\phi_t = \phi_0 + \int_0^t I_s \mathrm{d}s + M_t\,,$$

where M_t is a martingale and

$$I_t = \langle \nabla\varphi(\bar{x}_t), b(\bar{x}_t)\rangle + \mathbb{E}_\xi \varphi(\bar{x}_t + G(\bar{x}_t, \xi)) - \varphi(\bar{x}_t)\,.$$

Since the Lyapunov function is not the same, we need to explicit here each term. The first term writes:

$$\begin{aligned}\langle \nabla\varphi(\bar{x}_t), b(\bar{x}_t)\rangle =& \frac{1}{2}\frac{\mathrm{d}A_t}{\mathrm{d}t}\|x_t - x_*\|^2 + \frac{1}{2}\frac{\mathrm{d}B_t}{\mathrm{d}t}\|z_t - x_*\|_{\Sigma^{-1}}^2 \\ &+ A_t\eta_t\langle x_t - x_*, z_t - x_t\rangle + B_t\eta_t'\langle \Sigma^{-1}(z_t - x_*), x_t - z_t\rangle.\end{aligned}$$

Mimicking the proof of Theorem 3.2, we write

$$\frac{1}{2}\|x_t - x_*\|^2 \leqslant \|x_t - x_*\|^2 - \frac{\mu}{2}\|x_t - x_*\|_{\Sigma^{-1}}^2,$$

and

$$\begin{aligned}\langle \Sigma^{-1}(z_t - x_*), x_t - z_t\rangle &= \langle z_t - x_*, x_t - x_*\rangle_{\Sigma^{-1}} - \|z_t - x_*\|^2_{\Sigma^{-1}} \\ &\leqslant \frac{1}{2}(\|x_t - x_*\|^2_{\Sigma^{-1}} - \|z_t - x_*\|^2_{\Sigma^{-1}}).\end{aligned}$$

Hence,

$$\begin{aligned}\langle \nabla\varphi(\bar{x}_t), b(\bar{x}_t)\rangle \leqslant \frac{\mathrm{d}A_t}{\mathrm{d}t}\|x_t - x_*\|^2 + \left(B_t\eta_t' - \frac{\mathrm{d}A_t}{\mathrm{d}t}\mu\right)\frac{1}{2}\|x_t - x_*\|^2_{\Sigma^{-1}} \\ + \left(\frac{\mathrm{d}B_t}{\mathrm{d}t} - B_t\eta_t'\right)\frac{1}{2}\|z_t - x_*\|^2_{\Sigma^{-1}} + A_t\eta_t\langle x_t - x_*, z_t - x_t\rangle\,.\end{aligned}$$

Further,

$$\begin{aligned}\varphi(\bar{x}_t + G(\bar{x}_t)) - \varphi(\bar{x}_t) = \frac{A_t}{2}\left(\|x_t - \gamma_t\nabla f(x_t,\xi) - x_*\|^2 - \|x_t - x_*\|^2\right) \\ + \frac{B_t}{2}\left(\|(z_t - x_*) - \gamma_t' g(x_t,\xi)\|^2_{\Sigma^{-1}} - \|z_t - x_*\|^2_{\Sigma^{-1}}\right)\,.\end{aligned}$$

Then, expanding and taking expectation over ξ of the first term:

$$\begin{aligned}\mathbb{E}_\xi\left[\frac{1}{2}\|x_t - \gamma_t g(x_t,\xi) - x_*\|^2 - \frac{1}{2}\|x_t - x_*\|^2\right] &= \frac{\gamma_t^2}{2}\mathbb{E}_\xi\left[\|g(x_t,\xi)\|^2\right] - \gamma_t\langle\Sigma(x_t - x_*), x_t - x_*\rangle \\ &\leqslant \left(\frac{R^2\gamma_t^2}{2} - \gamma_t\right)\|x_t - x_*\|^2_\Sigma,\end{aligned}$$

where we used the definition of R^2 in Equation (3.19):

$$\begin{aligned}\mathbb{E}_\xi\left[\|g(x_t,\xi)\|^2\right] &= (x_t - x_*)^\top\mathbb{E}\left[aa^\top aa^\top\right](x_t - x_*) \\ &= (x_t - x_*)^\top\mathbb{E}\left[\|a\|^2 aa^\top\right](x_t - x_*) \\ &\leqslant R^2(x_t - x_*)^\top\Sigma(x_t - x_*).\end{aligned}$$

The second term writes:

$$\begin{aligned}\frac{1}{2}\mathbb{E}_\xi\left[\|(z_t - x_*) - \gamma_t' g(x_t,\xi)\|^2_{\Sigma^{-1}} - \|z_t - x_*\|^2_{\Sigma^{-1}}\right] &= \frac{\gamma_t'^2}{2}\mathbb{E}_\xi\left[\|g(x_t,\xi)\|^2_{\Sigma^{-1}}\right] - \gamma_t'\langle x_t - x_*, z_t - x_*\rangle \\ &\leqslant \frac{\tilde{\kappa}\gamma_t'^2}{2}\|x_t - x_*\|^2_\Sigma - \gamma_t'\langle x_t - x_*, z_t - x_*\rangle,\end{aligned}$$

where we used the definition of $\tilde{\kappa}$ in Equation (3.20):

$$\begin{aligned}\mathbb{E}_\xi\left[\|g(x_t,\xi)\|^2_{\Sigma^{-1}}\right] &= (x_t - x_*)^\top\mathbb{E}\left[aa^\top\Sigma^{-1}aa^\top\right](x_t - x_*) \\ &= (x_t - x_*)^\top\mathbb{E}\left[a\|a\|^2_{\Sigma^{-1}}a^\top\right](x_t - x_*) \\ &\leqslant \tilde{\kappa}(x_t - x_*)^\top\Sigma(x_t - x_*).\end{aligned}$$

Combining these inequalities gives the following upper-bound on I_t:

$$\begin{aligned}I_t \leqslant \left(\frac{\mathrm{d}A_t}{\mathrm{d}t} - A_t\eta_t\right)\|x_t - x_*\|^2 + \left(\frac{\mathrm{d}B_t}{\mathrm{d}t} - B_t\eta_t'\right)\frac{1}{2}\|z_t - x_*\|^2_{\Sigma^{-1}} \\ + (A_t\eta_t - B_t\gamma_t')\langle x_t - x_*, z_t - x_*\rangle + \left(B_t\eta_t' - \frac{\mathrm{d}A_t}{\mathrm{d}t}\mu\right)\frac{1}{2}\|x_t - x_*\|^2_{\Sigma^{-1}} \\ + \left(\tilde{\kappa}B_t\gamma_t'^2 - A_t\gamma_t\left(2 - R^2\gamma_t\right)\right)\frac{1}{2}\|x_t - x_*\|^2_\Sigma\end{aligned}$$

Since $I_t \leqslant 0$ is still a sufficient condition for ϕ_t to be a supermartingale, we choose parameters such that all prefactors are equal to 0. We first take $\gamma_t = \frac{1}{R^2}$, and we want to satisfy:

$$\frac{\mathrm{d}A_t}{\mathrm{d}t} = A_t\eta_t\,, \qquad \frac{\mathrm{d}B_t}{\mathrm{d}t} = B_t\eta_t' \qquad A_t\eta_t = B_t\gamma_t'\,, \qquad B_t\eta_t' = \frac{\mathrm{d}A_t}{\mathrm{d}t}\mu\,, \qquad B_t\gamma_t'^2 = \frac{A_t}{\tilde{\kappa}R^2}\,.$$

To satisfy that last equality, we choose:

$$\gamma_t' = \sqrt{\frac{A_t}{B_t\tilde{\kappa}R^2}}.$$

The rest of the proof then follows just as in the proof of Theorem 3.B.2.

3.C. Proof of Theorem 3.3

By integrating the ODE

$$\begin{aligned}\mathrm{d}x_t &= \eta_t(z_t - x_t)\mathrm{d}t\,,\\ \mathrm{d}z_t &= \eta_t'(x_t - z_t)\mathrm{d}t\,,\end{aligned}$$

between T_k and $T_{k+1}-$, we obtain that there exists τ_k, τ_k'', such that

$$\begin{aligned}\tilde{y}_k = x_{T_{k+1}-} &= x_{T_k} + \tau_k(z_{T_k} - x_{T_k}) = \tilde{x}_k + \tau_k(\tilde{z}_k - \tilde{x}_k)\,,\\ z_{T_{k+1}-} &= z_{T_k} + \tau_k''(x_{T_k} - z_{T_k}) = \tilde{z}_k + \tau_k''(\tilde{x}_k - \tilde{z}_k)\,.\end{aligned} \tag{3.33}$$

From the first equation, we have $\tilde{x}_k = \frac{1}{1-\tau_k}(\tilde{y}_k - \tau_k\tilde{z}_k)$, which gives by substitution in the second equation,

$$\begin{aligned}z_{T_{k+1}-} &= \tilde{z}_k + \tau_k''\left(\frac{1}{1-\tau_k}(\tilde{y}_k - \tau_k\tilde{z}_k) - \tilde{z}_k\right)\\ &= \tilde{z}_k + \tau_k'(\tilde{y}_k - \tilde{z}_k)\,,\end{aligned}$$

where $\tau_k' = \frac{\tau_k''}{1-\tau_k}$.

Further, from (3.4)-(3.5), we obtain the equations

$$\tilde{x}_{k+1} = x_{T_{k+1}} = x_{T_{k+1}-} - \gamma_{T_{k+1}}\nabla f(x_{T_{k+1}-}) = \tilde{y}_k - \gamma_{T_{k+1}}\nabla f(\tilde{y}_k)\,, \tag{3.34}$$

$$\tilde{z}_{k+1} = z_{T_{k+1}} = z_{T_{k+1}-} - \gamma'_{T_{k+1}}\nabla f(x_{T_{k+1}-}) = \tilde{z}_k + \tau_k'(\tilde{y}_k - \tilde{z}_k) - \gamma'_{T_{k+1}}\nabla f(\tilde{y}_k)\,. \tag{3.35}$$

The stated equation (3.10)-(3.12) are the combination of (3.33), (3.34) and (3.35).

(1) The parameters of Theorem 3.2.(1) are $\eta_t = \frac{2}{t}, \eta_t' = 0, \gamma_t = \frac{1}{L}$ and $\gamma_t' = \frac{t}{2L}$. In this case, the ODE

$$\begin{aligned}\mathrm{d}x_t &= \eta_t(z_t - x_t)\mathrm{d}t = \frac{2}{t}(z_t - x_t)\mathrm{d}t\,,\\ \mathrm{d}z_t &= \eta_t'(x_t - z_t)\mathrm{d}t = 0\,,\end{aligned}$$

can be integrated in closed form: for $t \geqslant t_0$,

$$\begin{aligned}x_t &= z_{t_0} + \left(\frac{t_0}{t}\right)^2(x_{t_0} - z_{t_0}) = x_{t_0} + \left(1 - \left(\frac{t_0}{t}\right)^2\right)(z_{t_0} - x_{t_0})\,,\\ z_t &= z_{t_0}\,.\end{aligned}$$

In particular, taking $t_0 = T_k$, $t = T_{k+1}-$, we obtain $\tau_k = 1 - \left(\frac{T_k}{T_{k+1}}\right)^2$, $\tau_k'' = 0$ and thus $\tau_k' = \frac{\tau_k''}{1-\tau_k} = 0$. Finally, $\tilde{\gamma}_k = \gamma_{T_k} = \frac{1}{L}$ and $\tilde{\gamma}_k' = \gamma'_{T_k} = \frac{T_k}{2L}$.

(2) The parameters of Theorem 3.2.(2) are $\eta_t = \eta'_t \equiv \sqrt{\frac{\mu}{L}}, \gamma_t \equiv \frac{1}{L}$ and $\gamma'_t \equiv \frac{1}{\sqrt{\mu L}}$. In this case, the ODE

$$\mathrm{d}x_t = \eta_t(z_t - x_t)\mathrm{d}t = \sqrt{\frac{\mu}{L}}(z_t - x_t)\mathrm{d}t\,,$$
$$\mathrm{d}z_t = \eta'_t(x_t - z_t)\mathrm{d}t = \sqrt{\frac{\mu}{L}}(x_t - z_t)\mathrm{d}t\,,$$

can also be integrated in closed form: for $t \geqslant t_0$,

$$\begin{aligned} x_t &= \frac{x_{t_0} + z_{t_0}}{2} + \frac{x_{t_0} - z_{t_0}}{2}\exp\left(-2\sqrt{\frac{\mu}{L}}(t - t_0)\right) \\ &= x_{t_0} + \frac{1}{2}\left(1 - \exp\left(-2\sqrt{\frac{\mu}{L}}(t - t_0)\right)\right)(z_{t_0} - x_{t_0})\,, \\ z_t &= \frac{x_{t_0} + z_{t_0}}{2} + \frac{z_{t_0} - x_{t_0}}{2}\exp\left(-2\sqrt{\frac{\mu}{L}}(t - t_0)\right) \\ &= z_{t_0} + \frac{1}{2}\left(1 - \exp\left(-2\sqrt{\frac{\mu}{L}}(t - t_0)\right)\right)(x_{t_0} - z_{t_0})\,. \end{aligned}$$

In particular, taking $t_0 = T_k$, $t = T_{k+1}-$, we obtain $\tau_k = \tau''_k = \frac{1}{2}\left(1 - \exp\left(-2\sqrt{\frac{\mu}{L}}(T_{k+1} - T_k)\right)\right)$ and thus $\tau'_k = \frac{\tau''_k}{1-\tau_k} = \tanh\left(\sqrt{\frac{\mu}{L}}(T_{k+1} - T_k)\right)$. Finally, $\tilde{\gamma}_k = \gamma_{T_k} = \frac{1}{L}$ and $\tilde{\gamma}'_k = \gamma'_{T_k} = \frac{1}{\sqrt{\mu L}}$.

3.D. Heuristic ODE scaling limit of the continuized acceleration

3.D.1. Convex case. With the choices of parameters of Theorem 3.2.(1), the continuized acceleration is

$$\mathrm{d}x_t = \frac{2}{t}(z_t - x_t)\mathrm{d}t - \frac{1}{L}\nabla f(x_t)\mathrm{d}N(t)\,,$$
$$\mathrm{d}z_t = -\frac{t}{2L}\nabla f(x_t)\mathrm{d}N(t)\,.$$

The ODE scaling limit is obtained by taking the limit $L \to \infty$ (so that the step-size $1/L$ vanishes) and rescaling the time $s = t/\sqrt{L}$. Some law of large number argument heuristically gives us that, as $L \to \infty$, $\mathrm{d}N(t) = \mathrm{d}N(\sqrt{L}s) \approx \sqrt{L}\mathrm{d}s$. Thus in the limit, we obtain

$$\mathrm{d}x_s = \frac{2}{\sqrt{L}s}(z_s - x_s)\sqrt{L}\mathrm{d}s - \frac{1}{L}\nabla f(x_s)\sqrt{L}\mathrm{d}s\,,$$
$$\mathrm{d}z_s = -\frac{\sqrt{L}s}{2L}\nabla f(x_s)\sqrt{L}\mathrm{d}s\,.$$

The second term of the first equation becomes negligible in the limit. Thus the equations simplify to

$$\frac{\mathrm{d}x_s}{\mathrm{d}s} = \frac{2}{s}(z_s - x_s)\,,$$
$$\frac{\mathrm{d}z_s}{\mathrm{d}s} = -\frac{s}{2}\nabla f(x_s)\,.$$

Thus

$$-\frac{s}{2}\nabla f(x_s) = \frac{\mathrm{d}z_s}{\mathrm{d}s} = \frac{\mathrm{d}}{\mathrm{d}s}\left(x_s + \frac{s}{2}\frac{\mathrm{d}x_s}{\mathrm{d}s}\right) = \frac{\mathrm{d}x_s}{\mathrm{d}s} + \frac{1}{2}\frac{\mathrm{d}x_s}{\mathrm{d}s} + \frac{s}{2}\frac{\mathrm{d}^2x_s}{\mathrm{d}s^2}\,,$$

and thus

$$\frac{\mathrm{d}^2 x_s}{\mathrm{d}s^2} + \frac{3}{s}\frac{\mathrm{d}x_s}{\mathrm{d}s} + \nabla f(x_s) = 0\,.$$

This is the same limiting ODE as the one found by Su et al. [2014] for Nesterov acceleration.

3.D.2. Strongly-convex case. With the choices of parameters of Theorem 3.2.(2), the continuized acceleration is

$$\mathrm{d}x_t = \sqrt{\frac{\mu}{L}}(z_t - x_t)\mathrm{d}t - \frac{1}{L}\nabla f(x_t)\mathrm{d}N(t)\,,$$
$$\mathrm{d}z_t = \sqrt{\frac{\mu}{L}}(x_t - z_t)\mathrm{d}t - \frac{1}{\sqrt{\mu L}}\nabla f(x_t)\mathrm{d}N(t)\,.$$

Again, we take joint scaling $L \to \infty$, $s = t/\sqrt{L}$, with the approximation $\mathrm{d}N(t) \approx \sqrt{L}\mathrm{d}s$. We obtain

$$\mathrm{d}x_s = \sqrt{\frac{\mu}{L}}(z_s - x_s)\sqrt{L}\mathrm{d}s - \frac{1}{L}\nabla f(x_s)\sqrt{L}\mathrm{d}s\,,$$
$$\mathrm{d}z_s = \sqrt{\frac{\mu}{L}}(x_s - z_s)\sqrt{L}\mathrm{d}s - \frac{1}{\sqrt{\mu L}}\nabla f(x_s)\sqrt{L}\mathrm{d}s\,.$$

As before, the second term of the first equation becomes negligible in the limit. Thus the equations simplify to

$$\frac{\mathrm{d}x_s}{\mathrm{d}s} = \sqrt{\mu}(z_s - x_s)\,, \tag{3.36}$$
$$\frac{\mathrm{d}z_s}{\mathrm{d}s} = \sqrt{\mu}(x_s - z_s) - \frac{1}{\sqrt{\mu}}\nabla f(x_s)\,. \tag{3.37}$$

From (3.36), we have $z_s = x_s + \frac{1}{\sqrt{\mu}}\frac{\mathrm{d}x_s}{\mathrm{d}s}$, and by substitution in (3.37), we obtain

$$\frac{\mathrm{d}^2 x_s}{\mathrm{d}s^2} + 2\sqrt{\mu}\frac{\mathrm{d}x_s}{\mathrm{d}s} + \nabla f(x_s) = 0\,.$$

This is the so-called "low-resolution" ODE for Nesterov acceleration of Shi et al. [2018].

CHAPTER 4

Polynomial Based Iteration Methods for Accelerated Gossip

We remind that the contents of this chapter were published in the journal article:

> R. Berthier, F. Bach, P. Gaillard. Accelerated Gossip in Networks of Given Dimension using Jacobi Polynomial Iterations, 2020, *SIAM Journal on Mathematics of Data Science (SIMODS)*.

As illustrated by Example 1.11 on the torus, the rate of the exponential convergence of the synchronous simple gossip algorithm worsens significantly in many networks of interest as the size of the network increases. More precisely, define the diameter D of the network as the largest number of communication links needed to connect any two agents. While obviously, D steps of averaging are needed for any gossip method to spread information in the network, the simple gossip method may require up to $\Theta(D^2)$ communication steps to estimate the average. To reach the $O(D)$ bound, a diverse set of ideas were proposed, including second-order recursions [Cao et al., 2006, Rebeschini and Tatikonda, 2017], message passing algorithms [Moallemi and Roy, 2005], lifted Markov chain techniques [Shah, 2009], methods using Chebychev polynomial iterations [Montijano et al., 2012, Scaman et al., 2017] or inspiration arising from advection-diffusion processes [Sardellitti et al., 2010]. To our knowledge, all of these accelerated methods assume that the agents hold additional information about the network graph, such as its spectral gap. For instance, the heavy ball method [Polyak, 1964] in optimization translates into the shift-register gossip algorithm [Cao et al., 2006]:

$$\begin{aligned} x_1(v) &= \sum_{w:w\sim v} W_{v,w} x_0(w)\,, \\ x_{n+1}(v) &= \omega \sum_{w:w\sim v} W_{v,w} x_n(w) + (1-\omega) x_{n-1}(v)\,, \qquad n \geqslant 1\,, \end{aligned}$$

where ω is some simple function of the spectral gap μ. This can be rewritten more compactly as

$$x_1 = W x_0\,, \qquad\qquad x_{n+1} = \omega W x_n + (1-\omega) x_{n-1}\,. \tag{4.1}$$

This iteration obtains optimal asymptotic convergence on many graphs, with a relaxation time of the linear convergence that scales like $1/\sqrt{\mu}$ as the spectral gap μ converges to 0.

Proposition 4.1 (from [Liu et al., 2013, Theorem 2])**.** Let x_0 be an arbitrary family of initial observations and x_n the iterates of shift-register gossip defined in (4.1) with parameter

$$\omega = 2\frac{1-\sqrt{\mu(1-\mu/4)}}{(1-\mu/2)^2}\,,$$

where μ is the spectral gap of the gossip matrix W. Then

$$\limsup_{n\to\infty} \|x_n - \bar{x}\mathbb{1}\|_2^{1/n} \leqslant 1 - 2\frac{\sqrt{\mu(1-\mu/4)}-\mu/2}{1-\mu}\,.$$

Moreover, the upper bound is reached if there exists an eigenvector u of W, corresponding to the eigenvalue $1-\mu$, such that $\langle x_0, u\rangle \neq 0$.

The important consequence of this result is that the rate of convergence of the shift-register method behaves like $1 - 2\sqrt{\mu} + o(\sqrt{\mu})$ as $\mu \to 0$. This differs from simple gossip where the rate of convergence behaves like $1 - \tilde{\mu}$, see Corollary 1.1. Shift-register enjoys an accelerated rate of convergence as opposed to simple gossip which has a diffusive rate. This effect on the asymptotic rate of convergence can be seen in Figures 4.2 and 4.3.

In this chapter, we develop a gossip method based not on the spectral gap μ, but on the spectral dimension d, i.e., roughly speaking, on the density of eigenvalues of W near the upper edge of the spectrum. Looking at the upper part of the spectrum at a broader scale allows us to improve the local averaging of the gossip algorithm in the regime $n < 1/\sqrt{\mu}$. This improvement is worthy as the spectral gap μ can get arbitrarily small in large graphs (like the torus, see Example 1.11) while the spectral dimension scales well in large graphs (see Example 1.13).

The network is of spectral dimension d if the number of eigenvalues of W in $[1-E, 1]$ decreases like $E^{d/2}$ for small E ($\mu \ll E \ll 1$), see Section 4.3.3 for rigorous definitions. We see with examples below that this definition coincides with our intuition of the dimension of the graph, which is the dimension of the manifold on which the agents live. For instance, the grid with nodes $\mathbb{Z}^d$ where the nodes at distance 1 are connected, is a graph of dimension d. Thus the parameter d is much easier to know than the spectral gap μ.

In real-world situations, the practitioner reasonably knows if the network on which she implements the gossip method is of finite dimension, and if so, she also knows the dimension d. In this paper, we argue that she should run a second-order iteration with time-dependent weights

$$x_1 = a_0 W x_0 + b_0 x_0 \,, \qquad\qquad x_{n+1} = a_n W x_n + b_n x_n - c_n x_{n-1} \,. \tag{4.2}$$

where the recurrence weights a_n, b_n, c_n are given by the formulas

$$\begin{aligned} a_0 &= \frac{d+4}{2(2+d)} \,, & b_0 &= \frac{d}{2(2+d)} \,, \\ a_n &= \frac{(2n+d/2+1)(2n+d/2+2)}{2(n+1+d/2)^2} \,, & b_n &= \frac{d^2(2n+d/2+1)}{8(n+1+d/2)^2(2n+d/2)} \,, \\ c_n &= \frac{n^2(2n+d/2+2)}{(n+1+d/2)^2(2n+d/2)} \,, & n &\geqslant 1 \,. \end{aligned} \tag{4.3}$$

The motivation for these choice of weights a_n, b_n, c_n should not be obvious at first sight. It follows from a *polynomial-based* point of view on gossip algorithms.

We define a polynomial gossip method as any method combining the past iterates of the simple gossip method:

$$x_n = P_n(W) x_0 \,, \tag{4.4}$$

where P_n is a polynomial of degree smaller or equal to n satisfying $P_n(1) = 1$. The constraint $P_n(1) = 1$ ensures that $x_n = \bar{x}\mathbb{1}$ if all initial observations are the same, i.e., $x_0 = \bar{x}\mathbb{1}$. The constraint $\deg P_n \leqslant n$ ensures that the iterate x_n can be computed in n time steps. This polynomial approach is inspired from similar work done in the resolution of linear systems [Fischer, 1996] and on the load balancing problem [Diekmann et al., 1999]. The choice of an iteration is reframed as the choice of a sequence of polynomials, and the performance of the resulting gossip method depends on the spectrum of W. In this paper, we design polynomial gossip methods whose polynomials P_n, $n \geqslant 0$ satisfy a second-order recursion. This key property ensures that the resulting iterates $x_n = P_n(W)x_0$ can be computed recursively.

Simple gossip (1.16) corresponds to the particular case of the monomials $P_n(\lambda) = \lambda^n$. Shift-register gossip is a polynomial gossip method whose corresponding polynomials that can be expressed using the Chebyshev polynomials (see Proposition 4.17). As motivated below, the spectral dimension of a graph motivates to consider another choice of polynomials: the *Jacobi polynomials*,

that are well-known in the literature on orthogonal polynomials (see the Definition 4.5 of the Jacobi polynomials). This actually leads to the iteration (4.2), that we call the *Jacobi polynomial iteration*.

The Jacobi polynomial iteration (4.2) improves the convergence of the gossip method in the transitive phase $n < 1/\sqrt{\mu}$, but looses the optimal rate of convergence of shift-register gossip, because it does not use the spectral gap μ. We argue that in most applications of gossip methods, the asymptotic rate of convergence is not relevant as there is noise in the initial data x_0, thus a high precision on the result would be useless. However, we also build a gossip iteration that uses both parameters d and μ and achieves both the efficiency in the non-transitive regime and the fast rate of convergence.

This resolution of the gossip problem with inner-product free polynomial-based iterations is new, and could lead to other interesting algorithms on other types of graphs. Here, the phrase "inner-product free" comes from the literature on polynomial-based iterations for linear systems [Fischer, 1996], and refers to the fact that recurrence coefficients a_n, b_n, c_n are computed without using the gossip matrix W (but parameterized using the knowledge of d). Indeed, as the knowledge of the gossip matrix W is distributed across the graph, it would be a challenging distributed problem to compute the recurrence coefficients if they depended on W.

Although our work is inspired by iterative methods for linear systems, the Jacobi iteration that we developed for gossip can be transposed into a new idea to this literature, which can be useful for the distributed resolution of Laplacian systems over multi-agent networks.

Finally, in Section 4.7, we show that the message passing gossip iteration of Moallemi and Roy [2005] can be interpreted as an inner-product free polynomial iteration. This point of view allows to derive convergence rates of the message passing gossip on regular graphs.

Outline of this chapter. In Section 4.1, we give simulations in different types of networks of dimension 2 and 3. We show that the recursion (4.2) brings important benefits over existing methods in the non-asymptotic regime, i.e., when the observations are far from being fully mixed in the graph.

In Sections 4.2-4.3, we develop the derivation of the Jacobi polynomial iteration. Section 4.2 describes an optimal way to design polynomial-based gossip algorithms, following the lines of Fischer [1996] and Diekmann et al. [1999], and discusses its feasibility. Section 4.3 uses the notion of spectral dimension of a graph to inspire the practical Jacobi polynomial iteration (4.2).

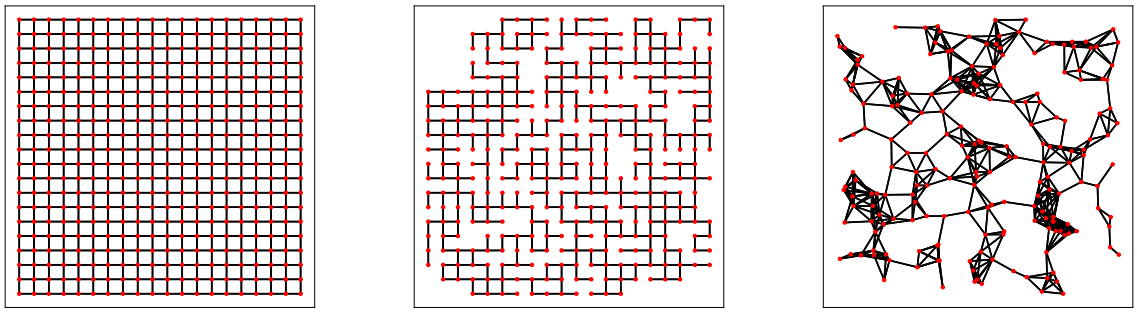
In Section 4.4, we prove some performance guarantees of the Jacobi polynomial iteration (4.2) under the assumption that the graph has spectral dimension d. As a corollary, we get performance results on two types of infinite graphs: the d-dimensional grid $\mathbb{Z}^d$ and supercritical percolation bonds in dimension d. This supports that the iteration (4.2) is robust to local perturbations of a graph.

In Section 4.5, we present the adaptation of the Jacobi polynomial iteration to the case where the spectral gap μ of W is given to improve the asymptotic rate of convergence.

In Section 4.6, we describe the parallel between gossip methods and iterative methods for linear systems, and discuss the contributions that our work can bring to the distributed resolution of Laplacian systems over networks.

In Section 4.7, we show how the message passing gossip algorithm can be interpreted as a polynomial gossip algorithm. We give the convergence rate of message passing in terms of the spectral gap μ.

Code. The code that generated the simulation results and the figures of this paper is available on the GitHub page `https://github.com/raphael-berthier/jacobi-polynomial-iterations`.



(A) Grid (B) Percolation bond (C) Random geometric graph

FIGURE 4.1. The three types of two-dimensional graphs considered in simulations.

4.1. Simulations: comparison of simple gossip, shift-register gossip and the Jacobi polynomial iteration

In this section, we run our methods on grids, percolation bonds and random geometric graphs, the latter being a widely used model for real-world networks [Penrose, 2003, Section 1.1]. In each case, we consider both the two-dimensional (2D) structure and its three-dimensional (3D) counterpart. We refer to Figure 4.1 for visualizations of the 2D structures, and to Appendix 4.A for details about the parameters used.

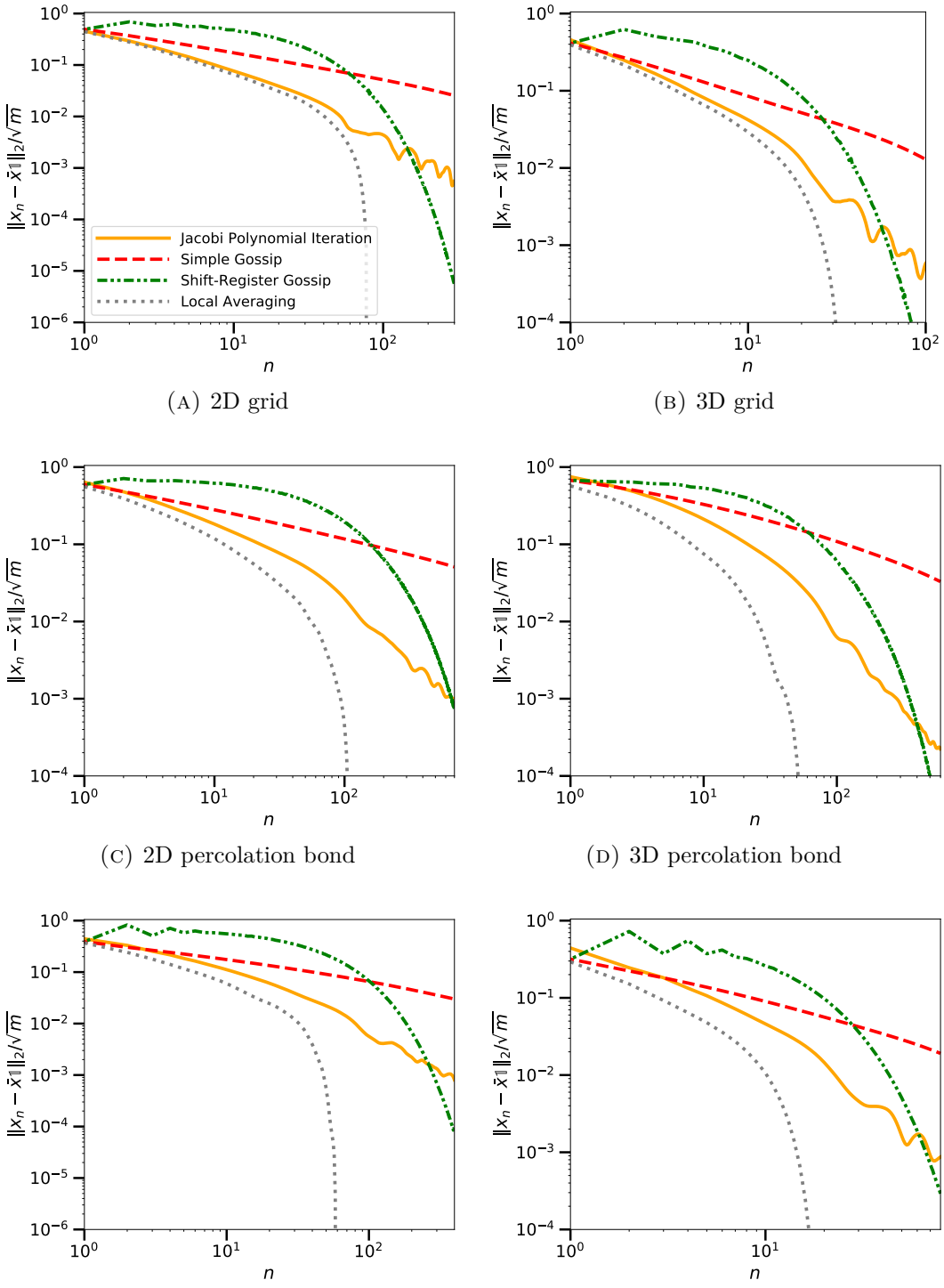
We compare our Jacobi polynomial iteration (4.2) with the simple gossip method (1.16) and the shift-register algorithm (4.1). We found experimentally that the behavior of the shift-register algorithm was typical of methods based on the spectral gap such as the splitting algorithm of Rebeschini and Tatikonda [2017] or the Chebychev polynomial acceleration scheme [Arioli and Scott, 2014, Scaman et al., 2017]; to avoid redundancy we do not present the similar behavior of these methods. We also compare with local averaging, which is given by the formula

$$x_n(v) = \frac{1}{|B_n(v)|} \sum_{w \in B_n(v)} x_0(v)\,,$$

where $|B_n(v)|$ denotes the ball in G, centered in v, of radius n, for the shortest path distance. Note that local averaging does not correspond in general to any computationally cheap iteration, as opposed to the algorithms we present here. Thus it should not be considered as a gossip method, but rather as a lower bound on the performance achievable by any gossip method. (This is made fully rigorous in the statistical gossip framework of Section 4.4.)

In our simulations, we change the graph G that we run our algorithms on, but we always sample $x_0(v) \sim_{\text{i.i.d.}} \mathcal{N}(0,1), v \in \mathcal{V}$ and measure the performance of gossip methods through the quantity $\|x_n - \bar{x}\mathbb{1}\|_2/\sqrt{m}$, where again m denotes the number of agents. Thus the performance of the algorithms is random because the initial values $x_0(v)$ are random, and also because percolation bonds and random geometric graphs are random. The results we present here are averaged over 10 realizations of the graph and the initial values, which is sufficient to give stable results.

Tuning. The optimal tuning of the shift-register gossip method as a function of the spectral gap was determined in [Liu et al., 2013, Theorem 2], it is given by the formula (4.1); this is the tuning that we use in our simulations. The Jacobi polynomial iteration is tuned by choosing $d = 2$ in 2D grid, 2D percolation bonds and 2D random geometric graphs, and $d = 3$ for their 3D analogs.



(A) 2D grid

(B) 3D grid

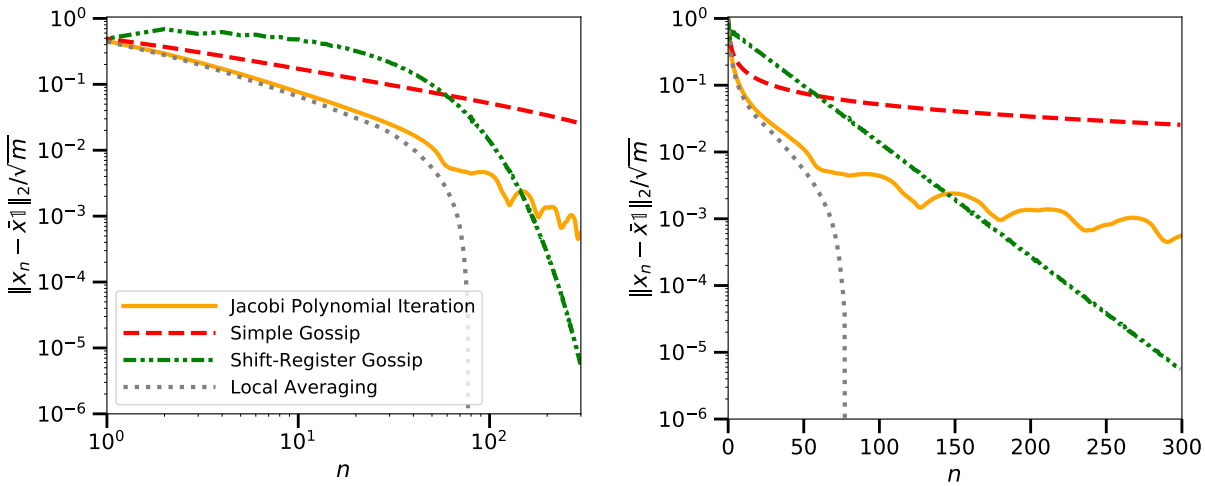
(C) 2D percolation bond

(D) 3D percolation bond

(E) 2D random geometric graph

(F) 3D random geometric graph

FIGURE 4.2. Performance of different gossip algorithms running on graphs with an underlying low-dimensional geometry, as measured by $\|x_n - \bar{x}\mathbb{1}\|_2/\sqrt{m}$.



(A) in logarithmic scale

(B) in semi-logarithmic scale

FIGURE 4.3. Performance of different gossip algorithms running on the 2D grid.

Interpretation of the results. The results of the simulations are exposed in Figure 4.2. The qualitative picture remains the same across different graphs. Simple gossip performs better than shift-register gossip in a first phase, but in a large n asymptotic, simple gossip converges slowly where shift-register gossip converges quickly. Instead, the Jacobi polynomial iteration accelerates over simple gossip for all values of n. The Jacobi polynomial iteration gets considerably closer to the local averaging optimal bound, especially in very regular structures like grids.

These results should be mitigated with the large n asymptotic: in Figure 4.3, we show the comparison of gossip methods on a longer time scale, in linear and log-scale y-axis. We only present the results on the 2D grid as they are typical of the behavior on other structures. We observe that shift-register gossip enjoys a much better asymptotic rate of convergence than simple gossip and the Jacobi polynomial iteration.

Methods that use the spectral gap are designed to achieve the best possible asymptotic (see [Cao et al., 2006, Rebeschini and Tatikonda, 2017]), thus the above observation is not surprising. These methods however fail in the non-asymptotic regime, where they are outperformed by the Jacobi polynomial iteration and simple gossip. We believe that in applications where a high precision on the average is not needed, the Jacobi polynomial iteration brings important improvements over existing methods, let alone the fact that it is considerably easier to tune. However, in Section 4.5, we present a Jacobi polynomial iteration that uses the spectral gap of the gossip matrix to obtain the accelerated convergence rate.

4.2. Design of best polynomial gossip iterations

We now turn to the design of efficient polynomial iterations of the form $x_n = P_n(W)x_0$. An important result of this section is that the best iterates of this form can be computed in an online fashion as they result from a second-order recurrence relation.

The approach presented in this section is similar to [Diekmann et al., 1999, Section 3.3], although therein it is applied to the slightly different problem of load balancing. We repeat here the derivations as we take a slightly different approach: here we derive the best polynomial P_n with fixed W and x_0; while in [Diekmann et al., 1999] the matrix W is fixed, but a polynomial P_n efficient uniformly over x_0 is sought. We then discuss why the resulting recursion may be impractical. The

next section introduces some approximation of the impractical scheme that leads to the practical iteration (4.2).

Our measure of performance of a polynomial gossip iteration is the sum of squared errors over the agents of the network:

$$\mathscr{E}(P_n) = \sum_{v\in\mathcal{V}} (x_n(v) - \bar{x})^2 = \|x_n - \bar{x}\mathbb{1}\|_2^2 = \|P_n(W)x_0 - \bar{x}\mathbb{1}\|_2^2 .$$

Denote $\lambda_1, \lambda_2, \ldots, \lambda_m$ the real eigenvalues of the symmetric matrix W and $u_1, u_2, \ldots, u_m$ are the associated eigenvectors, normalized such that $\|u_i\|_2 = 1$. The diagonalization of W gives the new expression of the error

$$\mathscr{E}(P_n) = \sum_{i=2}^{m} \langle x_0, u_i\rangle^2 P_n(\lambda_i)^2 = \int_{-1}^{1} P_n(\lambda)^2 \mathrm{d}\sigma(\lambda) , \qquad \mathrm{d}\sigma(\lambda) = \sum_{i=2}^{m} \langle x_0, u_i\rangle^2 \delta_{\lambda_i} , \tag{4.5}$$

where $\langle ., . \rangle$ denotes the canonical scalar product on $\mathbb{R}^n$ and δ_λ is the Dirac mass at λ.

The polynomial π_n minimizing the error $\mathscr{E}(P_n)$ must be chosen as

$$\pi_n \in \operatorname*{argmin}_{P(1)=1,\, \deg P \leqslant n} \int_{-1}^{1} P(\lambda)^2 \mathrm{d}\sigma(\lambda) . \tag{4.6}$$

We now show that the sequence of best polynomials $\pi_0, \pi_1, \pi_2, \ldots$ can be computed as the result of a second-order recursion, which leads to a second-order gossip method, whose coefficients depend on σ. As noted by Cao et al. [2006], having iterates x_n that satisfy a low-order recurrence relation is valuable as it ensures that they can be computed online with limited memory cost. In order to prove this property for our iterates, we use that these polynomials are orthogonal with respect to some measure τ.

Definition 4.1 (Orthogonal polynomials w.r.t. τ)**.** Let τ be a measure on $[-1, 1]$. Endow the set of polynomials $\mathbb{R}[X]$ with the scalar product

$$\langle P, Q\rangle_\tau = \int_{\mathbb{R}} P(\lambda)Q(\lambda)\mathrm{d}\tau(\lambda) .$$

Denote $\bar{n} \in \mathbb{N}\cup\{\infty\}$ the cardinal of the support of τ. Then there exists a family $\pi_0, \pi_1, ..., \pi_{\bar{n}-1}$ of polynomials, such that for all $n < \bar{n}$, $\pi_0, \pi_1, ..., \pi_n$ form an orthogonal basis of $(\mathbb{R}_n[X], \langle ., .\rangle_\tau)$, where $\mathbb{R}_n[X]$ denotes the set of polynomials of degree smaller or equal to n. In other words, for all $k, n < \bar{n}$,

$$\deg \pi_n = n , \qquad\qquad \langle \pi_k, \pi_n\rangle_\tau = 0 \qquad \text{if } k \neq n .$$

$\pi_0, \pi_1, ..., \pi_{\bar{n}-1}$ is called a sequence of *orthogonal polynomials with respect to τ (w.r.t. τ)*. Moreover, the family of orthogonal polynomials $\pi_0, \pi_1, ..., \pi_{\bar{n}-1}$ is unique up to a rescaling of each of the polynomials.

An extensive reference on orthogonal polynomials is the book [Szegö, 1939]. An introduction from the point of view of applied mathematics can be found in [Gautschi, 2004]. In Appendix 4.B, we recall the results from the theory of orthogonal polynomials that we use in this paper. The next proposition states that the optimal polynomials sought in (4.6) are orthogonal polynomials.

Proposition 4.2. Let σ be some finite measure on $[-1, 1]$ and let $\bar{n} \in \mathbb{N} \cup \{\infty\}$ be the cardinal of $\operatorname{Supp}\sigma - \{1\}$. For $0 \leqslant n \leqslant \bar{n} - 1$, the minimizer π_n of

$$\min_{P(1)=1,\, \deg P\leqslant n} \int_{-1}^{1} P(\lambda)^2 \mathrm{d}\sigma(\lambda)$$

is unique. Moreover, $\pi_0, \dots, \pi_{\bar{n}-1}$ is the unique sequence of orthogonal polynomials w.r.t. $\mathrm{d}\tau(\lambda) = (1-\lambda)\mathrm{d}\sigma(\lambda)$ normalized such that $\pi_n(1) = 1$.

This result is well-known and usually stated without proof [Nevai, 1986, Sections 3, 4.1], [Nevai, 1979, Section 2]; we give the short proof in Appendix 4.D. In the following, the phrase "*the* orthogonal polynomials w.r.t. τ" will refer to the unique family of orthogonal polynomials w.r.t. τ *and normalized such that* $\pi_n(1) = 1$.

Remark 4.1. When $\bar{n}$ is finite and $n \geqslant \bar{n}$, finding a minimizer of $\int_{-1}^{1} P(\lambda)^2 \mathrm{d}\sigma(\lambda)$ over the set of polynomials such that $P(1) = 1$, $\deg P \leqslant n$ is trivial. Indeed, one can consider the polynomial

$$\pi_{\bar{n}}(\lambda) = \frac{\prod_{\lambda' \in \operatorname{Supp}\sigma - \{1\}} (\lambda - \lambda')}{\prod_{\lambda' \in \operatorname{Supp}\sigma - \{1\}} (1 - \lambda')}$$

which is of degree $\bar{n}$, satisfies $\pi_{\bar{n}}(1) = 1$ and $\int_{-1}^{1} \pi_{\bar{n}}(\lambda)^2 \mathrm{d}\sigma(\lambda) = \sigma(\{1\})$. This is the best value that a polynomial P of any degree, such that $P(1) = 1$, can get.

A fundamental result on orthogonal polynomials states that they follow a second-order recursion.

Proposition 4.3 (Three-term recurrence relation, from [Szegö, 1939, Theorem 3.2.1])**.** Let $\pi_0, \dots, \pi_{\bar{n}-1}$ be a sequence of orthogonal polynomials w.r.t. some measure τ. There exist three sequences of coefficients $(a_n)_{1 \leqslant n \leqslant \bar{n}-2}$, $(b_n)_{1 \leqslant n \leqslant \bar{n}-2}$ and $(c_n)_{1 \leqslant n \leqslant \bar{n}-2}$ such that for $1 \leqslant n \leqslant \bar{n} - 2$,

$$\pi_{n+1}(\lambda) = (a_n \lambda + b_n)\pi_n(\lambda) - c_n \pi_{n-1}(\lambda)\,.$$

The classical proof of this proposition is given in Appendix 4.B.1. Taking σ to be the spectral measure of (4.5) in Proposition 4.2, we get that the best polynomial gossip algorithm is a second-order method whose coefficients are determined by the graph G, the gossip matrix W and the vertex v. Indeed, as $\pi_0, \dots, \pi_{\bar{n}-1}$ is a family of orthogonal polynomials, there exists coefficients a_n, b_n, c_n such that

$$\pi_{n+1}(\lambda) = (a_n \lambda + b_n)\pi_n(\lambda) - c_n \pi_{n-1}(\lambda)\,,$$

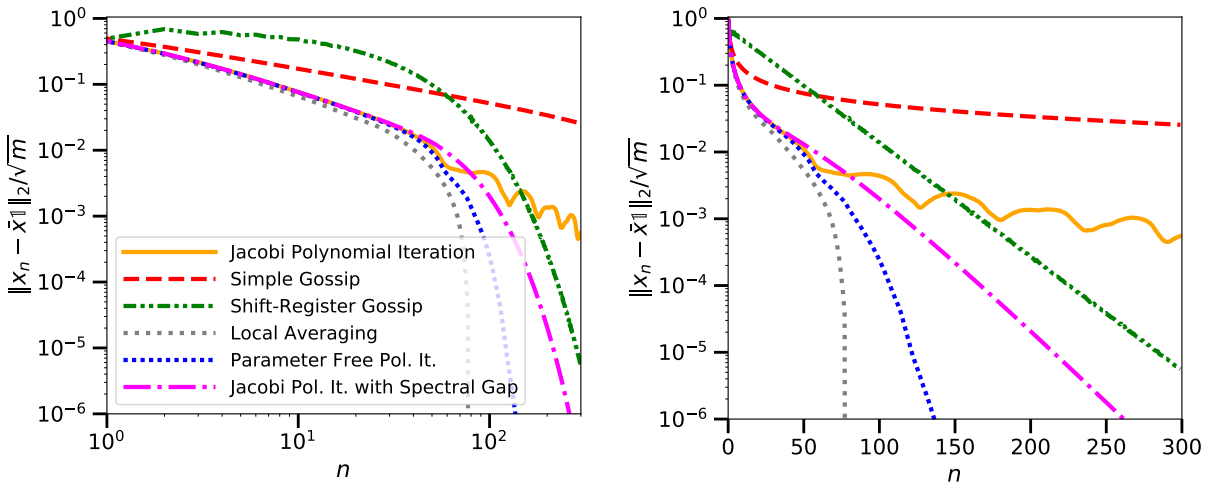
and thus

$$\pi_{n+1}(W) = a_n W \pi_n(W) + b_n \pi_n(W) - c_n \pi_{n-1}(W)\,.$$

Decomposing $\pi_1(\lambda) = a_0\lambda + b_0$ and applying the previous relation in x_0 gives the second-order recursion for the best polynomial estimators $x_n = \pi_n(W)x_0$:

$$x_1 = a_0 W x_0 + b_0 x_0\,, \qquad\qquad x_{n+1} = a_n W x_n + b_n x_n - c_n x_{n-1}\,. \tag{4.7}$$

Note that the dependence of the gossip method in the graph G, the gossip matrix W and the vertex v is entirely hidden in the coefficients a_n, b_n, c_n. Thus the choice of the coefficients is central. In [Diekmann et al., 1999], it is argued that the coefficients can be computed in a "preprocessing step". Indeed, the coefficients can be computed in a centralized or decentralized manner, at the cost of many extra communication steps. The gossip method that consists in computing the optimal coefficients a_n, b_n, c_n and running Eq. (4.7) will be refered to as *parameter-free polynomial iteration*, as it does not require any tuning of parameters, and by analogy with the terminology used in polynomial methods for the resolution of linear systems (see [Fischer, 1996, Section 6]). It corresponds to the optimal polynomial iteration. For a detailed exposition on the parameter-free polynomial iteration and a discussion of its practicability, see Appendix 4.E.



(A) in logarithmic scale

(B) in semi-logarithmic scale

FIGURE 4.4. Performance of different gossip algorithms running on the 2D grid.

However, in situations where a long preprocessing time is not granted (for instance because the network evolves during time), it is not a valid option to keep repeating the preprocessing step to update the coefficients a_n, b_n, c_n. Our approach consists in observing that there are sequences of coefficients like (4.3) that —albeit they are not optimal— work reasonably well on a large set of graphs. This implies that even if the details of the graph are not known to the algorithmic designer, she can make a choice of coefficients that have a fair performance.

More formally, we approximate the true spectral measure σ of the graph with a simpler measure $\tilde{\sigma}$, whose associated polynomials have known recursion coefficients a_n, b_n, c_n. We will show that in some cases, substituting the orthogonal polynomials w.r.t. σ with the ones orthogonal to $\tilde{\sigma}$ does not worsen the efficiency of the gossip method much. In the next sections, we argue for two choices of the approximating measure $\tilde{\sigma}$. The first uses only the spectral dimension d of the network, and gives the Jacobi polynomial iteration (4.2). The second one uses both the spectral dimension d and the spectral gap μ of W, and gives the Jacobi polynomial iteration with spectral gap.

Figure 4.4 reproduces Figure 4.3 and adds the performance of the parameter-free polynomial iteration and the Jacobi polynomial iteration with spectral gap. It shows that in linear scale, the performance of the parameter-free polynomial iteration is indistinguishable from the performance of the Jacobi polynomial iterations with or without spectral gap, which are obtained through approximations of the spectral measure σ. However, the figure in log-scale shows that the asymptotic convergence of the methods depends on the coarseness of the approximation. The relevance of this asymptotic convergence to the practice depends on the application.

Remark 4.2. The shift-register iteration $x_n = P_n(W)x_0$ defined in (4.1) can be seen as a best polynomial gossip iteration with some approximating measure. Indeed, the polynomials P_n, $n \geqslant 0$ are the orthogonal polynomials w.r.t. some measure whose support is strictly included in $[-1, 1]$ (see Proposition 4.18).

4.3. Design of polynomial gossip algorithms for graphs of given spectral dimension

4.3.1. The dimension d and the rate of decrease of the spectral measure near 1. We now assume that we are given a graph G on which we would like to run the optimal polynomial gossip

algorithm (4.7). However, we know neither the spectral measure σ, nor the coefficients a_n, b_n, c_n. In this section, we give a heuristic motivating an approximation $\tilde{\sigma}$ of the spectral measure σ using only the dimension d of the graph. The heuristic is supported by the simulations of Section 4.1 and some rigorous theoretical support in Section 4.4.

Our approximation is given by the following non-rigorous intuition:

$$\text{the graph } G \text{ is of dimension } d \qquad \Leftrightarrow \qquad \sigma([1-E,1]) \approx CE^{d/2} \qquad \text{as } E \ll 1\,, \tag{4.8}$$

for some constant C. Of course, we have not defined the dimension of a graph, nor given a rigorous signification of the symbols "$\approx$" and "$\ll$". We come back to these questions in Section 4.3.3, but for now we assume that the reader has an intuitive understanding of these notions and finish drawing the heuristic picture.

The intuition (4.8) describes the repartition of the mass of σ near 1. This mass near 1 challenges the design of polynomial methods as the gossip polynomials P are constrained to satisfy $P(1) = 1$ while minimizing $\int P^2 \mathrm{d}\sigma$. Moreover, eigenvalues of a graph close to 1 are known to describe the large-scale structure of the graph and thus must be central in the design of gossip methods. The traditional design of gossip algorithms considered the spectral gap μ between 1 and the second largest eigenvalue, a quantity that typically gets very small in large graphs. Intuition (4.8) also describes the behavior of the spectrum near 1, but on a larger scale than the spectral gap. It describes how the set of the largest eigenvalues is distributed around 1.

4.3.2. The Jacobi iteration for graphs of given dimension. When a spectral measure satisfies the edge estimate (4.8), we approximate it with a measure satisfying the same estimate, namely

$$\mathrm{d}\tilde{\sigma}(\lambda) = (1-\lambda)^{d/2-1} \mathbb{1}_{\{\lambda \in (-1,1)\}} \mathrm{d}\lambda\,.$$

Note that we do not elaborate on the normalization of the approximate measure $\mathrm{d}\tilde{\sigma}$ as it is only used to define an orthogonality relation between polynomials, in which the normalization does not matter. The orthogonal polynomials w.r.t the modified spectral measure $(1-\lambda)\mathrm{d}\tilde{\sigma}(\lambda) = (1-\lambda)^{d/2}\mathbb{1}_{\{\lambda\in(-1,1)\}}\mathrm{d}\lambda$ and their recursion coefficients are known as they correspond to the well-studied Jacobi polynomials [Szegö, 1939, Chapter IV]:

$$\begin{aligned}
a_0^{(d)} &= \frac{d+4}{2(2+d)}\,, & b_0^{(d)} &= \frac{d}{2(2+d)}\,,\\
a_n^{(d)} &= \frac{(2n+d/2+1)(2n+d/2+2)}{2(n+1+d/2)^2}\,, & b_n^{(d)} &= \frac{d^2(2n+d/2+1)}{8(n+1+d/2)^2(2n+d/2)}\,,\\
c_n^{(d)} &= \frac{n^2(2n+d/2+2)}{(n+1+d/2)^2(2n+d/2)}\,.
\end{aligned} \tag{4.9}$$

These coefficients are derived in Appendix 4.H.2. This approximation of the spectral measure gives the practical recursion

$$x_1 = a_0^{(d)} W x_0 + b_0^{(d)} x_0\,, \qquad x_{n+1} = a_n^{(d)} W x_n + b_n^{(d)} x_n - c_n^{(d)} x_{n-1}\,, \tag{4.10}$$

that only depends on d. It is just a rewriting of the Jacobi polynomial iteration (4.2) given in the introduction of this paper. The Jacobi polynomial $\pi_n^{(d/2,0)}(\lambda)$ such that $x_n = \pi_n^{(d/2,0)}(W)x_0$ is plotted in Figure 4.5 with $d = 2$ and $n = 6$, along with the polynomial λ^6 associated with simple gossip. The Jacobi polynomial is smaller in magnitude near the edge of the spectrum.

The shape of the diffusion of the Jacobi polynomial iteration on grids is shown in Figures 5.1-5.2; it is the subject of Chapter 5 to study it in detail.

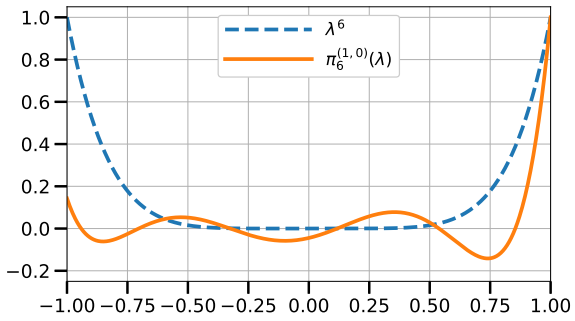


FIGURE 4.5. Comparison of the Jacobi polynomial $\pi_6^{(1,0)}(\lambda)$ with the polynomial of simple gossip λ^6.

4.3.3. Spectral dimension of a graph. In this section, we discuss the meaning of intuition (4.8). There are several definitions of the dimension of a graph.

When referring to the dimension of a graph, many authors actually refer to some quantity d that has been used in the construction of the graph. An example is the d-dimension grid $\{1,\dots,n\}^d$. Another example consists in removing edges in $\mathbb{Z}^d$ with probability $1-p$, independently of one another. The resulting graph G is called a percolation bond [Grimmett, 1999]. It is natural to consider that this graph is of dimension d. A more complicated example is the random geometric graph: choose $d \geqslant 1$, and sample n points uniformly in the d-dimensional cube $[0,1]^d$, and connect with an edge all pairs of points closer than some chosen distance $r > 0$. It is natural to say that this random geometric graph is d-dimensional as it is the dimension of the surface it is built on.

Mathematicians have developed more intrinsic definitions of the dimension of a graph [Durhuus, 2009]; here we use the notion of *spectral dimension*. This definition is of interest only for infinite graphs $G = (\mathcal{V}, \mathcal{E})$. Here, we consider only locally finite graphs, meaning that each node has only a finite number of neighbors. As with Definition 1.3, one can define a gossip matrix W with entries indexed by $\mathcal{V} \times \mathcal{V}$. If G is infinite, W is a doubly infinite array, but with only a finite number of non-zero elements in each line and column as the graph is locally finite.

The spectral dimension of a graph G is defined using a random walk on the graph, typically the simple random walk on G, but here we consider more generally the lazy random walk with transition matrix $\tilde{W} = (I + W)/2$. (We take the *lazy* random walk to avoid periodicity issues.)

Definition 4.2 (Spectral dimension)**.** Denote p_n the probability that the lazy random walk, when started from v, returns at v at iteration n. The spectral dimension of the graph is, if it exists and is finite, the limit

$$d_s = d_s(G, W, v) = -2 \lim_{n\to\infty} \frac{\log p_n}{\log n}\,.$$

If the graph is connected and W is the transition matrix of the simple random walk, this definition does not depend on the choice of the vertex v. Motivations for this definition are:

Proposition 4.4. The spectral dimension of $\left(\mathbb{Z}^d, W\right)$ with $W = A(\mathbb{Z}^d)/d$ is d.

Proposition 4.5 (The spectral dimension of the supercritical percolation cluster is d)**.** Let G_0 be a supercritical percolation bond in $\mathbb{Z}^d$ with edge probability $p \in (p_c, 1]$, meaning that a.s., there is an infinite connected component G in G_0. Endow G with the gossip matrix

$W = I + (A - D)/(2d)$, where A and D are respectively the adjacency and the degree matrices of G. Fix $v \in \mathcal{V}$. Then a.s. on the event $\{v \in G\}$, $d_s(G, W, v) = d$.

The proofs of Propositions 4.4, 4.5 are given in Appendix 4.F. The spectral dimension of a graph is related to the decay of the spectrum of W near 1.

Definition 4.3 (Spectral measure of a possibly infinite graph)**.** Let G be a graph and W its gossip matrix. Fix $v \in \mathcal{V}$. As W is an auto-adjoint operator, bounded by 1, acting on $\ell^2(\mathcal{V})$, there exists a unique positive measure $\sigma = \sigma(G, W, v)$ on $[-1, 1]$, called the *spectral measure*, such that for all polynomial P,

$$\langle e_v, P(W)e_v \rangle_{\ell^2(\mathcal{V})} = \int_{-1}^{1} P(\lambda) \mathrm{d}\sigma(\lambda) \,,$$

where $(e_w)_{w \in \mathcal{V}}$ is the canonical basis of $\mathbb{R}^{\mathcal{V}}$.

For a deeper presentation of spectral graph theory, see [Mohar and Woess, 1989] and references therein. Note that when the graph G is finite, it is easy to check that the spectral measure is the discrete measure $\sigma(G, W, v) = \sum_{i=1}^{m} (u_i(v))^2 \delta_{\lambda_i}$ where $\lambda_1, \ldots, \lambda_m$ are the eigenvalues of W and $u_1, \ldots, u_m$ are the associated normalized eigenvectors. However, when the graph G is infinite, the spectrum may exhibit a continuous part w.r.t. the Lebesgue measure.

Proposition 4.6 (The spectral dimension is the spectral decay)**.** Let G be a graph, W a gossip matrix on G and v a vertex. We denote $d_s = d_s(G, W, v)$ the spectral dimension and $\sigma = \sigma(G, W, v)$ the spectral measure. Then the limit $\lim_{E \to 0} \log \sigma([1 - E, 1]) / \log E$ exists and is finite if and only if d_s exists and is finite. In that case,

$$\lim_{E \to 0} \frac{\log \sigma([1 - E, 1])}{\log E} = \frac{d_s}{2} \,.$$

For a proof, see Appendix 4.G. This proposition gives a rigorous equivalent to intuition (4.8). It uses the spectral dimension of the graph, which is an intrinsic property of the graph and turns out to coincide with our intuition of the dimension of a graph in examples of interest. Note that in Section 4.2, the spectral measure σ is defined as $\mathrm{d}\sigma(\lambda) = \sum \langle x_0, u_i \rangle^2 \delta_{\lambda_i}$ whereas in this section, it is defined for finite graphs as $\mathrm{d}\sigma(\lambda) = \sum u_i(v)^2 \delta_{\lambda_i}$. Roughly speaking, intuition (4.8) is valid for the former if x_0 projects evenly on all eigenvectors u_i. It is the case if x_0 has random i.i.d. components for instance; this is used in Section 4.4.

4.4. Performance guarantees in graphs of spectral dimension d

In this section, we seek to give theoretical support to the empirical observations of Section 4.1: Jacobi polynomial gossip improves on the non-asymptotic phase over existing methods. This is challenging because the analysis of gossip methods is simpler in the asymptotic regime. In our case, we use asymptotic properties of the Jacobi polynomials as $n \to \infty$.

In order to be able to run an asymptotic analysis without falling in the asymptotic phase of exponential convergence, we run our method on infinite graphs $G = (\mathcal{V}, \mathcal{E})$. In infinite graphs, it is impossible for information to have reached every node in any finite time. In practice, the conclusions drawed on infinite graphs should be taken as approximations of the behavior on very large graphs.

Of course, it is impossible for any gossip method to estimate the average of the values in the infinite graphs: indeed, within time n the node v can only share information with nodes that are closer than n (w.r.t. the shortest path distance in the graph). Even worse, the average of an

infinite number of values is ill-defined. Thus additional assumptions on the observations $x_0(v)$ are needed. Several choices could be possible here, to keep the discussion simple we assume that the observations $x_0(v)$ are independent identically distributed (i.i.d.) samples from a probability law ν. The agents then seek to estimate the statistical mean $\mu = \int_{\mathbb{R}} x \, \mathrm{d}\nu(x)$ of ν.

In practice, to build good estimates, the nodes should average their samples, thus it is natural to run gossip algorithms in this situation. An estimator performs well if it averages a lot of samples and averages them uniformly. Thus the mean square error (MSE) of the estimators measures the capacity of a gossip methods to average locally in the graph.

This statistical gossip framework was already present in [Braca et al., 2008] and is not only used for its technical advantages. It is also a reasonable modeling of gossip of signals with a statistical structure in large networks. For instance, in sensor networks, observations are measurements of the environment corrupted by noise. The purpose of the gossip algorithm is to average observations to get a better estimate of the ground truth. Gossip algorithms are also used as building blocks in distributed statistical learning problems such as distributed optimization (see [Nedic and Ozdaglar, 2009, Scaman et al., 2017, Sayed, 2014, Sundhar Ram et al., 2010, Duchi et al., 2012, Chen and Sayed, 2012]) or distributed bandit algorithms (see [Szorenyi et al., 2013, Landgren et al., 2016, Korda et al., 2016]). All of these problems have a statistical structure that simplifies the underlying gossip problem. For instance, in sensor networks, good estimates of the mean may not require using observations from nodes extremely far in the network.

Let us now sum up the setting. The network of agents is modeled by a (possibly infinite, locally finite) graph $G = (\mathcal{V}, \mathcal{E})$, that we endow with a gossip matrix W. We consider a probability law ν on $\mathbb{R}$, and $\mu = \int_{\mathbb{R}} x \, \mathrm{d}\nu(x)$ its statistical mean. Each agent $v \in \mathcal{V}$ is given a sample from ν:

$$x_0(v), v \in \mathcal{V} \underset{\text{i.i.d.}}{\sim} \nu \, .$$

The following theorem gives the asymptotic MSE of the estimators built by the simple gossip method and the Jacobi polynomial iteration.

Theorem 4.1. Fix a vertex v and denote $d_s = d_s(G, W, v)$ the spectral dimension of the graph.

(1) Let x_n be the iterates of the simple gossip method (1.16), or the iterates of the shift-register gossip method (4.1) with some parameter $\omega \in [1, 2]$. Then

$$\liminf_{n\to\infty} \frac{\log \mathbb{E}[(x_n(v) - \mu)^2]}{\log n} \geqslant -\frac{d_s}{2} \, . \tag{4.11}$$

(2) Let x_n be the iterates of the Jacobi polynomial iteration (4.10) with parameter $d = d_s$. Then

$$\limsup_{n\to\infty} \frac{\log \mathbb{E}[(x_n(v) - \mu)^2]}{\log n} \leqslant -d_s \, . \tag{4.12}$$

See Appendix 4.I for a proof. The above theorem shows that the asymptotic MSE of the Jacobi polynomial iteration can be upper bounded using only the spectral dimension of the graph. The power decay of the MSE with the Jacobi polynomial iteration enjoys a better rate than with simple gossip and the shift-register iteration (regardless of the choice of ω). In some cases, this rate can be proved optimal using the Hausdorff dimension of the graph.

Definition 4.4 (Hausdorff dimension)**.** The Hausdorff dimension of the graph G at vertex v is, if it exists, the limit

$$d_h = d_h(G, v) = \lim_{n\to\infty} \frac{\log |B_n(v)|}{\log n}\,.$$

If G is connected, then d_h does not depend on the choice of v.

Proposition 4.7. Let $x_n = P_n(W)x_0$ be any polynomial gossip method on a graph G with Hausdorff dimension d_h. Then

$$\liminf_{n\to\infty} \frac{\log \mathbb{E}[(x_n(v) - \mu)^2]}{\log n} \geqslant -d_h\,. \tag{4.13}$$

See Appendix 4.J for a proof. Note that this lower bound it attained if x_n is the local average of values:

$$x_n(v) = \frac{1}{|B_n(v)|} \sum_{w\in B_n(v)} x_0(w)\,.$$

Thus reaching this lower bound means that the polynomial gossip method averages locally. Theorem 4.1 shows that it is the case with the Jacobi polynomial iteration if $d = d_s = d_h$.

Corollary 4.1. Assume that the spectral and the Hausdorff dimensions have the same value $d = d_h = d_s$. If x_n are the iterates of the Jacobi polynomial iteration (4.10), we obtain the optimal asymptotic convergence rate

$$\lim_{n\to\infty} \frac{\log \mathbb{E}[(x_n(v) - \mu)^2]}{\log n} = -d_h\,.$$

Application to the grid. Proposition 4.4 states that the spectral dimension of $\mathbb{Z}^d$ is d, which coincides the Hausdorff dimension.

Corollary 4.2. Let x_n be the iterates of the Jacobi polynomial iteration (4.10) on the grid $\mathbb{Z}^d$. Then we obtain the optimal asymptotic convergence rate

$$\lim_{n\to\infty} \frac{\log \mathbb{E}[(x_n(v) - \mu)^2]}{\log n} = -d\,.$$

Note that Theorem 4.1 also gives that if x_n are the iterates of the simple gossip method, then $\lim_{n\to\infty} \log \mathbb{E}[(x_n(v) - \mu)^2]/\log n = -d/2$. (The theorem actually only gives the lower bound, but the proof technique, combined with the fact that the spectrum of $\mathbb{Z}^d$ is symmetric, actually gives the result.) This result could have been anticipated intuitively as follows. Under the simple gossip iteration, the information of the measurement $x_0(v)$ diffuses following a simple random walk on the grid. According to the central limit theorem, at large n, the information is approximately distributed according to a Gaussian distribution of standard deviation $\sqrt{n}$, which is approximately supported by $\Theta(\sqrt{n}^d)$ nodes. This means that at time n, a node v gets the information of $\Theta(n^{d/2})$ neighbors. As a consequence, the MSE $\mathbb{E}[(x_n(v) - \mu)^2]$ scales like $n^{-d/2}$.

Application to the percolation bonds. Let G be the random infinite cluster of a supercritical percolation in $\mathbb{Z}^d$ as defined in Proposition 4.5. The proposition gives that the spectral dimension of G is a.s. d, which is also a lower bound for the Hausdorff dimension. But it is trivial that the Hausdorff dimension is smaller than d, thus the two coincide.

Corollary 4.3. Let G be the random infinite cluster of a supercritical percolation in $\mathbb{Z}^d$, and $v \in \mathbb{Z}^d$. Let x_n be the iterates of the Jacobi polynomial iteration (4.10). Then a.s. on the event $\{v \in G\}$,

$$\lim_{n\to\infty} \frac{\log \mathbb{E}_{x_0}[(x_n(v)-\mu)^2]}{\log n} = -d\,.$$

Remark 4.3. The Jacobi polynomial iteration (4.10) is derived so that $x_n = \pi_n^{(\alpha,\beta)}(W)x_0$, where $\pi_n^{(\alpha,\beta)}$ are the orthogonal polynomials w.r.t. the Jacobi measure $\sigma^{(\alpha,\beta)}(\mathrm{d}\lambda) = (1-\lambda)^\alpha(1+\lambda)^\beta \mathrm{d}\lambda$ on $[-1,1]$, with $\alpha = d/2, \beta = 0$, d is the spectral dimension. A curious reader could wonder what happens for other choices of α and β (while keeping d fixed). This question is investigated at length in Appendix 4.I.5. The conclusion is that the natural choice $\alpha = d/2, \beta = 0$ is optimal (up to constant factors) but there are other choices that are optimal.

4.5. The Jacobi polynomial iteration with spectral gap

In this section, we adapt the Jacobi polynomial iteration to the case where the spectral gap μ of the gossip matrix W is given. This allows to obtain accelerated asymptotic rates of convergence, that compete with the state-of-the-art accelerated algorithms for gossip.

We assume that we are given the spectral dimension d of the graph, which determines the density of eigenvalues near 1, and the spectral gap $\mu = 1 - \lambda_2(W)$, the distance between the largest and the second largest eigenvalue. Given these parameters, we can approximate the spectral measure of W with

$$\mathrm{d}\tilde{\sigma}(\lambda) = ((1-\mu)-\lambda)^{d/2-1}\mathbb{1}_{\{\lambda\in(-1,1-\mu)\}}\mathrm{d}\lambda\,.$$

Following the recommendation of Proposition 4.2, this means that we should consider the polynomial iteration associated with the orthogonal polynomials w.r.t. $(1-\lambda)\mathrm{d}\tilde{\sigma}(\lambda) = (1-\lambda)((1-\mu)-\lambda)^{d/2-1}\mathbb{1}_{\{\lambda\in(-1,1-\mu)\}}\mathrm{d}\lambda$. We do not know how to compute the recurrence formula for this measure, thus we used the orthogonal polynomials w.r.t. $((1-\mu)-\lambda)\mathrm{d}\tilde{\sigma}(\lambda) = ((1-\mu)-\lambda)^{d/2}\mathbb{1}_{\{\lambda\in(-1,1-\mu)\}}\mathrm{d}\lambda$, which is a rescaled version of a Jacobi measure. The corresponding polynomial method is called the *Jacobi polynomial iteration with spectral gap*.

A recursive formula for orthogonal polynomials w.r.t. $((1-\mu)-\lambda)\mathrm{d}\tilde{\sigma}(\lambda)$ is derived in Section 4.H.3. Taking $\alpha = d/2$ and $\beta = 0$ in equations (4.26), we get the practical recursion:

$$\begin{aligned}
&x_n = \frac{y_n}{\delta_n}\,,\\
&y_0 = x_0\,, \qquad \delta_0 = 1\,,\\
&y_1 = a_0^{(d,\mu)}Wx_0 + b_0^{(d,\mu)}x_0\,, \qquad \delta_1 = a_0^{(d,\mu)} + b_0^{(d,\mu)}\,,\\
&y_{n+1} = a_n^{(d,\mu)}Wy_n + b_n^{(d,\mu)}y_n - c_n^{(d,\mu)}y_{n-1}\,, \qquad t \geqslant 1\,,\\
&\delta_{n+1} = \left(a_n^{(d,\mu)} + b_n^{(d,\mu)}\right)\delta_n - c_n^{(d,\mu)}\delta_{n-1}\,, \qquad n \geqslant 1\,,\\
&a_n^{(d,\mu)} = a_n^{(d)}\left(1-\frac{\mu}{2}\right)^{-1}, \qquad b_n^{(d,\mu)} = b_n^{(d)} + \frac{\mu}{2}\left(1-\frac{\mu}{2}\right)^{-1}a_n^{(d)}\,, \qquad n \geqslant 0\,,\\
&c_n^{(d,\mu)} = c_n^{(d)}\,, \qquad n \geqslant 1\,,
\end{aligned} \tag{4.14}$$

where the coefficients $a_n^{(d)}, b_n^{(d)}, c_n^{(d)}$ are defined in (4.9).

Theorem 4.2 (Asymptotic rate of convergence)**.** Let $\mu > 0$ be a lower bound on the spectral gap of the gossip matrix W and d any positive real. Let x_0 be any family of initial

observations and x_n be the sequence of iterates generated by the Jacobi polynomial iteration with spectral gap (4.14). Then

$$\limsup_{n\to\infty} \|x_n - \bar{x}\mathbb{1}\|_2^{1/n} \leqslant \frac{1-\mu/2}{(1+\sqrt{\mu/2})^2}\,.$$

This shows that the Jacobi polynomial iteration with spectral gap enjoys linear convergence. The asymptotic rate of convergence is equivalent to $1-\sqrt{2\mu}$ as $\mu \to 0$. This justifies that we obtain an accelerated asymptotic rate of convergence that compares with the state-of-the art accelerated gossip methods (see Figure 4.4).

Note that the asymptotic rate of convergence does not depend on d. However, the choice of d may have an important effect during the non-asymptotic phase $n < 1/\sqrt{\mu}$. In this phase, the spectral gap μ can be neglected in the approximation of the spectral measure, and it is important that the densities of eigenvalues of σ and $\tilde{\sigma}$ match near the upper edge of the spectrum. This is why one should choose d as the spectral dimension of the graph.

4.6. The parallel between the gossip methods and distributed Laplacian solvers

There is a natural parallel between gossip methods and iterative methods that solve linear systems. Loosely speaking, simple gossip corresponds to gradient descent on the quadratic minimization problem associated to the linear system, shift-register gossip to Polyak's heavy-ball method (1.2) and the parameter-free polynomial iteration to the conjugate gradient algorithm (see [Fischer, 1996] or [Polyak, 1987] for references on these subjects). In this parallel, the fact that we can reach perfect gossip in n steps (see Remark 4.1) translates into the finite convergence of the conjugate gradient algorithm in a number of iterations equal to the dimension of the ambient space. In the distributed resolution of linear systems, the problem that the recursion coefficients a_n, b_n, c_n can not be computed in a centralized manner has also appeared and it motivated the development of inner-product free iterations.

The Jacobi polynomial iterations presented above were motivated by the facts that (a) the parameter-free polynomial iteration is not feasible in the distributed setting of gossip, and (b) the gossip matrix W exhibits a structure due to the low-dimension manifold on which the agents live. Interestingly, the literature on multi-agent systems deals with some minimization problems with the same properties. Examples are given by the estimation of quantities on graphs from relative measurements, in which the agents $v \in \mathcal{V}$ try to estimate some quantity $x(v), v \in \mathcal{V}$ defined over the graph, from noisy relative measurements over the edges of the graph:

$$\xi(v,w) = x(v) - x(w) + \eta(v,w)\,, \qquad \{v,w\} \in \mathcal{E}\,.$$

This problem has applications in network localization, where the $x(v)$ are the positions of the agents and the $\xi(v,w)$ come from measurements of the distances and directions between the neighbors. It also has similar applications in time synchronization of clocks over networks, where $x(v)$ is the offset of the clock of node v; and to motion consensus, where $x(v)$ is the speed of agent v. For an introduction to estimation on graphs from relative measurements and its applications, see [Barooah and Hespanha, 2008] and references therein. Note that the quantities $x(v)$ can only be determined up to a global constant from the measurements; either we seek the true solution up to a constant only, either we assume that some agents know their true value.

In natural approach to solve the problem is to determine estimates $y(v)$ of $x_(v)$ that minimize

$$\frac{1}{2}\sum_{v,w} W_{v,w}\left(\xi(v,w) - (y(v)-y(w))\right)^2\,,$$

where $W_{v,w}$ are some weights on the edges of the graph. Indeed, this corresponds to finding the maximum likelihood estimator if the noise $\eta(v,w)$ is i.i.d. Gaussian and $W_{v,w}$ is the inverse variance of $\eta(v,w)$. The above minimization problem is a quadratic problem whose covariance matrix is the Laplacian $I-W$. It can be solved using gradient descent or spectral-gap based accelerations like the heavy-ball method. However, the conjugate gradient algorithm can not be applied here as it involves centralized computations. The Jacobi polynomial iterations developed in this paper can be adapted to this situation in order to develop accelerations exploiting the structure of the Laplacian $I-W$. Experimenting how this performs in real-world situations is left for future work.

4.7. Message passing seen as a polynomial gossip algorithm

This section develops another application of the polynomial point-of-view on gossip algorithms. It is independent of the Jacobi polynomial iterations developed in Sections 4.3-4.5; we show that the message passing algorithm for gossip of Moallemi and Roy [2005] has a natural derivation as a polynomial gossip algorithm and uses this point of view to derive convergence rates.

The message passing algorithm of Moallemi and Roy [2005] (in its zero-temperature limit) defines quantities on the edges of the graph G with the following recursion: for $v,w\in\mathcal{V}$ linked by an edge in the graph G, it defines $K_0(v,w)=0$, $M_0(v,w)=0$, and

$$K_{n+1}(v,w) = 1+\sum_{u:u\sim v,\, u\neq w} K_n(u,v)\,, \tag{4.15}$$

$$M_{n+1}(v,w) = \frac{1}{K_{n+1}(v,w)}\Big(x_0(v)+\sum_{u:u\sim v,\, u\neq w} K_n(u,v)M_n(u,v)\Big)\,, \tag{4.16}$$

where $u\sim v$ denotes again that u and v are neighbors. $K(v,w)$ and $M(v,w)$ are interpreted as messages going from v to w in G: $M_n(v,w)$ corresponds to an average of observations gathered by v and transmitted to w; $K_n(v,w)$ is the corresponding number of observations. We recommend [Moallemi and Roy, 2005, Section II.A] and Lemma 4.8 for a detailed description of this intuition. At each time step n, the output of the algorithm is

$$x_n(v) = \frac{x_0(v)+\sum_{u:u\sim v} K_,(u,v)M_n(u,v)}{1+\sum_{u:u\sim v} K_n(u,v)}\,. \tag{4.17}$$

This gossip methods performs exact local averaging on trees, as shown by the following proposition.

Proposition 4.8. Assume that G is a tree. Then for all $n\geqslant 1$, $v\in\mathcal{V}$,

$$x_n(v) = \frac{1}{|B_v(n)|}\sum_{w\in B_v(n)} x_0(w)\,.$$

See Appendix 4.L for a proof. Nothing prevents from running the message passing recursion (4.16)-(4.17) in a graph G with loops. In the case of regular graphs, we are able to interpret the message passing algorithm as a polynomial gossip algorithm.

Theorem 4.3. Assume G is d-regular, meaning that each vertex has degree d, $d\geqslant 2$. Assume further that $W=A(G)/d$. Denote $\sigma(\mathbb{T}_d)=\sigma(\mathbb{T}_d,W,v)$ the spectral measure of the infinite d-regular tree at any vertex v (see Definition 4.3). Then the output x_n of the message passing algorithm (4.16)-(4.17) on G can also be obtained as $x_n=\pi_n(W)x_0$ where $\pi_0,\pi_1,\dots$ are the orthogonal polynomials w.r.t. $(1-\lambda)\sigma(\mathbb{T}_d)(\mathrm{d}\lambda)$.

See Appendix 4.M for a proof. In words, the theorem above states that message passing corresponds to the best polynomial gossip algorithm when one *believes* the graph is a tree. This is not surprising as message passing algorithms are often derived by neglecting loops in a graph.

An easy follow-up of this theorem is that the iterates x_n defined in (4.16)-(4.17) follow a second-order recursion (in d-regular graphs). Actually the spectral measure $\sigma(\mathbb{T}_d)$ of the infinite d-regular tree, also called the Kesten-McKay measure, can be computed explicitly (see [Sodin, 2007, Section 2.2]),

$$\sigma(\mathbb{T}_d)(\mathrm{d}\lambda) = \frac{d}{2\pi(1-\lambda^2)}\left(\frac{4(d-1)}{d^2} - \lambda^2\right)^{1/2} \mathbb{1}_{[-2\sqrt{d-1}/d, 2\sqrt{d-1}/d]}(\lambda)\mathrm{d}\lambda\,.$$

The recurrence relation of the modified Kesten-McKay measure $(1-\lambda)\sigma(\mathbb{T}_d)(\mathrm{d}\lambda)$ is derived in Appendix 4.H.4. It shows that

$$x_1 = a_0 W x_0 + b_0 x_0\,, \qquad x_{n+1} = a_n W x_n - c_n x_{n-1}\,,$$

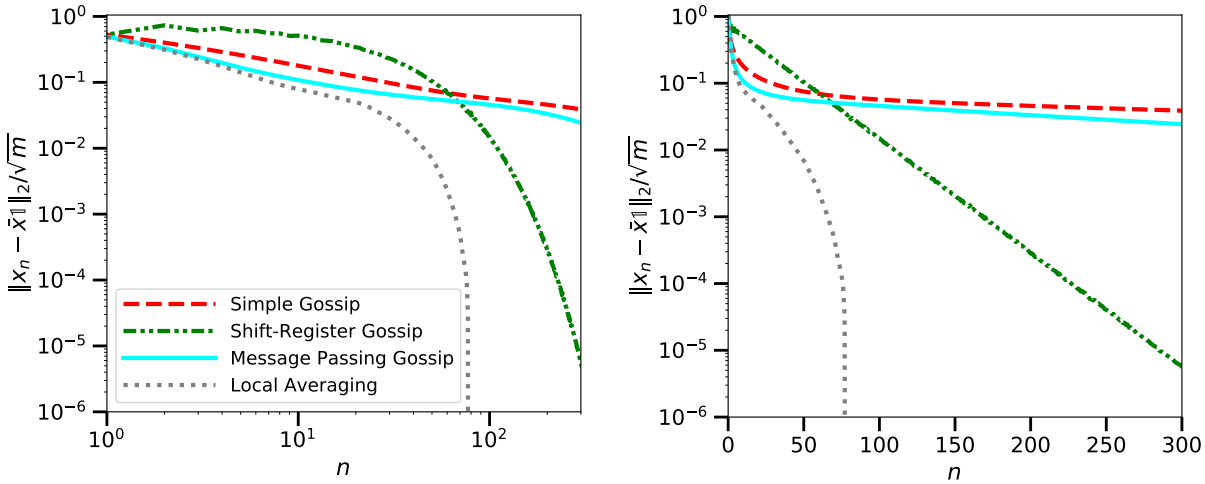
$$a_0 = \frac{d}{d+1}\,, \qquad b_0 = \frac{1}{d+1}\,, \qquad a_n = \frac{\frac{d}{d-1} - 2(d-1)^{-(n+1)}}{1 - \frac{2}{d}(d-1)^{-(n+1)}}\,, \qquad c_n = \frac{\frac{1}{d-1} - \frac{2}{d}(d-1)^{-n}}{1 - \frac{2}{d}(d-1)^{-(n+1)}}\,, \quad n \geqslant 1\,.$$

Theorem 4.3 gives a way to study the convergence of the message passing algorithms on d-regular graphs with loops. For instance, using asymptotic properties of the orthogonal polynomials w.r.t. $(1-\lambda)\sigma(\mathbb{T}_d)(\mathrm{d}\lambda)$, we obtain the convergence rate of the message passing algorithm as a function of the spectral gap of the matrix:

> **Theorem 4.4.** Assume G is d-regular, meaning that each vertex has degree d, $d \geqslant 3$. Assume further that $W = A(G)/d$, and denote $\tilde{\mu}$ its absolute spectral gap. Let x_0 be any family of initial observations and x_n be the sequence of iterates generated by equations (4.16)-(4.17). Then
>
> (1) If $\tilde{\mu} < 1 - 2\sqrt{d-1}/d$,
>
> $$\limsup_{n\to\infty} \|x_n - \bar{x}\mathbb{1}\|_2^{1/n} \leqslant \frac{(1-\tilde{\mu}) + \sqrt{(1-\tilde{\mu})^2 - 4(d-1)/d^2}}{1 + \sqrt{1 - 4(d-1)/d^2}}\,.$$
>
> Moreover, the upper bound is reached if there exists an eigenvector u corresponding to an eigenvalue of W of magnitude $1-\tilde{\mu}$ such that $\langle u, x_0\rangle \neq 0$.
>
> (2) If $\tilde{\mu} \geqslant 1 - 2\sqrt{d-1}/d$,
>
> $$\limsup_{n\to\infty} \|x_n - \bar{x}\mathbb{1}\|_2^{1/n} \leqslant \frac{2\sqrt{d-1}/d}{1 + \sqrt{1 - 4(d-1)/d^2}}\,.$$

A consequence of this theorem is that the rate of convergence of the message passing algorithm is $1 - c\tilde{\mu} + o(\tilde{\mu})$ as $\tilde{\mu} \to 0$, for some constant c. This proves that message passing has a diffusive (or unaccelerated) behavior on graphs with a small spectral gap. Figure 4.6 shows this diffusive convergence rate on the 2D grid.

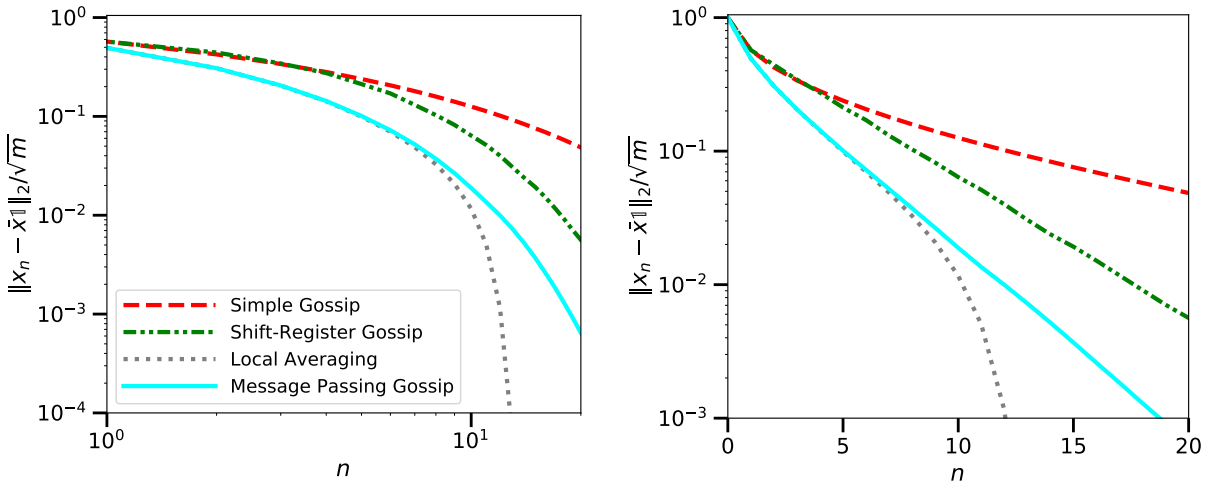
However, the message passing algorithm can be competitive in situations with a large spectral gap. For instance, McKay's Theorem [Sodin, 2007, Theorem 1.1] states that the spectral measure of a uniformly random d-regular graph on m vertices converges to the spectral measure $\sigma(\mathbb{T}_d)$ of the d-regular tree (in law, for the weak-convergence topology). This suggests that the message passing algorithm is well-suited for uniformly sampled large regular graphs. We give simulations in Figure 4.7 on uniformly sampled 3-regular graphs of size $m = 2000$. The results were averaged over 10 graphs. We observe that in this case, message passing matches closely the lower-bound. Note that in this case, we do not have a diffusive rate of convergence because the absolute spectral



(A) in logarithmic scale

(B) in semi-logarithmic scale

FIGURE 4.6. Performance of different gossip algorithms running on the 2D grid.



(A) in logarithmic scale

(B) in semi-logarithmic scale

FIGURE 4.7. Performance of different gossip algorithms running a uniformly random 3-regular graph of size $m = 2000$.

gap $\tilde{\mu}$ does not converge to 0 as $m \to \infty$ (see [Friedman, 2003] for a proof that $\mu \to 1 - 2\sqrt{d-1}/d$ in probability).

Appendix of Chapter 4

4.A. Details of the simulations of Section 4.1

2D grid. We run simulations on a 40×40 square lattice ($m = 1600$ vertices) endowed with the gossip matrix defined in (1.15) with $d_{\max} = 4$. The results are plotted in Figure 4.2A and a 20×20 grid is plotted in Figure 4.1A for visualization.

3D grid. We run simulations on a $12 \times 12 \times 12$ cubic lattice ($m = 1728$ vertices) endowed with the gossip matrix defined in (1.15) with $d_{\max} = 6$. The results are plotted in Figure 4.2B.

2D percolation bond. We build a 2D percolation bond by taking a 40×40 2D grid, and keep each edge independently with probability $p = 0.6$. To avoid connectivity issues, we consider G the largest connected component of the resulting graph, endowed with the gossip matrix defined in (1.15) with $d_{\max} = 4$. The results are plotted in Figure 4.2C and a 20×20 percolation bond is plotted in Figure 4.1B for visualization.

3D percolation bond. We build a 3D percolation bond by taking a $12 \times 12 \times 12$ 3D grid and keep each edge independently with probability $p = 0.4$. To avoid connectivity issues, we consider G the largest connected component of the resulting graph, endowed with the gossip matrix defined in (1.15) with $d_{\max} = 6$. The results are plotted in Figure 4.2D.

2D random geometric graph. We build a 2D random geometric graph G by sampling $m = 1600$ points uniformly in the unit square $[0, 1]^2$ and linking pairs closer than $1.5/\sqrt{m} = 0.0375$. To avoid connectivity issues, we consider G the largest connected component of the resulting graph. We build a gossip matrix W on G with the formulas: $W_{vw} = \max(\deg v, \deg w)^{-1}$ if $v \sim w$ and $W_{vv} = 1 - \sum_{w:w\sim v} \max(\deg v, \deg w)^{-1}$. The results are shown in Figure 4.2E.

3D random geometric graph. We build a 3D random geometric graph G by sampling $m = 1728$ points in the unit cube $[0, 1]^3$ and linking pairs closer than $1.5/m^{1/3} = 0.125$. To avoid connectivity issues, we consider G the largest connected component of the resulting graph. We build a gossip matrix W on G with the formulas: $W_{vw} = \max(\deg v, \deg w)^{-1}$ if $v \sim w$ and $W_{vv} = 1 - \sum_{w:w\sim v} \max(\deg v, \deg w)^{-1}$. The results are shown in Figure 4.2F.

4.B. Toolbox from orthogonal polynomials

In this appendix, we describe the tools from the theory of orthogonal polynomials that we use in this paper. The definition of the orthogonal polynomials π_n w.r.t. some measure τ is given in Definition 4.1. We start by giving some general properties of orthogonal polynomials in Section 4.B.1. We then describe two parameterized measures with respect to which orthogonal polynomials can be explicitly described: the Jacobi polynomials, in Section 4.B.2, and the polynomials orthogonal to some measure of the form $(1 - \lambda^2)^{1/2}/\rho(\lambda)$, where ρ is some polynomial, in Section 4.B.3. We finally give in Section 4.B.4 some asymptotic properties of the Jacobi polynomials as $n \to \infty$.

4.B.1. General properties.

Proposition 4.9 (from [Szegö, 1939, Theorem 3.3.1])**.** Let π_n be a family of orthogonal polynomials w.r.t. some measure τ on some interval $[a, b]$. Then the zeros of π_n are real, distinct and located in the interior of $[a, b]$.

In Proposition 4.3, it is stated that the orthogonal polynomials satisfy a three-term recurrence relation. We write here the short proof as it is used in Appendix 4.E.

Proof of Proposition 4.3. The polynomial $\lambda\pi_n(\lambda)$ of the variable λ is of degree $n+1$, thus it can be decomposed over the orthogonal basis $\pi_0(\lambda), \pi_1(\lambda), \dots, \pi_{n+1}(\lambda)$:

$$\lambda\pi_n(\lambda) = \sum_{k=0}^{n+1} \frac{\langle \lambda\pi_n, \pi_k\rangle_\tau}{\langle \pi_k, \pi_k\rangle_\tau}\pi_k(\lambda)\,.$$

Note that $\langle \lambda\pi_n, \pi_k\rangle_\tau = \int \lambda\pi_n(\lambda)\pi_k(\lambda)\mathrm{d}\tau(\lambda) = \langle \pi_n, \lambda\pi_k\rangle_\tau = 0$ when $k \leqslant n-2$ because in this case $\lambda\pi_k(\lambda) \in \mathbb{R}_{n-1}[X]$ and π_n is orthogonal to $\mathbb{R}_{t-1}[X]$. Thus

$$\lambda\pi_n(\lambda) = \frac{\langle \lambda\pi_n, \pi_{n+1}\rangle_\tau}{\langle \pi_{n+1}, \pi_{n+1}\rangle_\tau}\pi_{n+1}(\lambda) + \frac{\langle \lambda\pi_n, \pi_n\rangle_\tau}{\langle \pi_n, \pi_n\rangle_\tau}\pi_n(\lambda) + \frac{\langle \lambda\pi_n, \pi_{n-1}\rangle_\tau}{\langle \pi_{n-1}, \pi_{n-1}\rangle_\tau}\pi_{n-1}(\lambda)\,,$$

with the convention $\pi_{-1} = 0$. Note that $\langle \lambda\pi_n, \pi_{n+1}\rangle_\tau$ is non-zero as otherwise it would imply that $\lambda\pi_n$ is a polynomial of degree smaller or equal to n, which is absurd. We get the recursion formula by denoting

$$a_n = \frac{\langle \pi_{n+1}, \pi_{n+1}\rangle_\tau}{\langle \lambda\pi_n, \pi_{n+1}\rangle_\tau}\,, \quad b_n = -\frac{\langle \pi_{n+1}, \pi_{n+1}\rangle_\tau \langle \lambda\pi_n, \pi_n\rangle_\tau}{\langle \lambda\pi_n, \pi_{n+1}\rangle_\tau \langle \pi_n, \pi_n\rangle_\tau}\,, \quad c_n = \frac{\langle \pi_{n+1}, \pi_{n+1}\rangle_\tau \langle \lambda\pi_n, \pi_{n-1}\rangle_\tau}{\langle \lambda\pi_n, \pi_{n+1}\rangle_\tau \langle \pi_{n-1}, \pi_{n-1}\rangle_\tau}\,. \quad (4.18)$$

□

4.B.2. Jacobi polynomials.

Definition 4.5 (from [Szegö, 1939, Chapter IV])**.** Let $\alpha, \beta > -1$. The Jacobi polynomials $P_n^{(\alpha,\beta)}$ are the orthogonal polynomials w.r.t. the Jacobi measure

$$\sigma^{(\alpha,\beta)}(\mathrm{d}\lambda) = (1-\lambda)^\alpha(1+\lambda)^\beta \mathbb{1}_{\{\lambda\in(-1,1)\}}\mathrm{d}\lambda\,,$$

normalized such that $P_n^{(\alpha,\beta)}(1) = \binom{n+\alpha}{n}$.

Example 4.1 (from [Szegö, 1939, Section 2.4])**.** (1) The Chebyshev polynomials T_n of the first kind are the orthogonal polynomials w.r.t. $\sigma^{(-1/2,-1/2)}(\mathrm{d}\lambda) = (1-\lambda^2)^{-1/2}$ and normalized such that $T_n(1) = 1$. They are, up to some rescaling, a family of Jacobi polynomials. They satisfy the trigonometric formula

$$T_n(\cos\theta) = \cos(n\theta)\,.$$

(2) The Chebyshev polynomials U_n of the second kind are the orthogonal polynomials w.r.t. $\sigma^{(1/2,1/2)}(\mathrm{d}\lambda) = (1-\lambda^2)^{1/2}$ and normalized such that $U_n(1) = n+1$. They are, up to some rescaling, a family of Jacobi polynomials. They satisfy the trigonometric formula

$$U_n(\cos\theta) = \frac{\sin(n+1)\theta}{\sin\theta}\,.$$

A remarkable property of the Jacobi polynomials is that their recurrence relation can be computed explicitly.

Proposition 4.10 (from [Szegö, 1939, Section 4.4.5])**.** Let $\alpha, \beta > -1$. The Jacobi polynomials $P_n^{\alpha,\beta}$ satisfy the three recurrence formula

$$P_0^{(\alpha,\beta)}(\lambda) = 1\,, \quad P_1^{(\alpha,\beta)}(\lambda) = \frac{1}{2}(\alpha+\beta+2)\lambda + \frac{1}{2}(\alpha-\beta)\,,$$
$$\begin{aligned}2(n+1)(n+1+\alpha+\beta)(2n+\alpha+\beta)P_{n+1}^{(\alpha,\beta)}(\lambda)\\ = (2n+\alpha+\beta+1)[(2n+\alpha+\beta+2)(2n+\alpha+\beta)\lambda + \alpha^2 - \beta^2]P_n^{(\alpha,\beta)}(\lambda)\\ - 2(n+\alpha)(n+\beta)(2n+\alpha+\beta+2)P_{n-1}^{(\alpha,\beta)}(\lambda)\,.\end{aligned}$$

Example 4.2. The Chebyshev polynomials of the first and the second kind satisfy the same recurrence formula, but with different initializations:

$$\begin{aligned} &T_0(\lambda) = 1\,, && T_1(\lambda) = \lambda\,, && T_{n+1}(\lambda) = 2\lambda T_n(\lambda) - T_{n-1}(\lambda)\,,\\ &U_0(\lambda) = 1\,, && U_1(\lambda) = 2\lambda\,, && U_{n+1}(\lambda) = 2\lambda U_n(\lambda) - U_{n-1}(\lambda)\,.\end{aligned}$$

Proposition 4.11 (from [Szegö, 1939, Theorem 7.32.1])**.** Let $\alpha, \beta \geqslant 1/2$. Then

$$\max_{\lambda\in[-1,1]} \left|P_n^{(\alpha,\beta)}(\lambda)\right| = \binom{n+\max(\alpha,\beta)}{n}\,.$$

4.B.3. Polynomials orthogonal w.r.t. $(1-\lambda^2)^{1/2}/\rho(\lambda)$, ρ polynomial. In this section, we present how one can compute the recurrence relation for some orthogonal polynomials w.r.t. a weight of the form $(1-\lambda^2)^{1/2}/\rho(\lambda)$, ρ polynomial.

Proposition 4.12 (from [Szegö, 1939, Theorem 1.2.1])**.** Let ρ be a real polynomial of degree l which is non-negative for $\lambda \in [-1,1]$. Then there exists a polynomial h of degree l such that for all real θ, $\rho(\cos\theta) = |h(e^{i\theta})|^2$.

Proposition 4.13 (from [Szegö, 1939, Theorem 2.6])**.** Let ρ be a real polynomial of degree l taking positive values on the interval $[-1,1]$, and

$$\tau(\mathrm{d}\lambda) = \frac{(1-\lambda^2)^{1/2}}{\rho(\lambda)}\mathrm{d}\lambda\,.$$

Let h be a polynomial of degree l such that $\rho(\cos\theta) = |h(e^{i\theta})|^2$ (see Proposition 4.12), and decompose $h(e^{i\theta}) = c(\theta) + is(\theta)$, $c(\theta)$ and $s(\theta)$ real. Then the polynomials

$$\pi_n(\cos\theta) = c(\theta)U_n(\cos\theta) - \frac{s(\theta)}{\sin\theta}T_{n+1}(\cos\theta)$$

are orthogonal w.r.t. τ.

4.B.4. Asymptotics for the Jacobi polynomials. To prove the asymptotic performance guarantees of the polynomial iterations we build in this paper, we need the following asymptotic properties of the Jacobi polynomials.

Proposition 4.14 (from [Szegö, 1939, Theorem 8.21.7])**.** Let $\alpha, \beta > -1$, and $\lambda > 1$ a real number. Then there exists a positive constant $c = c(\alpha,\beta,\lambda)$ such that

$$P_n^{(\alpha,\beta)}(\lambda) \underset{n\to\infty}{\sim} \frac{c}{n^{1/2}}\left(\lambda + \sqrt{\lambda^2-1}\right)^n\,.$$

In the special case of the Chebyshev polynomials, we also have similar non-asymptotic bounds.

Lemma 4.1. For all $\lambda > 1$, for all $n \geqslant 0$,

$$\frac{1}{2}\left(\lambda + \sqrt{\lambda^2 - 1}\right)^n \leqslant T_n(\lambda) \leqslant \left(\lambda + \sqrt{\lambda^2 - 1}\right)^n , \tag{4.19}$$

$$\left(\lambda + \sqrt{\lambda^2 - 1}\right)^n \leqslant U_n(\lambda) \leqslant (n+1)\left(\lambda + \sqrt{\lambda^2 - 1}\right)^n . \tag{4.20}$$

Proof. We start by deriving a classic expression for the Chebyshev polynomials. The identities

$$T_n(\cos\theta) = \cos(n\theta)\,, \qquad\qquad U_n(\cos\theta) = \frac{\sin((n+1)\theta)}{\sin\theta}\,,$$

can be interpreted as

$$T_n\left(\frac{z+z^{-1}}{2}\right) = \frac{z^n + z^{-n}}{2}\,, \qquad U_n\left(\frac{z+z^{-1}}{2}\right) = \frac{z^{n+1} - z^{-(n+1)}}{z - z^{-1}}\,, \qquad \text{for } z = e^{i\theta}.$$

The above equations are equalities of holomorphic functions on the unit circle, it implies that the identities must be true for all complex numbers $z \neq 0$; we use it here for real numbers z.

For $\lambda > 1$, write $\lambda = (z + z^{-1})/2$, $z > 1$. This is equivalent to $z = \lambda + \sqrt{\lambda^2 - 1}$. Then

$$T_n(\lambda) = \frac{z^n + z^{-n}}{2} = \frac{1 + z^{-2n}}{2} z^n = \frac{1 + z^{-2n}}{2}\left(\lambda + \sqrt{\lambda^2 - 1}\right)^n .$$

As $z > 1$,

$$\frac{1}{2} \leqslant \frac{1 + z^{-2n}}{2} \leqslant 1\,.$$

This proves the inequalities (4.19). Further,

$$U_n(\lambda) = \frac{z^{n+1} - z^{-(n+1)}}{z - z^{-1}} = \frac{1 - z^{-2n-2}}{1 - z^{-2}} z^n = \frac{1 - z^{-2n-2}}{1 - z^{-2}}\left(\lambda + \sqrt{\lambda^2 - 1}\right)^n .$$

As $z > 1$,

$$1 \leqslant \frac{1 - z^{-2n-2}}{1 - z^{-2}} \leqslant n + 1\,.$$

This proves the inequalities (4.20). □

Proposition 4.15 (from [Szegö, 1939, Theorem 7.32.2])**.** Let $\alpha, \beta > -1$. There exists two constants $C_1, C_2 > 0$ such that,

$$\left|P_n^{(\alpha,\beta)}(\cos\theta)\right| \leqslant \begin{cases} C_1 \theta^{-\alpha-1/2} n^{-1/2} & \text{if } 1/n \leqslant \theta \leqslant \pi/2\,, \\ C_2 n^{\alpha} & \text{if } 0 \leqslant \theta \leqslant 1/n\,. \end{cases}$$

4.C. Some basic tools for the proofs

4.C.1. Comparing integrals using a domination of the cumulative distribution function.

Lemma 4.2. Let σ, τ be two positive measures on some interval $[a, b]$ such that for all $\lambda \in [a, b]$,

$$\sigma([\lambda, b]) \leqslant \tau([\lambda, b])\,. \tag{4.21}$$

Then for all continuous non-decreasing functions $f : [a, b] \to \mathbb{R}_{\geqslant 0}$,

$$\int_{[a,b]} f(\lambda) \mathrm{d}\sigma(\lambda) \leqslant \int_{[a,b]} f(\lambda) \mathrm{d}\tau(\lambda)\,.$$

PROOF. For any $u \in \mathbb{R}_{\geqslant 0}$, denote $\lambda^*(u) = \min\{\lambda | u \leqslant f(\lambda)\}$.

$$\int_{[a,b]} f(\lambda)\mathrm{d}\sigma(\lambda) = \int_{[a,b]} \int_{\mathbb{R}_{\geqslant 0}} \mathbb{1}_{\{u \leqslant f(\lambda)\}} f(\lambda)\mathrm{d}u\mathrm{d}\sigma(\lambda) = \int_{\mathbb{R}_{\geqslant 0}} \left(\int_{[a,b]} \mathbb{1}_{\{u \leqslant f(\lambda)\}} \mathrm{d}\sigma(\lambda) \right) \mathrm{d}u$$
$$= \int_{\mathbb{R}_{\geqslant 0}} \sigma([\lambda^*(u), b])\mathrm{d}u\,.$$

The proof is finished using (4.21) and similar equalities for τ. □

.

4.C.2. The gamma and beta function. The gamma function Γ and the beta function B are defined as [Olver et al., 2010, Section 5.2, Section 5.12]

$$\Gamma(z) = \int_0^\infty e^{-u} u^{z-1} \mathrm{d}u\,, \quad z \geqslant 0\,; \qquad B(a,b) = \int_0^1 s^{a-1}(1-s)^{b-1}\mathrm{d}s = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}\,, \quad a,b > 0\,.$$

The asymptotic ratios of the gamma functions are given in [Olver et al., 2010, Eq. 5.11.12]: for $c, d \in \mathbb{R}$,

$$\frac{\Gamma(z+c)}{\Gamma(z+d)} \sim z^{c-d} \quad \text{as } z \to +\infty\,.$$

This gives the asymptotic of the beta function

$$B(a,b) \sim \frac{\Gamma(b)}{a^b} \quad \text{as } a \to +\infty\,. \tag{4.22}$$

4.D. Proof of Proposition 4.2

Note first that as $\mathrm{d}\tau(\lambda) = (1-\lambda)\mathrm{d}\sigma(\lambda)$ is a measure on $[-1,1]$, if $\tilde{\pi}_0, \dots, \tilde{\pi}_{\bar{n}-1}$ is a sequence of orthogonal polynomials w.r.t. τ, then the zeros of the polynomials $\tilde{\pi}_0, \dots, \tilde{\pi}_{\bar{n}-1}$ are located in the interior of $[-1,1]$ (see Proposition 4.9). In particular, $\tilde{\pi}_n(1) \neq 0$, $n < \bar{n}$. Thus it is possible to build a family $\pi_0 = \tilde{\pi}_0/\tilde{\pi}_0(1), \dots, \pi_{\bar{n}-1} = \tilde{\pi}_{\bar{n}-1}/\tilde{\pi}_{\bar{n}-1}(1)$ of orthogonal polynomials normalized to take value 1 at 1, as it is done in Proposition 4.2.

The polynomial π_n satisfies $\pi_n(1) = 1$ and $\deg \pi_n = n$. We now consider some polynomial Q_n also satisfying $Q_n(1) = 1$ and $\deg Q_n = n$, and show that

$$\int \pi_n(\lambda)^2 \mathrm{d}\sigma(\lambda) \leqslant \int Q_n(\lambda)^2 \mathrm{d}\sigma(\lambda)\,, \qquad \text{i.e.} \qquad \langle \pi_n, \pi_n \rangle_\sigma \leqslant \langle Q_n, Q_n \rangle_\sigma\,.$$

The polynomial $Q_n - \pi_n$ vanishes at 1 thus there exists a polynomial R_{n-1} of degree at most $n-1$ such that $Q_n(\lambda) = \pi_n(\lambda) + (1-\lambda)R_{n-1}(\lambda)$. Then

$$\langle Q_n, Q_n \rangle_\sigma = \langle \pi_n, \pi_n \rangle_\sigma + 2\langle \pi_n, (1-\lambda)R_{n-1} \rangle_\sigma + \langle (1-\lambda)R_{n-1}, (1-\lambda)R_{n-1} \rangle_\sigma\,.$$

Note that $\langle \pi_n, (1-\lambda)R_{n-1} \rangle_\sigma = \langle \pi_n, R_{n-1} \rangle_\tau = 0$ because π_n is orthogonal to all polynomials of degree smaller or equal to $n-1$ w.r.t. $\langle ., . \rangle_\tau$. Moreover,

$$\langle (1-\lambda)R_{n-1}, (1-\lambda)R_{n-1} \rangle_\sigma = \int (1-\lambda)^2 R_{n-1}(\lambda)^2 \mathrm{d}\sigma(\lambda) \geqslant 0\,.$$

Thus $\langle Q_n, Q_n \rangle_\sigma \geqslant \langle \pi_n, \pi_n \rangle_\sigma$. This shows that π_n is a minimizer.

We now show that the minimizer π_n is unique. There is equality $\langle Q_n, Q_n \rangle_\sigma = \langle \pi_n, \pi_n \rangle_\sigma$ if and only if $\int (1-\lambda)^2 R_{n-1}(\lambda)^2 \mathrm{d}\sigma(\lambda) = 0$, i.e. $(1-\lambda)R_{n-1}$ vanishes on $\operatorname{Supp}\sigma$. But the cardinal of $\operatorname{Supp}\sigma$ is at least $\bar{n}$ while $(1-\lambda)R_{n-1}$ is a polynomial of degree at most $n \leqslant \bar{n} - 1$. Thus the equality case is reached if and only if $R_{n-1} = 0$, i.e. $Q_n = \pi_n$.

4.E. The parameter-free polynomial iteration

In this section, we give the details of the implementation of the parameter-free polynomial iteration in a centralized setting. We explicit the computation of the optimal coefficients a_n, b_n and c_n. It is used in the simulation of Figure 4.4.

The parameter free polynomial iteration is Eq. (4.7), where the coefficients a_n, b_n, c_n, $n \geqslant 1$, are determined in Eq. (4.18). The results are repeated here for convenience:

$$x_1 = a_0 W x_0 + b_0 x_0\,, \quad x_{n+1} = a_n W x_n + b_n x_n - c_n x_{n-1}\,,$$
$$a_n = \frac{\langle \pi_{n+1}, \pi_{n+1}\rangle_\tau}{\langle \lambda\pi_n, \pi_{n+1}\rangle_\tau}\,, \quad b_n = -\frac{\langle \pi_{n+1}, \pi_{n+1}\rangle_\tau \langle \lambda\pi_n, \pi_n\rangle_\tau}{\langle \lambda\pi_n, \pi_{n+1}\rangle_\tau \langle \pi_n, \pi_n\rangle_\tau}\,, \quad c_n = \frac{\langle \pi_{n+1}, \pi_{n+1}\rangle_\tau \langle \lambda\pi_n, \pi_{n-1}\rangle_\tau}{\langle \lambda\pi_n, \pi_{n+1}\rangle_\tau \langle \pi_{n-1}, \pi_{n-1}\rangle_\tau}\,.$$

where $\tau = (1-\lambda)\sigma$, σ is defined in (4.5). Note that the scalar products that appear in the formulas for a_n, b_n, c_n can be computed from the iterates $x_n = \pi_n(W)x_0$, $n \geqslant 0$. For instance,

$$\begin{aligned}\langle \lambda\pi_n, \pi_{n-1}\rangle_\tau &= \int \lambda\pi_n(\lambda)\pi_{n-1}(\lambda)(1-\lambda)\mathrm{d}\sigma(\lambda)\\ &= \sum_{i=1}^m \langle x_0, u_i\rangle^2 \lambda_i \pi_n(\lambda_i)\pi_{n-1}(\lambda_i)(1-\lambda_i)\\ &= \langle W\pi_n(W)x_0, (I-W)\pi_{n-1}(W)x_0\rangle\\ &= \langle Wx_n, x_{n-1} - Wx_{n-1}\rangle\,.\end{aligned}$$

Note that the last line requires the computation of a scalar product $\langle ., .\rangle$ over $\mathbb{R}^{\mathcal{V}}$, which means summing over $v \in \mathcal{V}$. This is possible in simulations where we can centralize the information of the nodes $v \in \mathcal{V}$. However in practical situation where the coordinates of x_n are distributed among the nodes, such a computation requires many additional communication steps. This makes the parameter free polynomial iteration impractical.

The computation of the other scalar products give

$$b_n = -a_n \frac{\langle x_n - Wx_n, Wx_n\rangle}{\langle x_n, x_n - Wx_n\rangle}\,, \qquad c_n = a_n \frac{\langle x_n, x_{n-1} - Wx_{n-1}\rangle}{\langle x_{n-1}, x_{n-1} - Wx_{n-1}\rangle}\,,$$

and as $a_n + b_n - c_n = 1$ (that follows from $\pi_n(1) = 1$ for all n), we get for $n \geqslant 1$,

$$\tilde{b}_n = -\frac{\langle x_n - Wx_n, Wx_n\rangle}{\langle x_n, x_n - Wx_n\rangle}\,, \qquad \tilde{c}_n = \frac{\langle x_n, x_{n-1} - Wx_{n-1}\rangle}{\langle x_{n-1}, x_{n-1} - Wx_{n-1}\rangle}\,,$$
$$x_{n+1} = \frac{1}{1+\tilde{b}_n - \tilde{c}_n}\left(Wx_n + \tilde{b}_n x_n - \tilde{c}_n x_{n-1}\right)\,.$$

Similarly, one can compute that

$$x_1 = \frac{1}{1+\tilde{b}^0}\left(Wx_0 + \tilde{b}_0 x_0\right)\,, \qquad \tilde{b}_0 = -\frac{\langle x_0 - Wx_0, Wx_0\rangle}{\langle x_0, x_0 - Wx_0\rangle}\,,$$

which gives the initialization of the parameter-free polynomial iteration.

4.F. Proofs of Propositions 4.4 and 4.5

4.F.1. Proof of Proposition 4.4. The return probability p_n of the lazy random walk on $\mathbb{Z}^d$ is equivalent to $C/n^{d/2}$ for some constant C. It is, for instance, a consequence of the local central limit theorem for random walks on $\mathbb{Z}^d$ [Lawler and Limic, 2010, Theorem 2.1.1]. Thus the spectral dimension of $\mathbb{Z}^d$ is d.

4.F.2. Proof of Proposition 4.5. The return probabilities of the random walk on the supercritical percolation cluster have rather been studied in continuous time. The continuous-time random walk is defined as follows: the random walk at w waits at an exponential time of parameter 1 before picking a site w' out of the $2d$ neighboring sites uniformly randomly. If there is an edge in the percolation configuration between w and w', the random walk jumps to w', otherwise it stays in w and starts again. Denote X_t the continuous-time random walk, and $\mathbb{P}_w$ the probability w.r.t. this random walk when it is started from some vertex w.

Lemma 4.3. There exists two constants $c = c(d,p), C = c(d,p) > 0$ such that, a.s. on the set $\{v \in G\}$, there exists a random time t_0 such that for $t \geqslant t_0$,

$$\frac{c}{t^{d/2}} \leqslant \mathbb{P}_v(X_t = v) \leqslant \frac{C}{t^{d/2}}\,.$$

Proof. The upper bound is proved by Mathieu and Remy [2004, Theorem 1.2]. As noted by Biskup et al. [2011, Lemma 5.1], the lower bound can be proved using a central limit theorem on X_t; we repeat the argument here as our random walk differs slightly from theirs. As X_t is reversible w.r.t. the uniform measure on G,

$$\mathbb{P}_v(X_{2t} = v) = \sum_{w\in G} \mathbb{P}_v(X_t = w)\mathbb{P}_w(X_t = v) = \sum_{w\in G} \mathbb{P}_v(X_t = w)^2\,.$$

By the Cauchy-Schwarz inequality,

$$\begin{aligned}
\mathbb{P}_v(\|X_t - v\|_2 \leqslant \sqrt{t})^2 &= \left(\sum_{x\in G} \mathbb{1}_{\{\|x-v\|_2\leqslant\sqrt{t}\}}\mathbb{P}_v(X_t = x)\right)^2 \\
&\leqslant \left|\{\|x \in G : x - v\|_2 \leqslant \sqrt{t}\}\right| \left(\sum_{w\in G} \mathbb{P}_v(X_t = w)^2\right) \\
&\leqslant C_1 t^{d/2}\mathbb{P}_v(X_{2t} = v)\,,
\end{aligned}$$

for some constant C_1. Now using [Andres et al., 2013, Theorem 1.1(a)], there exists a deterministic variance σ^2 such that the law of $(X_t - v)/\sqrt{t}$ converges a.s. on the event $\{v \in G\}$ to a centered Gaussian with variance σ^2. Thus there exists a deterministic constant $c_1 > 0$ and a random time t_1 such that for $t \geqslant t_1$, $\mathbb{P}_t(\|X_t - v\|_2 \leqslant \sqrt{t})^2 \geqslant c_1$. This finishes the proof of the lower bound. □

We now finish the proof of the proposition using Lemma 4.3. If μ_t denotes the law of X_t,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[\mu_t\right] = (W - I)\mu_t\,, \qquad \text{thus} \quad \mu_t = e^{t(W-I)}\mu_0\,.$$

Thus

$$\mathbb{P}_v(X_t = v) = \langle \delta_v, \mu_t\rangle = \langle \delta_v, e^{t(W-I)}\delta_v\rangle \overset{\text{(Definition 4.3)}}{=} \int e^{t(\lambda-1)}\mathrm{d}\sigma(\lambda)\,.$$

As a consequence, Lemma 4.3 translates into bounds on the Laplace transform of σ: a.s. on $\{v \in G\}$, for t large enough,

$$\frac{c}{t^{d/2}} \leqslant \int e^{t(\lambda-1)}\mathrm{d}\sigma(\lambda) \leqslant \frac{C}{t^{d/2}}\,.$$

Some bounds on the spectral density of σ near 1 easily follow (see [Müller and Stollmann, 2007, Lemma 4.5]): there exists constants $c', C' > 0$ such that a.s. on $\{v \in G\}$, for E small enough,

$$c'E^{d/2} \leqslant \sigma([1 - E, 1]) \leqslant C'E^{d/2}\,.$$

The proof is finished using Proposition 4.6.

4.G. Proof of Proposition 4.6

We start by assuming that $l = \lim_{E\to 0} \log \sigma([1-E,1])/\log E$ exists and is finite. We show that d_s exists and that $l = d_s/2$. To this end, we define

$$\underline{d}_s = -2 \limsup_{n\to\infty} \frac{\log p_n}{\log n}\,, \qquad\qquad \bar{d}_s = -2 \liminf_{n\to\infty} \frac{\log p_n}{\log n}\,,$$

where p_n is defined as in Definition 4.2. Note that

$$p_n = \left\langle e_v, \left(\frac{I+W}{2}\right)^n e_v \right\rangle \overset{\text{(Definition 4.3)}}{=} \int \left(\frac{1+\lambda}{2}\right)^n \mathrm{d}\sigma(\lambda)\,. \tag{4.23}$$

Proof that $\bar{d}_s/2 \leqslant l$. Consider $l_+ > l$. Then there exists constants $c_1, c_2 > 0$ such that for all $E \in [0,2]$,

$$\sigma([1-E,1]) \geqslant c_1 E^{l_+} = c_2 \sigma^{(l_+-1,0)}([1-E,1])\,,$$

where $\sigma^{(l_+-1,0)}(\mathrm{d}\lambda) = (1-\lambda)^{l_+-1}\mathrm{d}\lambda$. Then

$$\begin{aligned} p_n &\overset{(4.23)}{=} \int_{[-1,1]} \left(\frac{1+\lambda}{2}\right)^n \mathrm{d}\sigma(\lambda) \overset{\text{(Lemma 4.2)}}{\geqslant} c_2 \int_{-1}^1 \left(\frac{1+\lambda}{2}\right)^{n-1} (1-\lambda)^{l_+-1}\mathrm{d}\lambda \\ &\overset{(u=(1+\lambda)/2)}{=} c_3 \int_0^1 u^n (1-u)^{l_+-1}\mathrm{d}u = c_3 B\left(n+1, l_+\right) \underset{n\to\infty}{\overset{(4.22)}{\sim}} \frac{c_4}{n^{l_+}}\,, \end{aligned}$$

for some constant $c_3, c_4 > 0$. Thus

$$\liminf_{n\to\infty} \frac{\log p_n}{\log n} \geqslant -l_+\,, \qquad\qquad \text{i.e.} \quad \frac{\bar{d}_s}{2} \leqslant l_+\,.$$

This being true for all $l_+ > l$, this proves $\bar{d}_s/2 \leqslant l$.

Proof that $\underline{d}_s/2 \geqslant l$. Consider $l_- < l$. Then there exists constants C_1, C_2 such that for all $E \in [0,2]$,

$$\sigma([1-E,1]) \leqslant C_1 E^{l_-} = C_2 \sigma^{(l_--1,0)}([1-E,1])\,,$$

where $\sigma^{(l_--1,0)}(\mathrm{d}\lambda) = (1-\lambda)^{l_--1}\mathrm{d}\lambda$. Then

$$\begin{aligned} p_n &\overset{(4.23)}{=} \int_{[-1,1]} \left(\frac{1+\lambda}{2}\right)^n \mathrm{d}\sigma(\lambda) \overset{\text{(Lemma 4.2)}}{\leqslant} C_2 \int_{-1}^1 \left(\frac{1+\lambda}{2}\right)^n (1-\lambda)^{l_--1}\mathrm{d}\lambda \\ &\overset{(u=(1+\lambda)/2)}{=} C_3 \int_0^1 u^n (1-u)^{l_--1}\mathrm{d}u = C_3 B\left(n+1, l_-\right) \underset{n\to\infty}{\sim} \frac{C_4}{n^{l_-}}\,, \end{aligned}$$

for some constants C_3, C_4. Thus

$$\limsup_{n\to\infty} \frac{\log p_n}{\log n} \leqslant -l_-\,, \qquad\qquad \text{i.e.} \quad \frac{\underline{d}_s}{2} \geqslant l_-\,.$$

This being true for all $l_- < l$, this proves $\underline{d}_s/2 \geqslant l$.

Finally, we have proven $l \leqslant \underline{d}_s/2 \leqslant \bar{d}_s/2 \leqslant l$. Thus the limit $d_s = -2\lim_{n\to\infty} \log p_n/\log n$ exists and is equal to $2l$.

Conversely, we assume now that d_s exists and is finite. We show that $l = \lim_{E\to 0} \log \sigma([1-E,1])/\log E$ exists and that $l = d_s/2$. To this end, we define

$$\underline{l} = \liminf_{E\to 0} \frac{\log \sigma([1-E,1])}{\log E}\,, \qquad\qquad \bar{l} = \limsup_{E\to 0} \frac{\log \sigma([1-E,1])}{\log E}\,.$$

Proof that $\underline{l} \geqslant d_s/2$. For any $n \in \mathbb{N}$, we have $\mathbb{1}_{\{\lambda \geqslant 1-E\}} \leqslant (1-E/2)^{-n}\,((1+\lambda)/2)^n$, thus by integrating against $\mathrm{d}\sigma(\lambda)$,

$$\sigma([1-E,1]) \leqslant \left(1-\frac{E}{2}\right)^{-n} \int \left(\frac{1+\lambda}{2}\right)^n \sigma(\mathrm{d}\lambda)\,,$$

$$\frac{\log \sigma([1-E,1])}{\log E} \geqslant \frac{\log \int \left(\frac{1+\lambda}{2}\right)^n \sigma(\mathrm{d}\lambda)}{\log n}\frac{\log n}{\log E} - \frac{n \log\left(1-\frac{E}{2}\right)}{\log E}\,.$$

We choose $n(E) = \lfloor E^{-1} \rfloor$. Then we get

$$\underline{l} = \liminf_{E\to 0} \frac{\log \sigma([1-E,1])}{\log E} \geqslant -\frac{d_s}{2}(-1) - 0 = \frac{d_s}{2}$$

Proof that $\bar{l} \leqslant d_s/2$. For any $n \in \mathbb{N}$, we have $((1+\lambda)/2)^n - (1-E/2)^n \leqslant \mathbb{1}_{\{\lambda \geqslant 1-E\}}$, thus by integrating against $\mathrm{d}\sigma(\lambda)$,

$$\int \left(\frac{1+\lambda}{2}\right)^n \mathrm{d}\sigma(\lambda) - \left(1-\frac{E}{2}\right)^n \leqslant \sigma([1-E,1])\,.$$

Let $d > d_s$. There exists a constant $c > 0$ such that $\int((1+\lambda)/2)^n \mathrm{d}\sigma(\lambda) \geqslant c/n^{d/2}$. Then

$$\log\left(\frac{c}{n^{d/2}} - \left(1-\frac{E}{2}\right)^n\right) \leqslant \log \sigma([1-E,1])\,.$$

Let $\alpha > 1$. We choose $n(E) = \lceil E^{-\alpha} \rceil$. Then

$$\left(1-\frac{E}{2}\right)^{n(E)} = \exp\left(n(E)\log\left(1-\frac{E}{2}\right)\right) \leqslant \exp\left(-\frac{n(E)E}{2}\right) \leqslant \exp\left(-\frac{1}{2}E^{1-\alpha}\right)$$

decreases super-polynomially fast as $E \to 0$. Moreover

$$\frac{c}{n(E)^{d/2}} \underset{E\to 0}{\sim} cE^{\alpha d/2}\,.$$

Finally,

$$\bar{l} = \limsup_{E\to 0} \frac{\log \sigma([1-E,1])}{\log E} \leqslant \frac{\alpha d}{2}\,.$$

As this is true for all $\alpha > 1, d > d_s$, we have $\bar{l} \leqslant d_s/2$.

Finally, we have proven that $d_s/2 \leqslant \underline{l} \leqslant \bar{l} \leqslant d_s/2$. Then the limit $l = \lim_{E\to 0} \log \sigma([1-E,1])/\log E$ exists and $l = d_s/2$.

4.H. Computation of the recursion coefficients of some orthogonal polynomials

4.H.1. A rescaling lemma for orthogonal polynomials. We start with a lemma giving the change in the recursion coefficients of orthogonal polynomials when the underlying measure undergoes an affine transformation. It is used in the next subsections.

Lemma 4.4. Let σ be a measure on $\mathbb{R}$, $\pi_0, \ldots, \pi_{\bar{n}-1}$ a sequence of orthogonal polynomials w.r.t. σ and

$$\pi_{n+1}(\lambda) = (a_n\lambda + b_n)\pi_n(\lambda) - c_n\pi_{n-1}(\lambda)\,, \qquad n \geqslant 1\,, \tag{4.24}$$

their recurrence formula (see Definition 4.1 and Theorem 4.3).

Let $\varphi : \lambda \mapsto \alpha\lambda + \beta$, $\alpha \neq 0$ be a linear function and $\tilde{\sigma}$ be the image measure of σ by φ (which means that for all measurable set A, $\tilde{\sigma}(A) = \sigma(\varphi^{-1}(A))$). Then a sequence of orthogonal polynomials w.r.t. $\tilde{\sigma}$ is given by the formula

$$\tilde{\pi}_n(\tilde{\lambda}) := \pi_n\left(\varphi^{-1}(\tilde{\lambda})\right) = \pi_n\left(\frac{\tilde{\lambda}-\beta}{\alpha}\right).$$

These polynomials follow the recursion formula

$$\tilde{\pi}_{n+1}(\tilde{\lambda}) = (\tilde{a}_n\tilde{\lambda} + \tilde{b}_n)\tilde{\pi}_n(\tilde{\lambda}) - \tilde{c}_n\tilde{\pi}_{n-1}(\tilde{\lambda}),$$
$$\tilde{a}_n = \frac{a_n}{\alpha}, \qquad \tilde{b}_n = b_n - \frac{a_n\beta}{\alpha}, \qquad \tilde{c}_n = c_n.$$

Proof. By change of variable,

$$\begin{aligned}\int \tilde{\pi}_n(\tilde{\lambda})\tilde{\pi}_k(\tilde{\lambda})\mathrm{d}\tilde{\sigma}(\tilde{\lambda}) &= \int \pi_n\left(\varphi^{-1}(\tilde{\lambda})\right)\pi_k\left(\varphi^{-1}(\tilde{\lambda})\right)\mathrm{d}\tilde{\sigma}(\tilde{\lambda}) \\ &= \int \pi_n\left(\varphi^{-1}(\varphi(\lambda))\right)\pi_k\left(\varphi^{-1}(\varphi(\lambda))\right)\mathrm{d}\sigma(\lambda) \\ &= \int \pi_n(\lambda)\pi_k(\lambda)\mathrm{d}\sigma(\lambda) = \mathbb{1}_{\{n=k\}},\end{aligned}$$

and $\deg\tilde{\pi}_n = n$ thus $\tilde{\pi}_0, \ldots, \tilde{\pi}_{\bar{n}-1}$ are orthogonal polynomials w.r.t. $\tilde{\sigma}$. The recurrence relation for $\tilde{\pi}_n$ follows by evaluating the recurrence relation (4.24) for π_n in $(\tilde{\lambda}-\beta)/\alpha$. □

4.H.2. Jacobi polynomials. Let $\alpha, \beta > -1$. In this section, we derive, using the recurrence formula for the Jacobi polynomial $P_n^{(\alpha,\beta)}$ of Proposition 4.10, a similar recurrence relation for the polynomials $\pi_n^{(\alpha,\beta)}$ orthogonal w.r.t. the Jacobi measure $\sigma^{(\alpha,\beta)}$, but normalized such that $\pi_n^{(\alpha,\beta)}(1) = 1$.

Substituting $P_n^{(\alpha,\beta)} = \binom{n+\alpha}{t}\pi_n^{(\alpha,\beta)}$ in the recurrence relation of Proposition 4.10, we get

$$\begin{aligned}&2(n+1)(n+1+\alpha+\beta)(2n+\alpha+\beta)\binom{n+1+\alpha}{n+1}\pi_{n+1}^{(\alpha,\beta)}(\lambda) \\ &\quad= (2n+\alpha+\beta+1)[(2n+\alpha+\beta+2)(2n+\alpha+\beta)\lambda + \alpha^2 - \beta^2]\binom{n+\alpha}{n}\pi_n^{(\alpha,\beta)}(\lambda) \\ &\quad\quad- 2(n+\alpha)(n+\beta)(2n+\alpha+\beta+2)\binom{n-1+\alpha}{n-1}\pi_{n-1}^{(\alpha,\beta)}(\lambda).\end{aligned}$$

Using that $(n+1)\binom{n+1+\alpha}{n+1} = (n+1+\alpha)\binom{n+\alpha}{n}$ and $n\binom{n+\alpha}{n} = (n+\alpha)\binom{n-1+\alpha}{n-1}$, we can divide the above equation by $\binom{n+\alpha}{n}$. We get

$$\begin{aligned}&2(n+1+\alpha+\beta)(2n+\alpha+\beta)(n+1+\alpha)\pi_{n+1}^{(\alpha,\beta)}(\lambda) \\ &\quad= (2n+\alpha+\beta+1)[(2n+\alpha+\beta+2)(2n+\alpha+\beta)\lambda + (\alpha+\beta)(\alpha-\beta)]\pi_n^{(\alpha,\beta)}(\lambda) \\ &\quad\quad- 2n(n+\beta)(2n+\alpha+\beta+2)\pi_{n-1}^{(\alpha,\beta)}(\lambda).\end{aligned}$$

Summing up, we obtain the recursion formula

$$\pi_0(\lambda) = 1, \qquad \pi_1(\lambda) = a_0^{(\alpha,\beta)}\lambda + b_0^{(\alpha,\beta)},$$
$$\pi_{n+1}^{(\alpha,\beta)}(\lambda) = \left(a_n^{(\alpha,\beta)}\lambda + b_n^{(\alpha,\beta)}\right)\pi_n^{(\alpha,\beta)}(\lambda) - c_n^{(\alpha,\beta)}\pi_{n-1}^{(\alpha,\beta)}(\lambda),$$

with the recursion coefficients

$$\begin{aligned}
a_0^{(\alpha,\beta)} &= \frac{\alpha+\beta+2}{2(1+\alpha)}\,, \qquad b_0^{(\alpha,\beta)} = \frac{\alpha-\beta}{2(1+\alpha)}\,,\\
a_n^{(\alpha,\beta)} &= \frac{(2n+\alpha+\beta+1)(2n+\alpha+\beta+2)}{2(n+1+\alpha+\beta)(n+1+\alpha)}\,,\\
b_n^{(\alpha,\beta)} &= \frac{(2n+\alpha+\beta+1)(\alpha+\beta)(\alpha-\beta)}{2(n+1+\alpha+\beta)(n+1+\alpha)(2n+\alpha+\beta)}\,,\\
c_n^{(\alpha,\beta)} &= \frac{n(n+\beta)(2n+\alpha+\beta+2)}{(n+1+\alpha+\beta)(2n+\alpha+\beta)(n+1+\alpha)}\,.
\end{aligned} \tag{4.25}$$

4.H.3. Rescaled Jacobi polynomials. Let $\alpha, \beta > -1$. In this section, we determine a recursion formula for the orthogonal polynomials $\pi_n^{(\alpha,\beta,\mu)}$ w.r.t. the rescaled Jacobi measure

$$\mathrm{d}\sigma^{(\alpha,\beta,\mu)}(\lambda) = ((1-\mu)-\lambda)^{\alpha}(1+\lambda)^{\beta}\mathbb{1}_{\{\lambda\in(-1,1-\mu)\}}\mathrm{d}\lambda\,,$$

The polynomials $\pi_n^{(\alpha,\beta,\mu)}$ are normalized such that $\pi_n^{(\alpha,\beta,\mu)}(1) = 1$.

Note that, up to a rescaling, $\mathrm{d}\sigma^{(\alpha,\beta,\mu)}$ is the image measure of the Jacobi measure $\mathrm{d}\sigma^{(\alpha,\beta)}$ (defined in (4.5)) by the linear function $\varphi_\mu(\lambda) = (1-\mu/2)\lambda - \mu/2$. Thus Lemma 4.4 gives a family of orthogonal polynomials $P_n^{(\alpha,\beta,\mu)}$ w.r.t. $\mathrm{d}\sigma^{(\alpha,\beta,\mu)}$ and their recursion formula:

$$\begin{aligned}
&P_n^{(\alpha,\beta,\mu)}(\tilde\lambda) = \pi_n^{(\alpha,\beta)}\left(\varphi_\mu^{-1}(\tilde\lambda)\right)\,,\\
&P_{n+1}^{(\alpha,\beta,\mu)}(\tilde\lambda) = \left(a_n^{(\alpha,\beta,\mu)}\tilde\lambda + b_n^{(\alpha,\beta,\mu)}\right)P_n^{(\alpha,\beta,\mu)}(\tilde\lambda) - c_n^{(\alpha,\beta,\mu)}P_{n-1}^{(\alpha,\beta,\mu)}(\tilde\lambda)\,,\\
&a_n^{(\alpha,\beta,\mu)} = a_n^{(\alpha,\beta)}\left(1-\frac{\mu}{2}\right)^{-1}\,, \qquad b_n^{(\alpha,\beta,\mu)} = b_n^{(\alpha,\beta)} + \frac{\mu}{2}a_n^{(\alpha,\beta)}\left(1-\frac{\mu}{2}\right)^{-1}\,, \qquad c_n^{(\alpha,\beta,\mu)} = c_n^{(\alpha,\beta)}\,.
\end{aligned}$$

However, the polynomials $P_n^{(\alpha,\beta,\mu)}$ are not normalized such that $P_n^{(\alpha,\beta,\mu)} = 1$. Indeed, $P_n^{(\alpha,\beta,\mu)} = \pi_n^{(\alpha,\beta)}\left((1-\mu/2)^{-1}(1+\mu/2)\right)$. It is difficult to deduce the recurrence relation for $\pi_n^{(\alpha,\beta,\mu)} = P_n^{(\alpha,\beta,\mu)}/P_n^{(\alpha,\beta,\mu)}(1)$ from the recurrence relation for $P_n^{(\alpha,\beta,\mu)}$. One can circumvent this difficulty by using that the normalization $P_n^{(\alpha,\beta,\mu)}(1)$ also follows the recurrence relation

$$P_{n+1}^{(\alpha,\beta,\mu)}(1) = \left(a_n^{(\alpha,\beta,\mu)} + b_n^{(\alpha,\beta,\mu)}\right)P_n^{(\alpha,\beta,\mu)}(1) - c_n^{(\alpha,\beta,\mu)}P_{n-1}^{(\alpha,\beta,\mu)}(1)\,.$$

Summing things up, we get

$$\begin{aligned}
&\pi_n^{(\alpha,\beta,\mu)}(\lambda) = \frac{P_n^{(\alpha,\beta,\mu)}(\lambda)}{P_n^{(\alpha,\beta,\mu)}(1)}\,,\\
&P_0^{(\alpha,\beta,\mu)}(\lambda) = 1\,, \qquad P_0^{(\alpha,\beta,\mu)}(1) = 1\,,\\
&P_1^{(\alpha,\beta,\mu)}(\lambda) = a_0^{(\alpha,\beta,\mu)}\lambda + b_0^{(\alpha,\beta,\mu)}\,, \qquad P_1^{(\alpha,\beta,\mu)}(1) = a_0^{(\alpha,\beta,\mu)} + b_0^{(\alpha,\beta,\mu)}\,,\\
&P_{n+1}^{(\alpha,\beta,\mu)}(\lambda) = \left(a_n^{(\alpha,\beta,\mu)}\tilde\lambda + b_n^{(\alpha,\beta,\mu)}\right)P_n^{(\alpha,\beta,\mu)}(\lambda) - c_n^{(\alpha,\beta,\mu)}P_{n-1}^{(\alpha,\beta,\mu)}(\lambda)\,, \qquad n\geqslant 1\,,\\
&P_{n+1}^{(\alpha,\beta,\mu)}(1) = \left(a_n^{(\alpha,\beta,\mu)} + b_n^{(\alpha,\beta,\mu)}\right)P_n^{(\alpha,\beta,\mu)}(1) - c_n^{(\alpha,\beta,\mu)}P_{n-1}^{(\alpha,\beta,\mu)}(1)\,, \qquad n\geqslant 1\,,\\
&a_n^{(\alpha,\beta,\mu)} = a_n^{(\alpha,\beta)}\left(1-\frac{\mu}{2}\right)^{-1}\,, \qquad b_n^{(\alpha,\beta,\mu)} = b_n^{(\alpha,\beta)} + \frac{\mu}{2}a_n^{(\alpha,\beta)}\left(1-\frac{\mu}{2}\right)^{-1}\,, \qquad n\geqslant 0\,,\\
&c_n^{(\alpha,\beta,\mu)} = c_n^{(\alpha,\beta)}\,, \qquad n\geqslant 1\,.
\end{aligned} \tag{4.26}$$

These equations give a practical way to compute the polynomials $\pi_n^{(\alpha,\beta,\mu)}$ because all recursion coefficients can be computed explicitly using the formulas (4.25).

4.H.4. Polynomials orthogonal to the modified Kesten-McKay measure. In this section, we determine the recurrence formula for the orthogonal polynomials π_n w.r.t. the modified Kesten-McKay measure

$$(1-\lambda)\sigma(\mathbb{T}_d)(\mathrm{d}\lambda) = \frac{d}{2\pi(1+\lambda)}\left(\frac{4(d-1)}{d^2} - \lambda^2\right)^{1/2} \mathbb{1}_{[-2\sqrt{d-1}/d, 2\sqrt{d-1}/d]}(\lambda)\mathrm{d}\lambda\,.$$

The polynomials π_n are normalized such that $\pi_n(1) = 1$.

The measure $\mathrm{d}\sigma(\mathbb{T}_d)$ is, up to a rescaling factor, the image measure of

$$\mathrm{d}\tilde{\sigma}(\lambda) = \frac{(1-\lambda^2)^{1/2}}{d + 2\sqrt{d-1}\lambda}\mathbb{1}_{[-1,1]}(\lambda)\mathrm{d}\lambda \tag{4.27}$$

by the linear map $\varphi : \lambda \mapsto 2\sqrt{d-1}\lambda/d$. We thus compute a family of orthogonal polynomials w.r.t. $\tilde{\sigma}$ and then use Lemma 4.4.

The orthogonal polynomials w.r.t. $\tilde{\sigma}$ are given by Proposition 4.13. Following the cited theorem, we define $\rho(\lambda) = d + 2\sqrt{d-1}\lambda$ and

$$\begin{aligned}\rho(\cos\theta) &= 2\sqrt{d-1}\cos\theta + d = |h(e^{i\theta})|^2\,,\\ h(e^{i\theta}) &= \sqrt{d-1} + e^{i\theta} = \underbrace{\sqrt{d-1} + \cos\theta}_{:=c(\theta)} + i\underbrace{\sin\theta}_{:=s(\theta)}\,.\end{aligned}$$

Then we have the following family $\tilde{p}_t$ of orthogonal polynomials w.r.t. $\tilde{\sigma}$.

$$\begin{aligned}\tilde{p}_n(\cos\theta) &= c(\theta)U_n(\cos\theta) - \frac{s(\theta)}{\sin\theta}T_{n+1}(\cos\theta)\,,\\ \tilde{p}_n(\lambda) &= (\sqrt{d-1} + \lambda)U_n(\lambda) - T_{n+1}(\lambda)\,,\end{aligned}$$

where T_n and U_n denote the n-th Chebyshev polynomial of the first kind and the second kind respectively. As the Chebyshev polynomials T_n and U_n both satisfy the same recurrence relation

$$\begin{aligned}T_{n+1}(\lambda) &= 2\lambda T_n(\lambda) - T_{n-1}(\lambda)\,,\\ U_{n+1}(\lambda) &= 2\lambda U_n(\lambda) - U_{n-1}(\lambda)\,, \quad n \geqslant 1\,,\end{aligned}$$

the same relation follows for $\tilde{p}_n$:

$$\tilde{p}_{n+1}(\lambda) = 2\lambda\tilde{p}_n(\lambda) - \tilde{p}_{n-1}(\lambda)\,, \quad n \geqslant 1\,,$$

with initial condition $\tilde{p}_0(\lambda) = \sqrt{d-1}$ and $\tilde{p}_1(\lambda) = 2\sqrt{d-1}\lambda + 1$.

Lemma 4.4 gives the rescaled orthogonal polynomials $p_n(\lambda) = \tilde{p}_n\left(\varphi^{-1}(\lambda)\right)$ w.r.t. $\mathrm{d}\sigma(\mathbb{T}_d)$:

$$p_0(\lambda) = \sqrt{d-1}\,, \quad p_1(\lambda) = d\lambda + 1\,, \quad p_{n+1}(\lambda) = \frac{d}{\sqrt{d-1}}\lambda p_n(\lambda) - p_{n-1}(\lambda)\,, \quad n \geqslant 1\,. \tag{4.28}$$

As $\pi_n(\lambda) = p_n(\lambda)/p_n(1)$, it now remains to compute $p_n(1)$. The sequence $p_n(1)$, $n \geqslant 1$ satisfies a second-order recurrence relation with fixed coefficients, it thus can be solved explicitly

$$p_n(1) = \frac{1}{d-1}\left(d(d-1)^{(n+1)/2} - 2(d-1)^{(1-n)/2}\right)\,.$$

By substituting $p_n(\lambda) = p_n(1)\pi_n(\lambda)$ in (4.28), one obtains

$$\pi_0(\lambda) = 1\,, \qquad \pi_1(\lambda) = a_0\lambda + b_0\,, \qquad \pi_{n+1}(\lambda) = a_n\lambda\pi_n(\lambda) - c_n\pi_{n-1}(\lambda)\,, \quad n \geqslant 1\,,$$

$$a_0 = \frac{d}{d+1}\,, \qquad b_0 = \frac{1}{d+1}\,, \qquad a_n = \frac{\frac{d}{d-1} - 2(d-1)^{-(n+1)}}{1 - \frac{2}{d}(d-1)^{-(n+1)}}\,, \qquad c_n = \frac{\frac{1}{d-1} - \frac{2}{d}(d-1)^{-n}}{1 - \frac{2}{d}(d-1)^{-(n+1)}}\,, \quad n \geqslant 1\,.$$

4.I. Proof of Theorem 4.1

The proof is divided in four subsections. Appendix 4.I.1 develops tools that we use both in the proof of the theorem. We then prove the theorem in Sections 4.I.2, 4.I.3and 4.I.4. Finally, in Appendix 4.I.5, we discuss the choice of the parameters of the Jacobi polynomials in the Jacobi polynomial iteration. In all this appendix, we denote $\sigma = \sigma(G, W, v)$ the spectral measure of G.

4.I.1. Preliminaries. The first lemma relates the MSE of the estimator $x_n(v)$ to the spectral measure.

Lemma 4.5. Write $x_n = P_n(W)x_0$ using the polynomial gossip point of view. Then

$$\mathbb{E}[(x_n(v) - \mu)^2] = (\operatorname{var} \nu) \, \|P_n(W)e_v\|^2_{\ell^2(\mathcal{V})} = (\operatorname{var} \nu) \int P_n(\lambda)^2 \mathrm{d}\sigma(\lambda) \, .$$

Proof. As $P_n(1) = 1$, we have

$$\mathbb{E}[x_n] = \mathbb{E}[P_n(W)x_0] = P_n(W)\mathbb{E}[x_0] = P_n(W)\mu\mathbb{1} = P_n(1)\mu\mathbb{1} = \mu\mathbb{1} \, .$$

In words, the estimator $x_n(v)$ is unbiased. Thus

$$\begin{aligned} \mathbb{E}[(x_n(v) - \mu)^2] &= \operatorname{var} x_n(v) = \operatorname{var} \, \langle P_n(W)x_0, e_v\rangle_{\ell^2(\mathcal{V})} = \operatorname{var} \, \langle x_0, P_n(W)e_v\rangle_{\ell^2(\mathcal{V})} \\ &= (\operatorname{var} \nu) \, \|P_n(W)e_v\|^2_{\ell^2(\mathcal{V})} \, , \end{aligned}$$

using that W is symmetric and that the $x_0(w)$, $w \in \mathcal{V}$ are i.i.d. random variables. Then

$$\mathbb{E}[(x_n(v) - \mu)^2] = (\operatorname{var} \nu) \, \langle P_n(W)e_v, P_n(W)e_v\rangle_{\ell^2(\mathcal{V})} = (\operatorname{var} \nu) \left\langle e_v, P_n(W)^2 e_v \right\rangle_{\ell^2(\mathcal{V})} \, .$$

The proof is finished using the Definition 4.3 of the spectral measure. □

In the statement of Theorem 4.1, we have stated results in terms of the spectral dimension $d_s = 2 \lim_{E\to 0} \log \sigma([1-E, 1])/\log E$. In the proof here, we will be more precise. We show how the results of Theorem 4.1 actually depend on different definitions of the dimension.

Definition 4.6. Let τ be a probability measure on $[-1, 1]$. We define

(1) the right upper dimension $\overline{\dim}_{\rightarrow}\tau \in [0, \infty]$ of the measure τ as

$$\overline{\dim}_{\rightarrow}\tau = 2 \limsup_{E\to 0} \frac{\log \tau([1-E, 1])}{\log E} \, ,$$

(2) the right lower dimension $\underline{\dim}_{\rightarrow}\tau \in [0, \infty]$ of the measure τ as

$$\underline{\dim}_{\rightarrow}\tau = 2 \liminf_{E\to 0} \frac{\log \tau([1-E, 1])}{\log E} \, ,$$

(3) the left upper dimension $\overline{\dim}_{\leftarrow}\tau \in [0, \infty]$ of the measure τ as

$$\overline{\dim}_{\leftarrow}\tau = 2 \limsup_{E\to 0} \frac{\log \tau([-1, -1+E])}{\log E} \, .$$

(4) the left lower dimension $\underline{\dim}_{\leftarrow}\tau \in [0, \infty]$ of the measure τ as

$$\underline{\dim}_{\leftarrow}\tau = 2 \liminf_{E\to 0} \frac{\log \tau([-1, -1+E])}{\log E} \, .$$

4.I.2. Proof of Theorem 4.1: simple gossip. In the case of simple gossip, $P_n(\lambda) = \lambda^n$.

Proposition 4.16. Let τ be a probability measure on $[-1, 1]$. Then

$$\liminf_{n\to\infty} \frac{\int \lambda^{2n} \mathrm{d}\tau(\lambda)}{\log n} \geqslant -\frac{\min\left(\overline{\dim}_{\rightarrow}\tau, \overline{\dim}_{\leftarrow}\tau\right)}{2}$$

Proof. Let $d > \overline{\dim}_{\rightarrow}\tau$. As $\overline{\dim}_{\rightarrow}\tau = 2\limsup_{E\to 0} \log \sigma([1-E,1])/\log E$, there exists constants $c_1, c_2 > 0$ such that for all $E \in [0, 2]$,

$$\tau([1-E,1]) \geqslant c_1 E^{d/2} = c_2 \sigma^{(d/2-1,0)}([1-E,1]) \tag{4.29}$$

where $\sigma^{(d/2-1,0)}(\mathrm{d}\lambda) = (1-\lambda)^{d/2-1}\mathrm{d}\lambda$. Then using jointly Lemma 4.2 and Eq. (4.29),

$$\int \lambda^{2n} \mathrm{d}\tau(\lambda) \geqslant \int_{[0,1]} \lambda^{2n} \mathrm{d}\tau(\lambda) \geqslant c_1 \int_0^1 \lambda^{2n}(1-\lambda)^{d/2-1}\mathrm{d}\lambda = B(2n+1, d/2) \overset{(4.22)}{\underset{n\to\infty}{\sim}} \frac{c_3}{n^{d/2}},$$

for some constant c_3. Thus

$$\liminf_{n\to\infty} \frac{\int \lambda^{2n} \mathrm{d}\tau(\lambda)}{\log n} \geqslant -\frac{d}{2}.$$

This being true for all $d > \overline{\dim}_{\rightarrow}\tau$, this proves

$$\liminf_{n\to\infty} \frac{\int \lambda^{2n} \mathrm{d}\tau(\lambda)}{\log n} \geqslant -\frac{\overline{\dim}_{\rightarrow}\tau}{2}.$$

The proof at the other edge of the spectrum is the same by symmetry. □

The proof of Theorem 4.1 for simple gossip follows easily. Indeed, if $\tau = \sigma$ is the spectral measure of the graph, then $\overline{\dim}_{\rightarrow}\sigma = d_s$. Thus

$$\liminf_{n\to\infty} \frac{\log \mathbb{E}[(x_n(v)-\mu)^2]}{\log n} \overset{\text{(Lemma 4.5)}}{=} \liminf_{n\to\infty} \frac{\log \int \lambda^{2n} \mathrm{d}\sigma(\lambda)}{\log n} \overset{\text{(Proposition 4.16)}}{\geqslant} -\frac{d_s}{2}.$$

4.I.3. Proof of Theorem 4.1: shift-register. In the case of the shift-register gossip iteration, $P_t(\lambda)$ satisfies the second-order recurrence relation

$$P_0(\lambda) = 1, \qquad P_1(\lambda) = \lambda, \qquad P_{n+1}(\lambda) = \omega\lambda P_n(\lambda) + (1-\omega)P_{n-1}(\lambda). \tag{4.30}$$

The case $\omega = 1$ corresponds to simple gossip: it has been treated above. We now assume $\omega \in (1, 2]$.

Proposition 4.17. Let P_n be the polynomials defined in Eq. (4.30) with $\omega \in (1, 2]$. Then

$$P_n(\lambda) = (\omega-1)^{n/2}\left[\left(2 - \frac{2}{\omega}\right) T_n\left(\frac{\omega}{2\sqrt{\omega-1}}\lambda\right) + \left(\frac{2}{\omega} - 1\right) U_n\left(\frac{\omega}{2\sqrt{\omega-1}}\lambda\right)\right]$$

where T_n and U_n are the Chebyshev polynomials of the first and second kind respectively (see Example 4.1).

Proof. Consider the rescaled version Q_n of P_n given by the formula

$$P_n(\lambda) = (\omega-1)^{n/2} Q_n\left(\frac{\omega}{2\sqrt{\omega-1}}\lambda\right). \tag{4.31}$$

If follows from Eq. (4.30) that

$$Q_0(\lambda) = 1, \qquad Q_1(\lambda) = \frac{2}{\omega}\lambda, \qquad Q_{n+1}(\lambda) = 2\lambda Q_n(\lambda) - Q_{n-1}(\lambda).$$

Thus the sequence $Q_n, n \geqslant 0$ satisfies the same recurrence relation as the Chebyshev polynomials, but with a different initialization. As a consequence, it must be a linear combination of the two sequences of Chebyshev polynomials: there exists $\mu, \nu \in \mathbb{R}$ such that for all n,

$$Q_n(\lambda) = \mu T_n(\lambda) + \nu U_n(\lambda)$$

The computation of the weights μ, ν is straightforward from the initialization Q_0, Q_1. This proves the proposition. □

Proposition 4.18. Let $\omega \in (1, 2]$. The polynomials $P_n, n \geqslant 0$ defined in Eq. (4.30) are the orthogonal polynomials w.r.t. the measure

$$\tau(\mathrm{d}\lambda) = \frac{\left(\left(2\sqrt{\omega - 1}/\omega\right)^2 - \lambda^2\right)^{1/2}}{1 - \lambda^2}\,.$$

Proof. The orthogonal polynomials w.r.t. the measure

$$\tilde{\tau}(\mathrm{d}\lambda) = \frac{\left(1 - \lambda^2\right)^{1/2}}{\left(\omega/(2\sqrt{\omega - 1})\right)^2 - \lambda^2}$$

are computed using Proposition 4.13 with

$$\rho(\cos\theta) = \frac{\omega^2}{4(\omega - 1)} - \cos^2\theta\,.$$

Simple computations give that $\rho(\cos\theta) = |h(e^{i\theta})|^2$ with

$$h(e^{i\theta}) = \left|\frac{\omega}{2\sqrt{\omega - 1}}\left[\left(2 - \frac{2}{\omega}\right)\frac{1 - e^{2i\theta}}{2} + \left(\frac{2}{\omega} - 1\right)\right]\right|^2\,.$$

Proposition 4.13 then gives that the polynomials $(2 - 2/\omega)T_n + (2/\omega - 1)U_n$ are orthogonal w.r.t. $\tilde{\tau}$. But these polynomials are the polynomials Q_n defined in Eq. (4.31). We then use Lemma 4.4 to prove that P_n is orthogonal w.r.t. τ. □

Lemma 4.6. Let P_n be the polynomials defined in Eq. (4.30) with $\omega \in (1, 2]$ and τ a measure on $[-1, 1]$. Then

$$\liminf_{n\to\infty} \frac{\int P_n(\lambda)^2 \mathrm{d}\tau(\lambda)}{\log n} \geqslant -\frac{\min\left(\overline{\dim}_{\rightarrow}\tau, \overline{\dim}_{\leftarrow}\tau\right)}{2}$$

Proof.

$$\begin{aligned}
\int P_n(\lambda)^2 \mathrm{d}\tau(\lambda) &\geqslant \int_{[2\sqrt{\omega-1}/\omega, 1]} P_n(\lambda)^2 \mathrm{d}\tau(\lambda) \\
&\overset{\text{(Proposition 4.17)}}{\geqslant} c_1(\omega - 1)^n \int_{[2\sqrt{\omega-1}/\omega, 1]} T_n\left(\frac{\omega}{2\sqrt{\omega - 1}}\lambda\right)^2 \mathrm{d}\tau(\lambda) \\
&\overset{(4.19)}{\geqslant} c_2(\omega - 1)^n \int_{[2\sqrt{\omega-1}/\omega, 1]} \left(\frac{\omega}{2\sqrt{\omega - 1}}\lambda + \sqrt{\frac{\omega^2}{4(\omega - 1)}\lambda^2 - 1}\right)^{2n} \mathrm{d}\tau(\lambda)\,.
\end{aligned}$$

where $c_i > 0$ is a constant independent of n. Let $d > \overline{\dim}_{\rightarrow}\tau$. As $\overline{\dim}_{\rightarrow}\tau = 2\limsup_{E\to 0} \log\sigma([1 - E, 1])/\log E$, there exists constants $c_3, c_4 > 0$ such that for all $E \in [0, 2]$,

$$\tau([1 - E, 1]) \geqslant c_3 E^{d/2} = c_4 \sigma^{(d/2-1,0)}([1 - E, 1]) \tag{4.32}$$

where $\sigma^{(d/2-1,0)}(\mathrm{d}\lambda) = (1-\lambda)^{d/2-1}\mathrm{d}\lambda$. Then using jointly Lemma 4.2 and Eq. (4.32),

$$\int P_n(\lambda)^2 \mathrm{d}\tau(\lambda) \geqslant c_5(\omega-1)^n \int_{2\sqrt{\omega-1}/\omega}^1 \left(\frac{\omega}{2\sqrt{\omega-1}}\lambda + \sqrt{\frac{\omega^2}{4(\omega-1)}\lambda^2 - 1}\right)^{2n} (1-\lambda)^{d/2-1}\mathrm{d}\lambda$$
$$\geqslant c_6(\omega-1)^n \int_0^{\cosh^{-1}(\omega/(2\sqrt{\omega-1}))} e^{2nu}\left(1 - \frac{2\sqrt{\omega-1}}{\omega}\cosh u\right)^{d/2-1} \sinh u \,\mathrm{d}u\,.$$

where in the last step we made the change of variable $\omega/(2\sqrt{\omega-1})\lambda + \sqrt{\omega^2/(4(\omega-1))\lambda^2-1} = e^u$, i.e. $\lambda = 2\sqrt{\omega-1}/\omega \cosh u$. Denote $u_{\max} = \cosh^{-1}(\omega/(2\sqrt{\omega-1}))$ to shorten notations. As cosh is a convex function, for $u \in [0, u_{\max}]$,

$$\cosh u_{\max} - \cosh u \leqslant \frac{\cosh u_{\max} - 1}{u_{\max}}(u_{\max} - u)\,,$$
$$\Leftrightarrow 1 - \frac{2\sqrt{\omega-1}}{\omega}\cosh u \leqslant \left(1 - \frac{2\sqrt{\omega-1}}{\omega}\right)\left(1 - \frac{u}{u_{\max}}\right)$$

Moreover, choose some constant $u_{\min} \in (0, u_{\max})$ so that we can lower bound with a constant c_7: for all $u \in [u_{\min}, u_{\max}]$, $\sinh u \geqslant c_7$. This finally gives:

$$\int P_n(\lambda)^2 \mathrm{d}\tau(\lambda) \geqslant c_8(\omega-1)^n \int_{u_{\min}}^{u_{\max}} e^{2nu}\left(1 - \frac{u}{u_{\max}}\right)^{d/2-1} \mathrm{d}u\,.$$

After the change of variable $w = 2n(u_{\max} - u)$, this gives

$$\int P_n(\lambda)^2 \mathrm{d}\tau(\lambda) \geqslant c_8(\omega-1)^n \int_0^{2n(u_{\max}-u_{\min})} e^{2nu_{\max}} e^{-w}\left(\frac{w}{2nu_{\max}}\right)^{d/2-1} \frac{1}{2n}\,\mathrm{d}w\,.$$

Note that $e^{2nu_{\max}} = (\omega-1)^{-n}$, thus there exists a constant $c_9 > 0$ such that

$$\int P_n(\lambda)^2 \mathrm{d}\tau(\lambda) \geqslant c_9 \frac{1}{n^{d/2}} \int_0^{2n(u_{\max}-u_{\min})} e^{-w} w^{d/2-1}\,\mathrm{d}w\,.$$

This proves that

$$\liminf_{n\to\infty} \frac{\int P_n(\lambda)^2 \mathrm{d}\tau(\lambda)}{\log n} \geqslant -\frac{d}{2}\,.$$

This being true for all $d > \overline{\dim}_{\rightarrow}\tau$, this proves

$$\liminf_{n\to\infty} \frac{\int P_n(\lambda)^2 \mathrm{d}\tau(\lambda)}{\log n} \geqslant -\frac{\overline{\dim}_{\rightarrow}\tau}{2}\,.$$

The proof at the other edge of the spectrum is the same by symmetry. □

4.I.4. Proof of Theorem 4.1: Jacobi polynomial iteration. In this section, we use again the notation of Appendix 4.H.2: in the case of the Jacobi polynomial iteration (4.10), we have $x_n = \pi_n^{(d_s/2,0)}(W)x_0$, where $\pi_n^{(\alpha,\beta)}$ is the rescaled Jacobi polynomial; $\pi_n^{(\alpha,\beta)} = P_n^{(\alpha,\beta)}/\binom{n+\alpha}{n}$ where $P_n^{(\alpha,\beta)}$ is the traditional Jacobi polynomial. Lemma 4.5 suggests to study the quantity $\int \pi_n^{(d_s/2,0)}(\lambda)^2 \mathrm{d}\sigma(\lambda)$. However we study here the behavior of $\int \pi_n^{(\alpha,\beta)}(\lambda)^2 \mathrm{d}\sigma(\lambda)$ for any (α,β). This will be useful in Appendix 4.I.5 to give a motivation for the choice $\alpha = d/2, \beta = 0$ complementary to the intuition developed in Section 4.3, and will allow us to discuss the performance of other choices.

Proposition 4.19. Let τ be a probability measure on $[-1,1]$ and $\alpha, \beta \geqslant -1/2$. Then

$$\limsup_{n\to\infty} \frac{\log \int \pi_n^{(\alpha,\beta)}(\lambda)^2 \mathrm{d}\tau(\lambda)}{\log n} \leqslant -\min\left(2\alpha+1, \underline{\dim}_{\rightarrow}\tau, 2(\alpha-\beta) + \underline{\dim}_{\leftarrow}\tau\right)\,.$$

Before proving this proposition, we use it to finish the proof of the theorem. If $\tau = \sigma$ is the spectral measure of the graph, then $\underline{\dim}_{\rightarrow}\sigma = d_s$. Thus taking $\alpha = d_s/2$, $\beta = 0$ in Proposition 4.19, we get

$$\limsup_{n\to\infty} \frac{\log \int \pi_n^{(d_s/2,0)}(\lambda)^2 \mathrm{d}\sigma(\lambda)}{\log n} \leqslant -\min(d_s + 1, d_s, d_s + \underline{\dim}_{\leftarrow}\sigma) = -d_s\,, \tag{4.33}$$

as $\underline{\dim}_{\leftarrow}\sigma \geqslant 0$. One can conclude the proof using Lemma 4.5.

We now turn to the the proof of Proposition 4.19.

Lemma 4.7. Let τ be a probability measure on $[-1, 1]$ and $\alpha \geqslant -1/2$, $\beta > -1$. Then

$$\limsup_{n\to\infty} \frac{\log \int_{[0,1]} P_n^{(\alpha,\beta)}(\lambda)^2 \mathrm{d}\tau(\lambda)}{\log n} \leqslant -\min(1, \underline{\dim}_{\rightarrow}\tau - 2\alpha)\,.$$

Proof. Let $d < \underline{\dim}_{\rightarrow}\tau$. As $\underline{\dim}_{\rightarrow}\tau = 2\liminf_{E\to 0} \log \tau([1-E, 1])/\log E$, there exists constants C_1, C_2 such that for all $E \in [0, 2]$,

$$\tau([1-E, 1]) \leqslant C_1 E^{d/2} = C_2 \sigma^{(d/2-1,0)}([1-E, 1]) \tag{4.34}$$

where $\sigma^{(d/2-1,0)}(\mathrm{d}\lambda) = (1-\lambda)^{d/2-1}\mathrm{d}\lambda$.

For the proof of this result, we use the asymptotic bound on the Jacobi polynomials given by Proposition 4.15, thus we divide the integral

$$\int_{[0,1]} P_n^{(\alpha,\beta)}(\lambda)^2 \mathrm{d}\tau(\lambda) = \int_{[\cos 1/n, 1]} P_n^{(\alpha,\beta)}(\lambda)^2 \mathrm{d}\tau(\lambda) + \int_{[0,\cos 1/n[} P_n^{(\alpha,\beta)}(\lambda)^2 \mathrm{d}\tau(\lambda)\,,$$

and treat the two terms separately.

(a)

$$\begin{aligned}\int_{[\cos 1/n, 1]} P_n^{(\alpha,\beta)}(\lambda)^2 \mathrm{d}\tau(\lambda) &\leqslant C_3 n^{2\alpha} \tau\left(\left[\cos\frac{1}{n}, 1\right]\right) \leqslant C_1 C_3 n^{2\alpha}\left(1 - \cos\frac{1}{n}\right)^{d/2} \\ &\leqslant C_4 n^{2\alpha - d}\end{aligned}$$

for some constants C_3, C_4. Thus

$$\limsup_{n\to\infty} \frac{\log \int_{[\cos 1/n, 1]} P_n^{(\alpha,\beta)}(\lambda)^2 \mathrm{d}\tau(\lambda)}{\log n} \leqslant 2\alpha - d\,. \tag{4.35}$$

(b)

$$\int_{[0,\cos 1/n[} P_n^{(\alpha,\beta)}(\lambda)^2 \mathrm{d}\tau(\lambda) \leqslant C_5 n^{-1} \int_{[0,\cos(1/n)[} (\arccos\lambda)^{-2\alpha-1} \mathrm{d}\tau(\lambda) \tag{4.36}$$

We then use jointly Eq. (4.34) and Lemma 4.2 with the function

$$f(\lambda) = (\arccos\lambda)^{-2\alpha-1}\mathbb{1}_{\{\lambda < \cos 1/n\}} + n^{2\alpha+1}\mathbb{1}_{\{\lambda \geqslant \cos 1/n\}}\,.$$

Note that f is non-decreasing as $\alpha \geqslant 1/2$. We get

$$\begin{aligned}&\int_{[0,\cos(1/n)[} (\arccos\lambda)^{-2\alpha-1} \mathrm{d}\tau(\lambda) \\ &\qquad \leqslant C_2 \int_{[0,\cos(1/n)[} (\arccos\lambda)^{-2\alpha-1}(1-\lambda)^{d/2-1}\mathrm{d}\lambda + C_1 n^{2\alpha+1}\left(1 - \cos\frac{1}{n}\right)^{d/2}\end{aligned}$$

Now using the simple inequality $\arccos\lambda \geqslant \sqrt{2}\sqrt{1-\lambda}$, we get

$$\begin{aligned}\int_{[0,\cos(1/n)[} &(\arccos\lambda)^{-2\alpha-1}\mathrm{d}\tau(\lambda)\\ &\leqslant C_5\int_{[0,\cos(1/n)[}(1-\lambda)^{-\alpha+d/2-3/2}\mathrm{d}\lambda + C_6 n^{2\alpha+1-d}\end{aligned} \tag{4.37}$$

for some constants C_5, C_6. Now if β is a real number,

$$\lim_{n\to\infty}\frac{\log\int_0^{\cos 1/n}(1-\lambda)^\beta\mathrm{d}\lambda}{\log n} = \max(0,-2\beta-2)\,. \tag{4.38}$$

Indeed, if $\beta\neq -1$,

$$\begin{aligned}\int_0^{\cos 1/n}(1-\lambda)^\beta\mathrm{d}\lambda &= \left[-\frac{(1-\lambda)^{\beta+1}}{\beta+1}\right]_0^{\cos 1/n} = \frac{1}{\beta+1}\left[1-\left(1-\cos\frac{1}{n}\right)^{\beta+1}\right]\\ &\underset{n\to\infty}{=} \frac{1}{\beta+1}\left[1-n^{-2\beta-2}+o(n^{-2\beta-2})\right] \sim C(\beta)n^{\max(0,-2\beta-2)}\,.\end{aligned}$$

for some constant $C(\beta)$ depending on β. This proves the statement (4.38) for $\beta\neq -1$. The result for $\beta=-1$ follows easily by noting that both terms in (4.38) are decreasing in β.

Merging finally Eqs. (4.36), (4.37) and (4.38), we get

$$\begin{aligned}\limsup_{n\to\infty}\frac{\log\int_{[0,\cos 1/n[}P_n^{(\alpha,\beta)}(\lambda)^2\mathrm{d}\tau(\lambda)}{\log n} &\leqslant -1+\max(0,2\alpha+1-d)\\ &= \max(-1,2\alpha-d) = -\min(1,d-2\alpha)\,.\end{aligned} \tag{4.39}$$

Finally

$$\begin{aligned}\limsup_{n\to\infty}&\frac{\log\int_{[0,1]}P_n^{(\alpha,\beta)}(\lambda)^2\mathrm{d}\tau(\lambda)}{\log n}\\ &\leqslant \limsup_{n\to\infty}\frac{2\max\left(\int_{[\cos 1/n,1]}P_n^{(\alpha,\beta)}(\lambda)^2\mathrm{d}\tau(\lambda),\int_{[0,\cos 1/n[}P_n^{(\alpha,\beta)}(\lambda)^2\mathrm{d}\tau(\lambda)\right)}{\log n}\\ &\leqslant \max\left(\limsup_{n\to\infty}\frac{\log\int_{[\cos 1/n,1]}P_n^{(\alpha,\beta)}(\lambda)^2\mathrm{d}\tau(\lambda)}{\log t},\limsup_{n\to\infty}\frac{\log\int_{[0,\cos 1/n[}P_n^{(\alpha,\beta)}(\lambda)^2\mathrm{d}\tau(\lambda)}{\log n}\right)\\ &\overset{(4.35),(4.39)}{\leqslant} \max(2\alpha-d,-\min(1,d-2\alpha)) = -\min(1,d-2\alpha)\,.\end{aligned}$$

As this is true for all $d<\underline{\dim}_{\rightarrow}\tau$, the lemma is proved. □

Proof of Proposition 4.19. If we denote $\tilde{\tau}$ the symmetric measure of τ w.r.t. 0 (i.e. the image measure of τ by the map $\lambda\mapsto-\lambda$), we have

$$\int_{[-1,0]}P_n^{(\alpha,\beta)}(\lambda)^2\mathrm{d}\tau(\lambda) = \int_{[0,1]}P_n^{(\alpha,\beta)}(-\lambda)^2\mathrm{d}\tilde{\tau}(\lambda) = \int_{[0,1]}P_n^{(\beta,\alpha)}(\lambda)^2\mathrm{d}\tilde{\tau}(\lambda)$$

Thus according to Lemma 4.7,

$$\limsup_{t\to\infty}\frac{\log\int_{[-1,0]}P_n^{(\alpha,\beta)}(\lambda)^2\mathrm{d}\tau(\lambda)}{\log n} \leqslant -\min(1,\underline{\dim}_{\rightarrow}\tilde{\tau}-2\beta) = -\min(1,\underline{\dim}_{\leftarrow}\tau-2\beta)\,. \tag{4.40}$$

Finally, using that $\pi_n^{(\alpha,\beta)} = P_n^{(\alpha,\beta)}/\binom{n+\alpha}{n}$,

$$\begin{aligned}
\limsup_{n\to\infty} \frac{\log \int_{[-1,1]} \pi_n^{(\alpha,\beta)}(\lambda)^2 \mathrm{d}\tau(\lambda)}{\log n} &\leqslant \limsup_{n\to\infty} \frac{\log \int_{[-1,1]} P_n^{(\alpha,\beta)}(\lambda)^2 \mathrm{d}\tau(\lambda)}{\log n} - 2\limsup_{n\to\infty} \frac{\log \binom{n+\alpha}{n}}{\log n} \\
&\leqslant \limsup_{n\to\infty} \frac{\log\left(2\max\left(\int_{[-1,0]} P_n^{(\alpha,\beta)}(\lambda)^2 \mathrm{d}\tau(\lambda), \int_{[0,1]} P_n^{(\alpha,\beta)}(\lambda)^2 \mathrm{d}\tau(\lambda)\right)\right)}{\log n} - 2\alpha \\
&\leqslant \max\left(\limsup_{n\to\infty} \frac{\log \int_{[-1,0]} P_n^{(\alpha,\beta)}(\lambda)^2 \mathrm{d}\tau(\lambda)}{\log n}, \limsup_{n\to\infty} \frac{\log \int_{[0,1]} P_n^{(\alpha,\beta)}(\lambda)^2 \mathrm{d}\tau(\lambda)}{\log n}\right) - 2\alpha \\
&\overset{((4.40),\text{Lemma 4.7})}{\leqslant} \max\left(-\min(1, \underline{\dim}_{\leftarrow}\tau - 2\beta), -\min(1, \underline{\dim}_{\rightarrow}\tau - 2\alpha)\right) - 2\alpha \\
&\leqslant -\min(1, \underline{\dim}_{\leftarrow}\tau - 2\beta, \underline{\dim}_{\rightarrow}\tau - 2\alpha) - 2\alpha \\
&= -\min(2\alpha+1, 2(\alpha-\beta) + \underline{\dim}_{\leftarrow}\tau, \underline{\dim}_{\rightarrow}\tau)\,.
\end{aligned}$$

□

4.I.5. Tuning of the parameters α and β. In this section, we discuss the performance of the polynomial gossip iteration $x_n = \pi_n^{(\alpha,\beta)}(W)x_0$ using the tools developed in the proof above. The Jacobi polynomial iteration introduced in Section 4.3.2 corresponds to the specific choice $\alpha = d_s/2, \beta = 0$, where d_s is the spectral measure of the graph. Thanks to the tools developed in the proof above, we can explore analytically the effect of changing α and β. In Figures 5.3 and Section 5.1.4, we also explore the effect of changing α by studying the shape of the diffusion on grids.

Inspired by (4.33), we define optimality as follows.

Definition 4.7. Let $\alpha, \beta > -1, d_{\leftarrow}, d_{\rightarrow} \geqslant 0$. We say that (α, β) is optimal for $(d_{\leftarrow}, d_{\rightarrow})$ if for any spectral measure σ such that $\underline{\dim}_{\rightarrow}\sigma = d_{\rightarrow}$ and $\underline{\dim}_{\leftarrow}\sigma = d_{\leftarrow}$,

$$\limsup_{n\to\infty} \frac{\log \int \pi_n^{(\alpha,\beta)}(\lambda)^2 \mathrm{d}\sigma(\lambda)}{\log n} \leqslant -d_{\rightarrow}\,.$$

The following theorem is an analogue of the optimality theorem 4.1(2) in the general case.

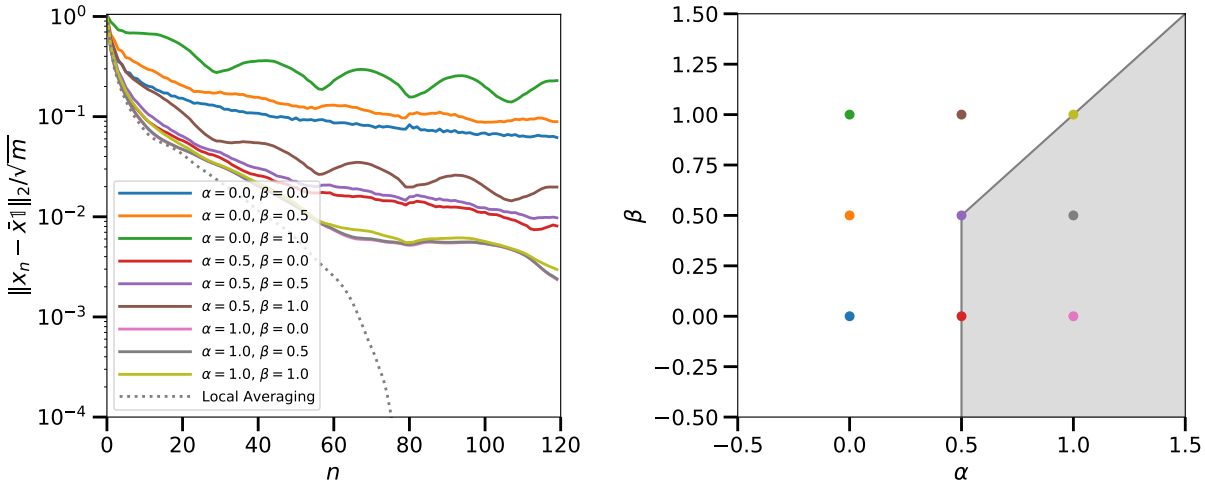
Theorem 4.5. Consider a graph G, a gossip matrix W and a vertex v. Denote $\sigma = \sigma(G, W, v)$ the spectral measure of the graph. Let $x_0(v), v \in \mathcal{V}$ be i.i.d. samples from a distribution of mean μ.

Let $\alpha, \beta > -1$ and define the polynomial iteration $x_n = \pi_n^{(\alpha,\beta)}(W)x_0$. If (α, β) is optimal for $(\underline{\dim}_{\leftarrow}\sigma, \underline{\dim}_{\rightarrow}\sigma)$, then

$$\limsup_{n\to\infty} \frac{\log \mathbb{E}\left[(x_n(v) - \mu)^2\right]}{\log n} \leqslant -\underline{\dim}_{\rightarrow}\sigma\,. \tag{4.41}$$

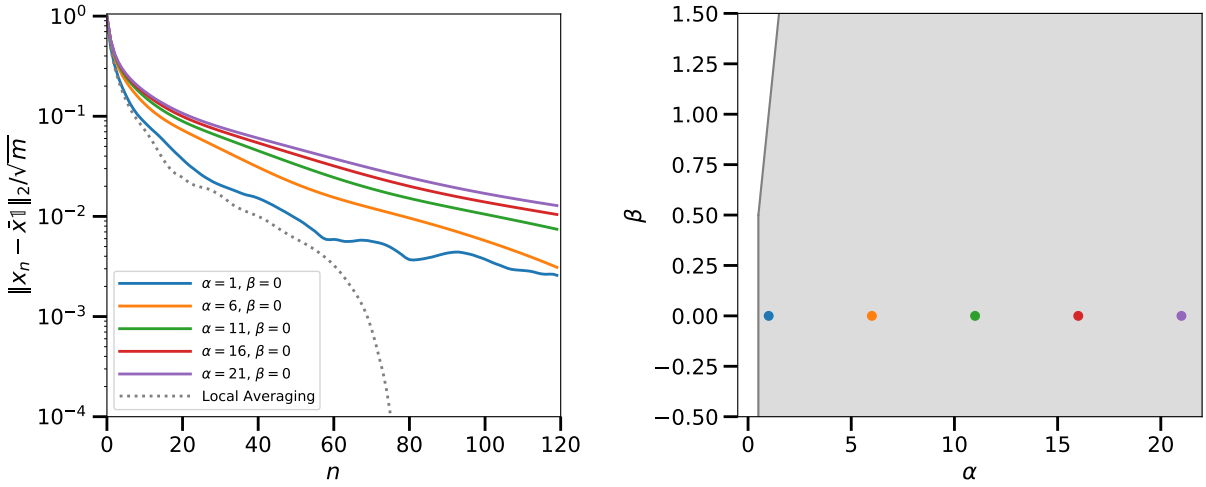
In the section above, we prove that $(d_{\rightarrow}/2, 0)$ is optimal for $(d_{\leftarrow}, d_{\rightarrow})$ (for any $d_{\leftarrow}, d_{\rightarrow} \geqslant -1/2$). We now explore other choices. According to Proposition 4.19, to prove that (α, β) is optimal for $(d_{\leftarrow}, d_{\rightarrow})$, it is sufficient to prove that

$$\min(2\alpha+1, d_{\rightarrow}, 2(\alpha-\beta) + d_{\leftarrow}) = d_{\rightarrow} \qquad \Leftrightarrow \qquad \begin{cases} 2\alpha + 1 \geqslant d_{\rightarrow} \\ 2(\alpha-\beta) + d_{\leftarrow} \geqslant d_{\rightarrow} \end{cases}$$



(A) Performance curves (B) Parameter space

FIGURE 4.8. Simulations of polynomial iterations using Jacobi polynomials with different parameters (α, β): frontier tightness.



(A) Performance curves (B) Parameter space

FIGURE 4.9. Simulations of polynomial iterations using Jacobi polynomials with different parameters (α, β): large α asymptotic.

$$\Leftrightarrow \quad \begin{cases} \alpha \geqslant \frac{1}{2}(d_{\rightarrow} - 1) \\ \beta \leqslant \alpha + \frac{d_{\leftarrow} - d_{\rightarrow}}{2} \end{cases} . \tag{4.42}$$

This gives a wide range of optimal parameters. For instance, the parameter α can be chosen arbitrarily large. In Figures 4.8B and 4.9B, the shaded regions corresponds to region for (α, β) defined by (4.42) with $(d_{\leftarrow}, d_{\rightarrow}) = (2, 2)$.

Note however that we have only proved that (4.42) are *sufficient* conditions for the optimality Theorem 4.5 to hold. To explore the tightness of our condition, we present in Figure 4.8 the results of simulations on the 2D grid. The setting is the same as in Section 4.1 (see also Appendix 4.A for details). Note that for the 2D infinite grid, $d_{\leftarrow} = d_{\rightarrow} = 2$ (see Proposition 4.4 and the symmetry of the spectrum of $\mathbb{Z}^d$ that follows from [Woess, 2000, Eq.(7.4)]). The curves in Figure 4.8A closest to the local averaging are those satisfying the condition (4.42), thus our condition seems tight.

Finally, note that the result (4.41) of Theorem 4.5 gives the rate of the power decay of the MSE, but neglects constants and sub-polynomial factors. These factors depend on (α, β) and can be significant for extreme values of (α, β). For instance, in Figure 4.9, we run simulations in the same setting as before, but for choices of parameters deeper in the optimality zone (4.42). The performance worses as α gets bigger. So contrarily to what is suggested by (4.42) and Theorem 4.1, taking large values for α is a bad idea in practice. This can also be hinted at by the limit [Olver et al., 2010, Eq. (18.6.2)]

$$\lim_{\alpha\to\infty} \pi_n^{(\alpha,\beta)}(\lambda) = \left(\frac{1+\lambda}{2}\right)^n .$$

This means that, as $\alpha \to \infty$, the polynomial gossip $x_n = \pi_n^{(\alpha,\beta)}(W)x_0$ converges to the simple gossip $x_n = \tilde{W}^n x_0$ with the gossip matrix $\tilde{W} = (I + W)/2$. We know from Theorem 4.1(1) that simple gossip is suboptimal.

Overall, theory and practice suggest that the choice $\alpha = \underline{\dim}_{\rightarrow}\sigma/2$, $\beta = 0$ that we make in Section 4.3.2 is relevant.

4.J. Proof of Proposition 4.7

The intuition lying behind the proposition is very simple: the unbiased estimator $x_n(v)$ are linear combination of observations corresponding to vertices in the ball $B_v(n)$, thus it must have variance greater than $\operatorname{var}\nu/|B_v(n)| \approx \operatorname{var}\nu/n^{d_h}$.

A more rigorous argument goes as follows: using that W is a gossip matrix, it is easy to show by induction that for all $k \geqslant 0$ and $v, w \in \mathcal{V}$, if $(W^k)_{vw} > 0$, then there exists a path of length k linking v to w in G. As $\deg P_n \leqslant n$, this implies that $P_n(W)e_v$ has at most $|B_v(n)|$ non-zero entries. Furthermore, the entries of $P_n(W)e_v$ sum to 1 because $W\mathbb{1} = \mathbb{1}$ and $P_n(1) = 1$. Thus, using the Cauchy-Schwarz inequality,

$$\begin{aligned}
1 &= \left(\sum_{w\in\mathcal{V}} (P_n(W)e_v)_w\right)^2 = \left(\sum_{w\in V} (P_n(W)e_v)_w \mathbb{1}_{\{(P_n(W)e_v)_w>0\}}\right)^2 \\
&\leqslant \|P_n(W)e_v\|^2_{\ell^2(\mathcal{V})} \sum_{w\in V} \mathbb{1}_{\{(P_n(W)e_v)_w>0\}} \leqslant \|P_n(W)e_v\|^2_{\ell^2(\mathcal{V})}\, |B_v(n)| \\
&\overset{\text{(Lemma 4.5)}}{=} \mathbb{E}[(x_n(v)-\mu)^2]|B_v(n)| .
\end{aligned}$$

Thus

$$\liminf_{n\to\infty} \frac{\log \mathbb{E}[(x_n(v)-\mu)^2]}{\log n} \geqslant \liminf_{n\to\infty} -\frac{\log|B_v(n)|}{\log n} = -d_h .$$

4.K. Proof of Theorem 4.2

In this section, we use the notation of Appendix 4.H.3. As $x_n = \pi_n^{(d/2,0,\mu)}(W)x_0$, we have

$$\|x_n - \bar{x}\mathbb{1}\|_2^2 = \sum_{i=2}^m \langle x_0, u_i\rangle^2 \pi_n^{(d/2,0,\mu)}(\lambda_i)^2 \leqslant \|x_0 - \bar{x}\mathbb{1}\|_2^2 \left(\sup_{\lambda\in[-1,1-\mu]} |\pi_n^{(d/2,0,\mu)}(\lambda)|\right)^2 , \tag{4.43}$$

where $\lambda_2, \dots, \lambda_m$ are the eigenvalues of W different from 1, that lie in $[-1, 1-\mu]$ by definition of μ, and $u_2, \dots, u_m$ are the corresponding normalized eigenvectors.

$$\begin{aligned}
\sup_{\lambda\in[-1,1-\mu]} |\pi_n^{(d/2,0,\mu)}(\lambda)| &\leqslant \frac{1}{|P_n^{(d/2,0,\mu)}(1)|} \sup_{\lambda\in[-1,1-\mu]} |P_n^{(d/2,0,\mu)}(\lambda)| \\
&= \frac{1}{|\pi_n^{(d/2,0)}(\varphi_\mu^{-1}(1))|} \sup_{\lambda\in\varphi_\mu^{-1}([-1,1-\mu])} |\pi_n^{(d/2,0)}(\lambda)| \\
&= \frac{1}{\left|\pi_n^{(d/2,0)}\left(\frac{1+\mu/2}{1-\mu/2}\right)\right|} \sup_{\lambda\in[-1,1]} |\pi_n^{(d/2,0)}(\lambda)| \\
&= \frac{1}{\left|P_n^{(d/2,0)}\left(\frac{1+\mu/2}{1-\mu/2}\right)\right|} \sup_{\lambda\in[-1,1]} |P_n^{(d/2,0)}(\lambda)| \qquad (4.44)
\end{aligned}$$

where $P_n^{(\alpha,\beta)}$ is the Jacobi polynomial, see Appendix 4.H.2.

By Proposition 4.11,

$$\sup_{\lambda\in[-1,1]} |P_n^{(d/2,0)}(\lambda)| = \binom{n+d/2}{n} \underset{n\to\infty}{\sim} n^{d/2}\,, \qquad (4.45)$$

an by Proposition 4.14 applied in $x = \frac{1+\mu/2}{1-\mu/2}$, there exists a positive constant c such that,

$$P_n^{(d/2,0)}\left(\frac{1+\mu/2}{1-\mu/2}\right) \underset{n\to\infty}{\sim} cn^{-1/2}\left(\frac{(1+\sqrt{\mu/2})^2}{1-\mu/2}\right)^n. \qquad (4.46)$$

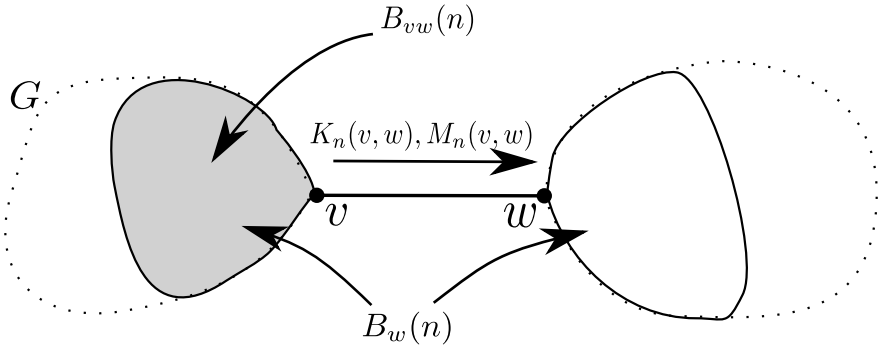
Combining (4.44), (4.45) and (4.46), we get that there exists a constant C such that

$$\sup_{\lambda\in[-1,1-\mu]} |\pi_n^{(d/2,0,\mu)}(\lambda)| \leqslant Cn^{(d+1)/2}\left(\frac{1-\mu/2}{(1+\sqrt{\mu/2})^2}\right)^n,$$

and we conclude using (4.43).

4.L. Proof of Proposition 4.8

Let $n \geqslant 0$ and $v, w \in \mathcal{V}$ be two vertices linked by an edge in G. Define $B_{vw}(n)$ as the set of vertices u in $B_w(n)$ such that all paths in the tree G going from u to w pass though v.



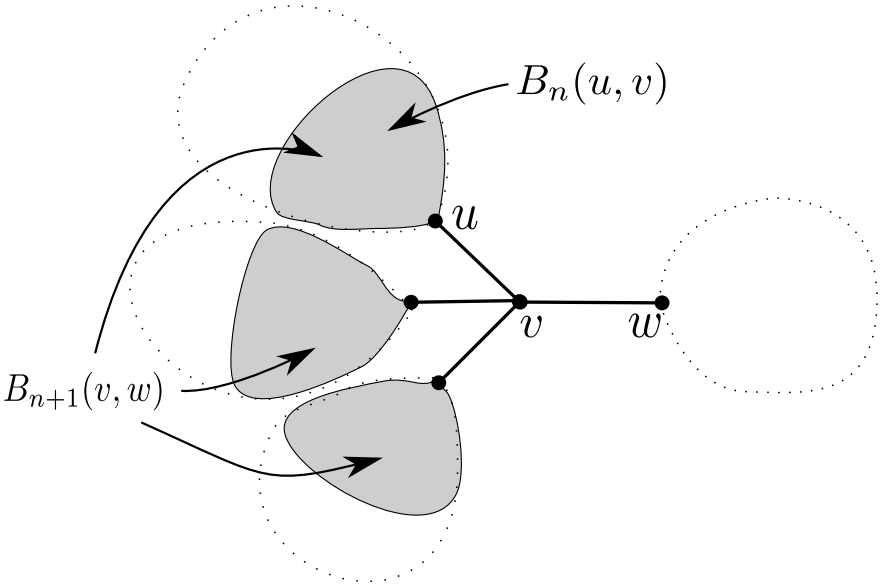
Lemma 4.8. For all $n \geqslant 0$, for all $v, w \in \mathcal{V}$ linked by an edge in G,

$$K_n(v,w) = |B_{vw}(n)|\,, \qquad \text{and} \quad \text{if } n \geqslant 1, \quad M_n(v,w) = \frac{1}{|B_{vw}(n)|}\sum_{u\in B_{vw}(n)} x_0(u)\,.$$

PROOF. The proof goes by induction. The statement is trivial for $n = 0, 1$. For the induction, assume the result at time n and note that

$$B_{vw}(n+1) = \{v\} \cup \left(\bigcup_{u:u\sim v,\, u\neq w} B_{uv}(n) \right) , \tag{4.47}$$

where all unions are disjoint. This essentially comes from the fact that G has no loops.



Taking cardinal, we get that

$$|B_{vw}(n+1)| \overset{(4.47)}{=} 1 + \sum_{u:u\sim v,\, u\neq w} |B_{uv}(n)| \overset{\text{(induction)}}{=} 1 + \sum_{u:u\sim v,\, u\neq w} K_n(u,v) \overset{(4.16)}{=} K_{n+1}(v,w)\,.$$

This proves the induction for the first equality. The proof for the second equality is similar:

$$\begin{aligned}\frac{1}{|B_{vw}(n+1)|} \sum_{u\in B_{vw}(n+1)} x_0(u) &\overset{(4.47)}{=} \frac{1}{K_{n+1}(v,w)} \left(x_0(v) + \sum_{u:u\sim v,\, u\neq w} \sum_{z\in B_n(u,v)} x_0(z) \right) \\ &\overset{\text{(induction)}}{=} \frac{x_0(v) + \sum_{u:u\sim v,\, u\neq w} |B_{uv}(n)| M_n(u,v)}{K_{n+1}(v,w)} \\ &\overset{\text{(induction)}}{=} \frac{x_0(v) + \sum_{u:u\sim v,\, u\neq w} K_n(u,v) M_n(u,v)}{K_{n+1}(v,w)} \overset{(4.16)}{=} M_{n+1}(v,w)\,.\end{aligned}$$

□

We now end the proof of Proposition 4.8. As $B_v(n) = \{v\} \cup (\bigcup_{u:u\sim v} B_{uv}(n))$ with disjoint unions, using Lemma 4.8, we get

$$\begin{aligned}\frac{1}{|B_v(n)|} \sum_{w\in B_v(n)} x_0(w) &= \frac{x_0(v) + \sum_{u:u\sim v} \sum_{w\in B_n(u,v)} x_0(w)}{1 + \sum_{u:u\sim v} |B_n(u,v)|} = \frac{x_0(v) + \sum_{u:u\sim v} K_n(u,v) M_n(u,v)}{1 + \sum_{u:u\sim v} K_n(u,v)} \\ &\overset{(4.17)}{=} x_n(v)\,.\end{aligned}$$

4.M. Proof of Theorem 4.3

As noted by Rebeschini and Tatikonda [2017], the message passing iteration (4.16)-(4.17) indexed by the edges of the graph can be written as an iteration indexed by the vertices of the graph. We repeat here the elementary derivation of this statement in our particular case of d-regular graphs.

First, because G is d-regular, it is an easy check from (4.16) that $K_n(v,w)$ does not depend on the edge (v,w) (thus we denote it K_n) and it satisfies the recursion $K_0 = 0$, $K_{n+1} = 1 + (d-1)K_n$.

Let us now denote $S_n(v) = x_0(v) + \sum_{u:u\sim v} K_n M_n(u,v)$ and $L_n = 1 + dK_n$ so that $x_n(v) = S_n(v)/L_n$. We now find recursions for L_n and S_n:

$$L_{n+1} = 1 + dK_{t+1} \overset{(4.16)}{=} 1 + d(1 + (d-1)K_n) = 2 + (d-1)(1 + dK_n) = 2 + (d-1)L_n\,,$$

and

$$\begin{aligned} S_{n+1}(v) &= x_0(v) + \sum_{u:u\sim v} K_{n+1}M_{n+1}(u,v) \overset{(4.16)}{=} x_0(v) + \sum_{u:u\sim v}\left(x_0(u) + \sum_{w:w\sim u,\, w\neq v} K_n M_n(w,u)\right) \\ &= x_0(v) + \sum_{u:u\sim v}\left(S_n(u) - K_n M_n(v,u)\right)\,. \end{aligned}$$

As

$$\begin{aligned} \sum_{u:u\sim v} K_n M_n(v,u) &\overset{(4.16)}{=} dx_0(v) + \sum_{u:u\sim v}\sum_{w:w\sim v,\, w\neq u} K_{n-1}M_{n-1}(w,v) \\ &= dx_0(v) + (d-1)\sum_{w:w\sim v} K_{n-1}M_{n-1}(w,v) = x_0(v) + (d-1)S_{n-1}(v)\,, \end{aligned}$$

we finally get

$$S_{n+1} = A(G)S_n - (d-1)S_{n-1}\,.$$

To sum up, we now have the simpler formulas for the message passing algorithm:

$$\begin{aligned} &L_{n+1} = 2 + (d-1)L_n\,, && L_0 = 1\,, \\ &S^{n+1} = dWS^n - (d-1)S^{n-1}\,, && S^0 = x_0\,,\quad S^1 = x_0 + dWx_0\,, \\ &x_n = S_n/L_n\,. \end{aligned} \tag{4.48}$$

In Appendix 4.H.4, it is proved that $\pi_n(\lambda) = p_n(\lambda)/p_n(1)$ where p_n satisfies the recursion formula

$$p_0(\lambda) = \sqrt{d-1}\,, \qquad p_1(\lambda) = d\lambda + 1\,, \qquad p_{n+1}(\lambda) = \frac{d}{\sqrt{d-1}}\lambda p_n(\lambda) - p_{n-1}(\lambda)\,, \quad n \geqslant 1\,.$$

Denote $q_n = (d-1)^{(n-1)/2}p_n$. It is an easy check that

$$q_0(\lambda) = 1\,, \qquad q_1(\lambda) = d\lambda + 1\,, \qquad q_{n+1}(\lambda) = d\lambda q_n(\lambda) - (d-1)q_{n-1}(\lambda)\,, \quad n \geqslant 1\,.$$

Using (4.48), one sees that for all n, $S^n = q_n(W)x_0$ and $L_n = q_n(1)$. Thus

$$x_n = \frac{S_n}{L_n} = \frac{q_n(W)x_0}{q_n(1)} = \frac{p_n(W)x_0}{p_n(1)} = \pi_n(W)x_0\,.$$

4.N. Proof of Theorem 4.4

Theorem 4.3 states that $x_n = \pi_n(W)x_0$ where the π_n are the orthogonal polynomials w.r.t. the modified Kesten-McKay measure $(1-\lambda)\sigma(\mathbb{T}_d)(\mathrm{d}\lambda)$. Then

$$\|x_n - \bar{x}\mathbb{1}\|_2^2 = \sum_{i=2}^m \langle x_0, u_i\rangle^2 \pi_n(\lambda_i)^2 \leqslant \|x_0 - \bar{x}\mathbb{1}\|_2^2 \left(\sup_{\lambda\in[-(1-\tilde{\mu}),(1-\tilde{\mu})]} |\pi_n(\lambda)|\right)^2\,, \tag{4.49}$$

where $\lambda_2, \dots, \lambda_n$ are the eigenvalues of W different from 1, that lie in $[-(1-\tilde{\mu}), (1-\tilde{\mu})]$ by definition of the absolute spectral gap $\tilde{\mu}$, and $u_2, \dots, u_n$ are the corresponding normalized eigenvectors.

In Section 4.H.4, we show that

$$\begin{aligned}
\pi_n(\lambda) &= \frac{p_n(\lambda)}{p_n(1)}\,, \\
p_n(\lambda) &= \tilde{p}_n(\varphi^{-1}(\lambda))\,, \qquad \varphi(\lambda) = \frac{2\sqrt{d-1}}{d}\lambda\,, \\
\tilde{p}_n(\lambda) &= \left(\sqrt{d-1}+\lambda\right) U_n(\lambda) - T_{n+1}(\lambda)\,,
\end{aligned}$$

where T_n and U_n are the Chebyshev polynomials of the first kind and of the second kind respectively. Denote $D = d/(2\sqrt{d-1})$. Then

$$\sup_{\lambda\in[-(1-\tilde{\mu}),(1-\tilde{\mu})]} |\pi_n(\lambda)| = \frac{1}{|\tilde{p}_n(D)|} \sup_{\lambda\in[-(1-\tilde{\mu})D,(1-\tilde{\mu})D]} |\tilde{p}_n(\lambda)|\,. \tag{4.50}$$

If $\lambda \in [-1,1]$, $|T_n(\lambda)| \leqslant 1$ and $|U_n(\lambda)| \leqslant n+1$. Thus

$$\sup_{\lambda\in[-1,1]} |\tilde{p}_n(\lambda)| \leqslant \left(\sqrt{d-1}+1\right)(n+1)+1\,. \tag{4.51}$$

We now discuss the different cases of the theorem.

(1) We assume $\tilde{\mu} < 1 - 2\sqrt{d-1}/d$. As $\tilde{p}_n$ are orthogonal polynomials w.r.t. some measure on $[-1,1]$, all zeros of p_n are real, distinct and located in the interior of $[-1,1]$ (see Proposition 4.9). It follows that

$$\sup_{\lambda\in(1,(1-\tilde{\mu})D]} |\tilde{p}_n(\lambda)| = |\tilde{p}_n\left((1-\tilde{\mu})D\right)|\,, \qquad \sup_{\lambda\in[-(1-\tilde{\mu})D,-1)} |\tilde{p}_n(\lambda)| = |\tilde{p}_n\left(-(1-\tilde{\mu})D\right)|\,. \tag{4.52}$$

Merging Eqs. (4.50)-(4.52), we obtain

$$\sup_{\lambda\in[-(1-\tilde{\mu}),(1-\tilde{\mu})]} |\pi_n(\lambda)| \leqslant \frac{1}{|\tilde{p}_n(D)|} \max\left(|\tilde{p}_n((1-\tilde{\mu})D)|, |\tilde{p}_n(-(1-\tilde{\mu})D)|, (\sqrt{d-1}+1)(n+1)+1\right)\,.$$

Lemma 4.9. (1) If $\lambda > 1$, then there exists a constant $C(d,\lambda) \neq 0$ such that

$$\tilde{p}_n(\lambda) \underset{n\to\infty}{\sim} C(d,\lambda)\left(\lambda+\sqrt{\lambda^2-1}\right)^n\,. \tag{4.53}$$

(2) If $\lambda < -1$, then there exists a constant $C(d,\lambda) \neq 0$ such that

$$\tilde{p}_n(\lambda) \underset{n\to\infty}{\sim} C(d,\lambda)\left(\lambda-\sqrt{\lambda^2-1}\right)^n\,. \tag{4.54}$$

Proof. In the proof of Lemma 4.1, we developed the following formulas for the Chebyshev polynomials:

$$T_n\left(\frac{z+z^{-1}}{2}\right) = \frac{z^n+z^{-n}}{2}\,, \qquad U_n\left(\frac{z+z^{-1}}{2}\right) = \frac{z^{n+1}-z^{-(n+1)}}{z-z^{-1}}\,.$$

We write $\lambda = (z+z^{-1})/2$ with $|z| > 1$. If $\lambda > 1$, then $z = \lambda + \sqrt{\lambda^2-1}$, and if $\lambda < -1$, then $z = \lambda - \sqrt{\lambda^2-1}$. Then

$$\begin{aligned}
\tilde{p}_n(\lambda) &= \left(\sqrt{d-1}+\lambda\right) U_n(\lambda) - T_{n+1}(\lambda) \\
&= \left(\sqrt{d-1}+\frac{z+z^{-1}}{2}\right)\frac{z^{n+1}-z^{-(n+1)}}{z-z^{-1}} - \frac{z^{n+1}+z^{-(n+1)}}{2}
\end{aligned}$$

$$\underset{n\to\infty}{\sim} \left[\left(\sqrt{d-1}+\frac{z+z^{-1}}{2}\right)\frac{1}{z-z^{-1}}-\frac{1}{2}\right] z^{n+1}\,.$$

The constant that appears is non-zero, thus the result is proved. □

Using Lemma 4.9, we get that there exists a constant $C(d)$ such that

$$\sup_{\lambda\in[-(1-\tilde{\mu}),(1-\tilde{\mu})]} |\pi_n(\lambda)| \leqslant C(d)\left(\frac{(1-\tilde{\mu})D+\sqrt{((1-\tilde{\mu})D)^2-1}}{D+\sqrt{D^2-1}}\right)^n.$$

Finally, using (4.49), this gives

$$\|x_n-\bar{x}\mathbb{1}\|_2 \leqslant \|x_0-\bar{x}\mathbb{1}\|_2 C(d)\left(\frac{(1-\tilde{\mu})D+\sqrt{((1-\tilde{\mu})D)^2-1}}{D+\sqrt{D^2-1}}\right)^n.$$

Dividing the numerator and the denominator of the fraction by D, we get the desired result.

We now turn to the second part of the statement. Let u be an eigenvector of W corresponding to an eigenvalue λ of magnitude $1-\tilde{\mu}$ such that $\langle x_0,u\rangle\neq 0$. Then

$$\|x_n-\bar{x}\mathbb{1}\|_2 \geqslant |\langle x_0,u\rangle||\pi_n(\lambda)| = |\langle x_0,u\rangle|\frac{|\tilde{p}_n(\lambda)|}{|\tilde{p}_n(D)|}\,,$$

Using as before Lemma 4.9, we get the desired lower bound.

(2) We now assume $\tilde{\mu}\geqslant 1-2\sqrt{d-1}/d$. This means that $(1-\tilde{\mu})D\leqslant 1$, and thus

$$\sup_{\lambda\in[-(1-\tilde{\mu})D,(1-\tilde{\mu})D]}|\tilde{p}_n(\lambda)| \leqslant \sup_{\lambda\in[-1,1]}|\tilde{p}_n(\lambda)| \overset{(4.51)}{\leqslant} \left(\sqrt{d-1}+1\right)(n+1)+1\,.$$

Combining with (4.49) and (4.50), we get

$$\|x_n-\bar{x}\mathbb{1}\|_2 \leqslant \|x_0-\bar{x}\mathbb{1}\|_2\frac{1}{|\tilde{p}_n(D)|}\left(\left(\sqrt{d-1}+1\right)(n+1)+1\right),$$

which gives the desired result using Lemma 4.9.

CHAPTER 5

Scaling Limits of Synchronous Gossip Algorithms to Partial Differential Equations

We recall that this chapter is a preliminary version of joint work with Mufan Li, currently a PhD student at the University of Toronto. He suggested looking at the scaling limit of the Jacobi polynomial equation and helped identifying the Euler–Poisson–Darboux equation.

In this chapter, we study synchronous gossip algorithms on $\mathbb{Z}^d$ endowed with a translation-invariant gossip operation. As $\mathbb{Z}^d$ is infinite, it is unclear what the average of an initial vector $x_0 = (x_0(v))_{v\in\mathbb{Z}^d}$ is; the averaging problem is thus ill-posed. Here, our strategy is to study the decay to 0 of the algorithms when initialized from $\mathbb{1}_0$, the vector with entry $\mathbb{1}_0(0) = 1$ and all other entries $\mathbb{1}_0(v)$, $v \in \mathbb{Z}^d\backslash\{0\}$, equal to 0. By analogy with partial differential equations (PDEs), this is the fundamental solution of the gossip iterations; the solutions for other initializations x_0 can be obtained by convoluting x_0 with the fundamental solution.

Restricting ourselves to $\mathbb{Z}^d$ with a translation-invariant gossip operation provides two advantages. First, we can use the Fourier transform on $\mathbb{Z}^d$ to analyze the behavior of the iterates. Second, the graph $\mathbb{Z}^d$ is naturally embedded in $\mathbb{R}^d$: this enables to rescale the iterates in space. These two tools enable analyses finer than those of the previous chapters.

Specifically, we show that gossip algorithms converge to PDEs when appropriately rescaled simultaneously in time and space. In Section 5.1, we give the heuristic derivations of these scaling limits. We show that the simple gossip algorithm converges to the heat equation (see, e.g., [Evans, 1998])

$$\partial_t u = \frac{1}{2}\nabla_y \cdot (Q\nabla_y u)\,, \qquad u = u(t,y)\,. \tag{5.1}$$

Here, ∇_y and $\nabla_y\cdot$ denote respectively the gradient and the divergence operator in the variable y. Q is a $d\times d$ matrix quantifying the potential anisotropy of the diffusion: it is a function of the local averaging operation. The fundamental solution of the heat equation (5.1), i.e., the solution when initialized at the Dirac mass $u(0,.) = \delta_0$, is

$$u(t,y) = \frac{1}{(2\pi)^{d/2}t^{d/2}(\det Q)^{1/2}} \exp\left(-\frac{1}{2t}\left\langle y, Q^{-1}y\right\rangle\right)\,. \tag{5.2}$$

The formula above shows the sub-optimality of the simple gossip method: the mass spreads on a typical scale $\|y\| \approx \sqrt{t}$, while we would like the scale to be $\|y\| \approx t$; indeed, the gossiped information can travel at most at distance $\Theta(t)$ in a time t, and we would like our gossip algorithms to match this optimal speed of diffusion. Equivalently, the solution decays to 0 at the rate $1/t^{d/2}$ in $\|.\|_\infty$, while we would like the rate to be $1/t^d$.

We design an accelerated second-order gossip iteration: we choose the recursion coefficients so that the iteration scales to the Euler–Poisson–Darboux (EPD) equation [Euler, 1770, Poisson, 1823, Darboux, 1896]

$$\partial_{tt}u + \frac{d+1}{t}\partial_t u = \nabla_y \cdot (Q\nabla_y u)\,. \tag{5.3}$$

See Appendix 5.A or [Bresters, 1973] for an introduction to the EPD equation in a more general form. The EPD equation is a subtle combination of the wave equation and of the heat equation. The wave component gives inertia to the diffusion so that the resulting PDE mixes faster. Remarkably, for this precise value of the coefficient $\frac{d+1}{t}$, the fundamental solution is

$$u(0,.) = \delta_0\,, \qquad \partial_t u(0,.) = 0\,, \qquad u(t,y) = \frac{\Gamma(d/2+1)}{\pi^{d/2}(\det Q)^{1/2}} \frac{1}{t^d} \mathbb{1}_{\{\langle y, Q^{-1}y\rangle \leqslant t^2\}}\,. \tag{5.4}$$

This method thus has an optimal scaling: the mass spreads on a typical scale $\|y\| \approx t$ and the solution decays to 0 at the rate $1/t^d$. But the fundamental solution also has the perfect shape: the averaging is uniform on the ellipsoid $\{\langle y, Q^{-1}y\rangle \leqslant t^2\}$.

The recursion coefficients of the accelerated second-order method only need to satisfy some asymptotic properties for the method to scale to the EDP equation (5.3). Interestingly, the Jacobi polynomial iteration (4.2)-(4.3) satisfies these asymptotics. We thus provide a different motivation for the Jacobi polynomial iteration, that is derived in Chapter 4 through algebraic methods on polynomials. However, the new derivation suggests that the precise formula (4.3) for the Jacobi polynomial iteration does not matter: only certain asymptotics must be satisfied.

In Figures 5.1-5.2, we provide simulations in dimension $d = 1$ and $d = 2$. They show that the limiting PDEs are sharp in describing the behavior of gossip algorithms as the number of iterations grows, and that the accelerated methods achieve faster diffusion.

The derivations described above are only heuristic in Section 5.1: we do not specify how we measure the convergence of gossip iterations to the limiting PDEs. In Section 5.2, we provide rigorous meanings to these convergences.

The simple gossip algorithm can be seen as the iteration of the law of a random walk on $\mathbb{Z}^d$; the convergence to a Gaussian random variable is made rigorous by the central limit theorem and the local central limit theorem, see Section 5.2.1. In Section 5.2.2, we provide analog results for the convergence of the Jacobi polynomial iteration to the EPD equation (5.3): a weak limit theorem and a stronger result of local type. Finally, in Section 5.2.3, we apply the latter result to obtain an asymptotic equivalent of convergence rate (not only a domination)

$$\sum_{v\in\mathbb{Z}^d} x_n(v)^2 \underset{n\to\infty}{\sim} \frac{1}{(\det Q)^{1/2}|B(0,1)|} \frac{1}{n^d}$$

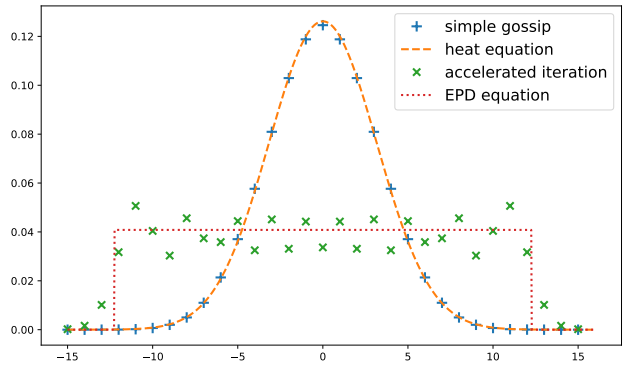
for the Jacobi polynomial iteration on $\mathbb{Z}^d$.

Notation. For $v \in \mathbb{Z}^d$, we denote $\mathbb{1}_v = (\mathbb{1}_v(w))_{w\in\mathbb{Z}^d}$ the vector with entry $\mathbb{1}_v(v) = 1$ and all other entries equal to 0. We denote $e_1, \ldots, e_d$ the canonical basis of $\mathbb{R}^d$. $\lfloor s \rfloor$ denotes the integer part of a real number s.

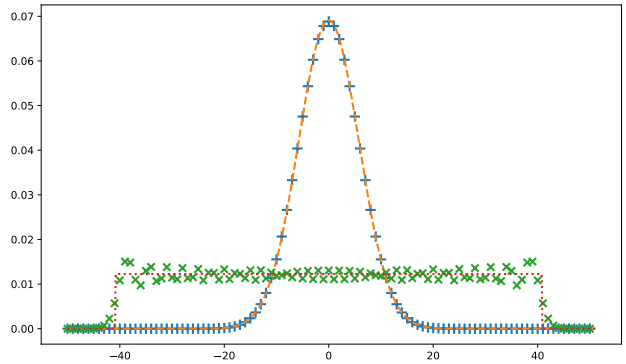
5.1. Heuristic derivation of the Euler–Poisson–Darboux gossip algorithm

Let $\omega = (\omega(v))_{v\in\mathbb{Z}^d}$ be a vector of non-negative reals, representing a local averaging operation on $\mathbb{Z}^d$. We assume that ω has finite support and that $\sum_{v\in\mathbb{Z}^d} \omega(v) = 1$. In this section, our synchronous gossip operator W is the convolution by ω. If $x = (x(v))_{v\in\mathbb{Z}^d}$ and $*$ denotes the convolution on $\mathbb{Z}^d$, $\omega * x$ represents the computation, by each node v, of a weighted linear combination of the values of its neighbors:

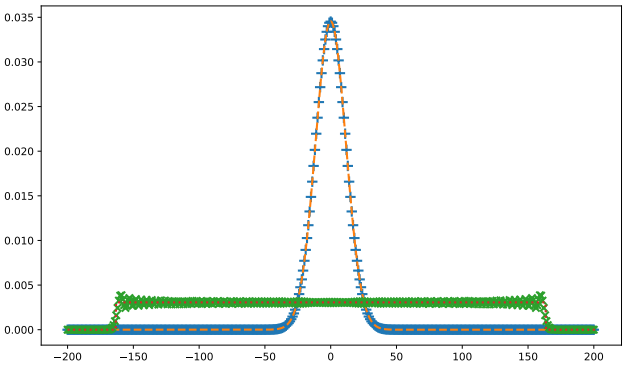
$$(\omega * x)(v) = \sum_{w\in\mathbb{Z}^d} \omega(w) x(v-w)\,.$$



(A) n = 15

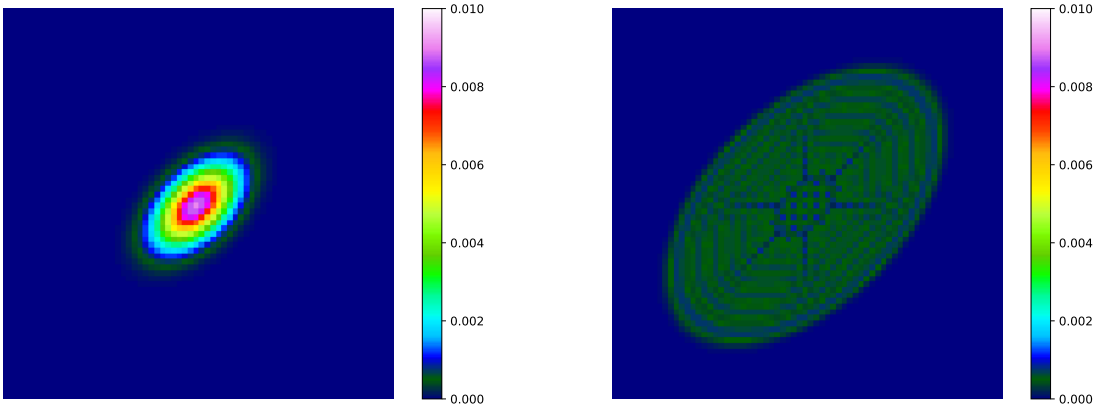


(B) n = 50



(C) n = 200

FIGURE 5.1. Comparison between gossip algorithms and their scaling limits: the simple gossip and the heat equation, the accelerated Jacobi polynomial iteration and the EPD equation. All iterations were run on $\mathbb{Z}$ ($d = 1$) and initialized from $x_0 = \mathbb{1}_0$. We show the results $x_n(v)$ as a function of $v \in \mathbb{Z}$ for different numbers of iterations $n = 15, 50, 200$. Note that as the number of iteration increases, the description through the scaling limits becomes sharp. The accelerated Jacobi polynomial iteration diffuses faster; it has a different scaling than the simple gossip algorithm.



(A) Simple gossip (B) Accelerated Jacobi polynomial iteration

FIGURE 5.2. Comparison between simple gossip and the accelerated Jacobi polynomial iteration on the triangular lattice (5.6). We initialize from $x_0 = \mathbb{1}_0$ and we show the iterates $x_{30} = (x_{30}(v))_{v\in\mathbb{Z}^2}$ using a color scale. The accelerated Jacobi polynomial iteration diffuses faster than simple gossip: the mass is distributed more evenly and on a larger ellipsoid.

The support of ω represents the range of the possible communications in $\mathbb{Z}^d$. Typically, we can take

$$\omega = \frac{1}{2d}\sum_{i=1}^{d}(\mathbb{1}_{e_i} + \mathbb{1}_{-e_i})\,. \tag{5.5}$$

This corresponds to allowing only nearest neighbors communicate in $\mathbb{Z}^d$. In this case the underlying graph is the standard lattice on $\mathbb{Z}^d$. The triangular lattice in dimension 2 can be obtained by taking

$$\omega = \frac{1}{6}\left(\mathbb{1}_{(1,0)} + \mathbb{1}_{(-1,0)} + \mathbb{1}_{(0,1)} + \mathbb{1}_{(0,-1)} + \mathbb{1}_{(1,1)} + \mathbb{1}_{(-1,-1)}\right)\,. \tag{5.6}$$

Vectors ω with a larger support require larger ranges of communications.

Throughout this chapter, we assume that ω is centered, i.e.,

$$\sum_{v\in\mathbb{Z}^d}\omega(v)v = 0\,,$$

and we denote Q the covariance of ω, i.e.,

$$Q = \sum_{v\in\mathbb{Z}^d}\omega(v)vv^\top\,.$$

We assume that Q has full rank. Heuristically, this ensures that we average in all directions.

5.1.1. Scaling limit of the simple gossip algorithm to the heat equation. The synchronous simple gossip algorithm iterates the local averaging operation:

$$x_{n+1}(v) = (\omega * x_n)(v) = \sum_{w\in\mathbb{Z}^d}\omega(w)x_n(v-w)\,,\qquad v\in\mathbb{Z}^d\,. \tag{5.7}$$

Let $\Delta t, \Delta y > 0$ denote two scaling parameters. For $t \in (\Delta t)\mathbb{N} = \{(\Delta t)n,\, n \in \mathbb{N}\}$ and $y \in (\Delta y)\mathbb{Z}^d = \{(\Delta y)v,\, v \in \mathbb{Z}^d\}$, we define the scaled field

$$u(t,y) = x_{\frac{t}{\Delta t}}\left(\frac{y}{\Delta y}\right).$$

The iteration (5.7) can be reformulated in terms of u:

$$u(t+\Delta t, y) = \sum_{w\in\mathbb{Z}^d} \omega(w) u(t, y-(\Delta y)w)\,.$$

We now show that under a proper scaling for $\Delta t, \Delta y \to 0$, the above equation converges to a PDE in u. Recall that $\sum_{v\in\mathbb{Z}^d}\omega(v) = 1$, thus

$$u(t+\Delta t,y) - u(t,y) = \sum_{w\in\mathbb{Z}^d} \omega(w)\left[u(t,y-(\Delta y)w) - u(t,y)\right].$$

We take $\Delta t, \Delta y \to 0$ and make Taylor expansions of the differences:

$$u(t+\Delta t,y) - u(t,y) = (\Delta t)\partial_t u + o(\Delta t)\,,$$

$$u(t,y-(\Delta y)w) - u(t,y) = -(\Delta y)\left\langle \nabla_y u, w\right\rangle + \frac{(\Delta y)^2}{2}\left\langle w, \left(\nabla_y^2 u\right) w\right\rangle + o\left((\Delta y)^2\right),$$

where all derivatives are taken in (t,y). Note that we make a second-order expansion in space: this is due to the fact that the first-order terms cancel below. We obtain

$$(\Delta t)\partial_t u + o(\Delta t) = -(\Delta y)\left\langle \nabla_y u, \sum_{w\in\mathbb{Z}^d}\omega(w) w\right\rangle + \frac{(\Delta y)^2}{2}\sum_{w\in\mathbb{Z}^d}\omega(w)\left\langle w, \left(\nabla_y^2 u\right) w\right\rangle + o\left((\Delta y)^2\right).$$

As ω is centered, the first term of the right-hand side is zero. Moreover, we can rewrite

$$\sum_{w\in\mathbb{Z}^d}\omega(w)\left\langle w, \left(\nabla_y^2 u\right) w\right\rangle = \mathrm{Tr}\left(Q\nabla_y^2 u\right) = \nabla_y\cdot(Q\nabla_y u)\,.$$

We obtain

$$(\Delta t)\partial_t u + o(\Delta t) = \frac{(\Delta y)^2}{2}\nabla_y\cdot(Q\nabla_y u) + o\left((\Delta y)^2\right).$$

We choose the scaling $\Delta t = (\Delta y)^2$ and by identifying the highest-order terms, we obtain the scaling to the heat equation

$$\partial_t u = \frac{1}{2}\nabla_y\cdot(Q\nabla_y u)\,.$$

Here, Q quantifies the potential anisotropy of the diffusion. In the case of (5.5), we have $Q = \frac{1}{d}\,\mathrm{Id}$ and thus we obtain an isotropic heat equation $\partial_t u = \frac{1}{2d}\Delta_y u$, where Δ_y denotes the Laplacian in the variable y.

5.1.2. Second-order iteration scaling to the Euler–Poisson–Darboux equation. We now consider second-order iterations of the form

$$x_{n+1}(v) = a_n \sum_{w\in\mathbb{Z}^d}\omega(w)x_n(v-w) + b_n x_n(v) - c_n x_{n-1}(v)\,. \tag{5.8}$$

We impose $a_n + b_n - c_n = 1$ so that the sum of the coordinates of the vectors x_n remains constant. We show that, under specific asymptotics for a_n, b_n, c_n, the iteration (5.8) scales to the EPD equation. As in Section 5.1.1, we introduce scaling parameters $\Delta t, \Delta y > 0$ and the rescaled iterates

$$u(t,y) = x_{\frac{t}{\Delta t}}\left(\frac{y}{\Delta y}\right).$$

The iteration (5.8) can be reformulated in terms of u:

$$u(t+\Delta t,y)=a_n\sum_{w\in\mathbb{Z}^d}\omega(w)u(t,y-(\Delta y)w)+b_nu(t,y)-c_nu(t-\Delta t,y)\,.$$

Substracting $u(t,y)$ and using $a_n+b_n-c_n=1$, we obtain

$$u(t+\Delta t,y)-u(t,y)=a_n\sum_{w\in\mathbb{Z}^d}\omega(w)\left[u(t,y-(\Delta y)w)-u(t,y)\right]-c_n\left[u(t-\Delta t,y)-u(t,y)\right]\,.$$

We make the Taylor expansions of u, but this time a second-order expansion in t is necessary:

$$\begin{aligned}u(t+\Delta t,y)-u(t,y)&=(\Delta t)\partial_tu+\frac{(\Delta t)^2}{2}\partial_{tt}u+o(\Delta t)\,,\\ u(t-\Delta t,y)-u(t,y)&=-(\Delta t)\partial_tu+\frac{(\Delta t)^2}{2}\partial_{tt}u+o(\Delta t)\,,\\ u(t,y-(\Delta y)w)-u(t,y)&=-(\Delta y)\left\langle\nabla_yu,w\right\rangle+\frac{(\Delta y)^2}{2}\left\langle w,\left(\nabla_y^2u\right)w\right\rangle+o\left((\Delta y)^2\right)\,.\end{aligned}$$

We obtain

$$\frac{(\Delta t)^2}{2}(1+c_n)\partial_{tt}u+(\Delta t)(1-c_n)\partial_tu=a_n\frac{(\Delta y)^2}{2}\nabla_y\cdot(Q\nabla_yu)\,.$$

To have the scaling to the Euler–Poisson–Darboux (EPD) equation, we take $\Delta t=\Delta y$, and

$$a_n\xrightarrow[n\to\infty]{}2\,,\qquad\qquad c_n=1-\frac{d+1}{n}+o\left(\frac{1}{n}\right)\,.\tag{5.9}$$

Indeed, as $t=n\Delta t$, we have $1-c_n\sim\frac{d+1}{t}\Delta t$ and thus

$$\frac{(\Delta t)^2}{2}(1+1+o(1))\partial_{tt}u+(\Delta t)^2\left(\frac{d+1}{t}+o(1)\right)\partial_tu=(2+o(1))\frac{(\Delta y)^2}{2}\nabla_y\cdot(Q\nabla_yu)\,,$$

thus by identifying higher-order terms,

$$\partial_{tt}u+\frac{d+1}{t}\partial_tu=\nabla_y\cdot(Q\nabla_yu)\,.$$

Note that there is the implicit condition $b_n\xrightarrow[n\to\infty]{}0$ implied by (5.9) as $a_n+b_n-c_n=1$.

Different scalings. Note that in this section, the scaling is $\Delta t=\Delta y$ while for the simple gossip, the scaling is $\Delta t=(\Delta y)^2$. This is another illustration that the iteration of this section diffuses faster: to scale to a non-degenerate object, it needs go though a more important rescaling in space.

5.1.3. Probabilistic interpretation. For the sake of mathematical curiosity, let us make an aside on the probabilistic interpretations of the heat equation and of the EPD equation. It is well-known that the heat equation represents the evolution of the probability density function of Brownian motion in $\mathbb{R}^d$. Meanwhile, Kac [1974] showed that the telegrapher's equation

$$\partial_{tt}u+2a\partial_tu=\partial_{yy}u$$

represents the evolution of a persistent random walk: $u(t,.)$ is the probability density function of a random walker in $\mathbb{R}^d$, that moves according to a fixed unit speed, and, at a Poisson rate a, resamples the direction of its speed uniformly over the unit sphere. We can confidently extrapolate this result to the case where $a=(d+1)/(2t)$ depends on time: the EPD equation (5.3) is the density of a persistent random walk that gets more and more persistent over time. The rate $a=(d+1)/(2t)$ of the speed resampling is chosen so that the law of the random walker is uniform on the ball of radius t around the initial point. Note that as the random walker has unit speed, it can not be at distance from the origin larger than the elapsed time t.

This probabilistic point of view comforts the high-level idea that acceleration is achieved by giving inertia to the gossiped information.

5.1.4. Relation to the Jacobi polynomial iteration. For the convenience of the reader, we recall here the Jacobi polynomial iteration (4.2)-(4.3) introduced in Chapter 4 to accelerate gossip algorithms:

$$x_1 = a_0\omega * x_0 + b_0 x_0\,, \qquad x_{n+1} = a_n\omega * x_n + b_n x_n - c_n x_{n-1}, \tag{5.10}$$

$$a_0 = \frac{d+4}{2(2+d)}\,, \qquad b_0 = \frac{d}{2(2+d)}\,, \tag{5.11}$$

$$a_n = \frac{(2n+d/2+1)(2n+d/2+2)}{2(n+1+d/2)^2}\,, \qquad b_n = \frac{d^2(2n+d/2+1)}{8(n+1+d/2)^2(2n+d/2)}\,, \tag{5.12}$$

$$c_n = \frac{n^2(2n+d/2+2)}{(n+1+d/2)^2(2n+d/2)}\,, \qquad n \geqslant 1\,. \tag{5.13}$$

As explained in Chapter 4, this iteration is associated to the Jacobi polynomials $\pi_n^{(\alpha,\beta)}$ with $\alpha = d/2$ and $\beta = 0$. It is of the form (5.8) with coefficients satisfying (5.9). Thus the Jacobi polynomial iteration scales to the EDP equation. However, note the large difference between the approaches of Chapters 4 and 5: in Chapter 4, we use the geometry of the graph to approximate the spectrum of the gossip problem and design a polynomial-based method adapted to this approximate spectrum; in Chapter 5, we also use the geometry of the graph but to view gossip algorithms as PDEs when rescaled. It is remarkable that the two approaches lead to similar results.

The PDE perspective enriches our understanding of the Jacobi polynomial iteration. For instance, in the spirit of Appendix 4.I.5, one can explore the effect of using the Jacobi polynomial $\pi_n^{(\alpha,\beta)}$ for a different value than $(\alpha,\beta) = (d/2, 0)$ used in the Jacobi polynomial iteration. From (4.25), it follows that in this case,

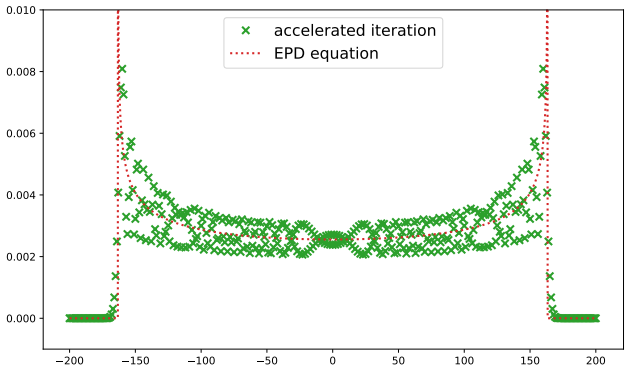
$$a_n \xrightarrow[n\to\infty]{} 2\,, \qquad c_n = 1 - \frac{2\alpha+1}{n} + o\left(\frac{1}{n}\right)\,,$$

thus, repeating the computations of Section 5.1.2, the iteration converges to the more general EPD equation (5.14). Consider its fundamental solution (5.15). If $\alpha > d/2$, the mass concentrates at the center of the ball of radius t. On the contrary, if $\alpha < d/2$, the mass concentrates at the edge of the ball. Both effects are undesirable as uniform averaging is the optimal strategy. These effects are simulated in Figure 5.3.

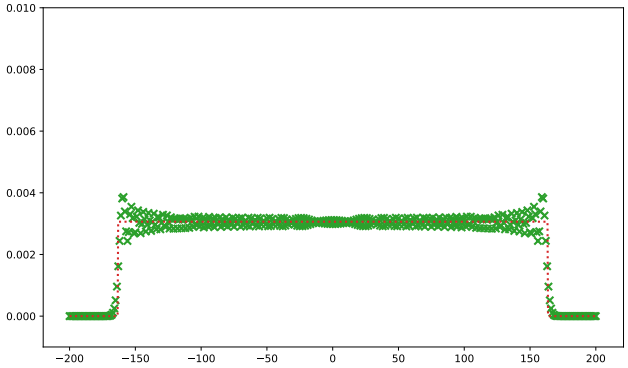
5.1.5. Open problems: other geometries, stochastic case. An important weakness of this chapter is that it only applies to synchronous gossip on a regular lattice. It is natural to ask what could happen in an asynchronous setting, or when the graph is microscopically perturbed (percolation graph, random geometric graph, etc).

For the simple gossip, or equivalently, for the random walk or for heat diffusion, answering this question is the subject of the field of homogenization, see, e.g., [Armstrong et al., 2019, Armstrong and Dario, 2018, Biskup et al., 2011]. The heuristic is that on a large scale and for long diffusion times, microscopic fluctuations of the connectivity (in space and in time) are homogenized: the process scales to an homogeneous diffusion with some constant effective diffusion matrix Q.

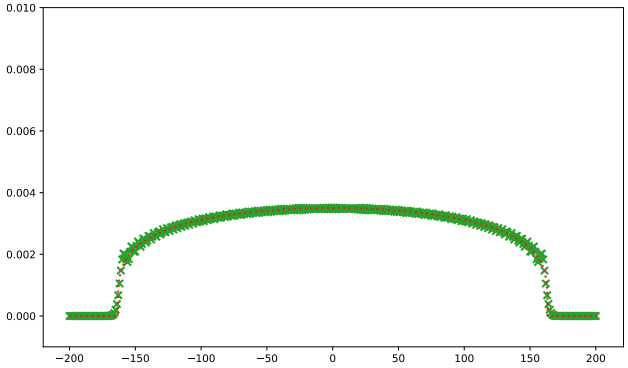
Our work raises the following question: is there homogenization for the EPD equation? Chapter 4 proves that there is some robustness of the Jacobi polynomial iteration to microscopic details of the graphs, as the rates are the same on all graphs of spectral dimension d. However, we do not know if the process scales to the same limit on those graphs.



(A) $\alpha = \frac{1}{4} < \frac{1}{2} = \frac{d}{2}$



(B) $\alpha = \frac{1}{2} = \frac{d}{2}$



(C) $\alpha = \frac{3}{4} > \frac{1}{2} = \frac{d}{2}$

FIGURE 5.3. Same simulation as in Figure 5.1(C), but we now study the effect of varying the parameter α of the Jacobi polynomial iteration. Varying α also changes the fundamental solution (5.15) of the EPD equation (5.14).

5.2. Rigorous convergence results

In this section, we give some rigorous ground to the heuristic derivations of Section 5.1. In Section 5.2.1, we start with the convergence of simple gossip to the heat equation. This case is simple as it is equivalent to the central limit theorem: we obtain a weak convergence result. A stronger convergence result, of local type, is deduced from the local central limit theorem.

Section 5.2.1 illustrates that two types of convergence are possible: weak and local. In Section 5.2.2, we prove analog results for the convergence of the Jacobi polynomial iteration to the EPD equation. We restrict to the Jacobi polynomial iteration—and not to any method satisfying (5.9)—for technical reasons: we use fine asymptotic properties of the Jacobi polynomials. However, we end this section with a remark on why we conjecture the same scaling for all iterations satisfying (5.9).

In Section 5.2.3, we apply the local convergence result to obtain convergence rates of the Jacobi polynomial iteration. These rates are sharp up to constants.

5.2.1. Simple gossip and the heat equation. Consider the simple gossip iteration

$$x_0 = \mathbb{1}_0\,, \qquad\qquad x_{n+1} = \omega * x_n\,.$$

The iteration x_n can be interpreted as the probability density function of a random walk on $\mathbb{Z}^d$, initialized from 0, with increments of density ω. As ω is centered, the random walk is unbiased; the matrix Q is the covariance of the increments. The asymptotic law x_n is described by the central limit theorems: here, we interpret them with our notations. Let $u(t, y)$ denote the fundamental solution (5.2) of the heat equation (5.1). We denote δ_y the Dirac mass at $y \in \mathbb{R}^d$.

> **Theorem 5.1** (central limit theorem, see, e.g., [Billingsley, 2008])**.** We have the following convergence in the space of positive measures: for any $t \geqslant 0$,
>
> $$\sum_{v\in\mathbb{Z}^d} x_{\lfloor t/\varepsilon^2\rfloor}(v)\delta_{\varepsilon v} \xrightarrow[\varepsilon\to 0]{} u(t, y)\mathrm{d}y\,.$$

A stronger local result holds assuming that ω is aperiodic, i.e., that the random walk with increments ω is an aperiodic Markov chain on $\mathbb{Z}^d$ [Billingsley, 2008, Section 8]. For instance, the vector ω of Equation (5.6), corresponding to the triangular lattice, is aperiodic, while the vector ω of Equation (5.5), corresponding to the regular grid, is not.

> **Theorem 5.2** (local central limit theorem, [Gnedenko, 1948])**.** Assume that ω is aperiodic. Then
>
> $$\sup_{v\in\mathbb{Z}^d} |x_n(v) - u(n, v)| = o\left(\frac{1}{n^{d/2}}\right) \qquad \text{as } n \to \infty.$$

A pedagogical introduction to the local central limit theorem is provided by Curien [2020]. The beauty of the local central limit theorem is that no rescaling is required: we simply discretize the heat equation in time and space.

5.2.2. The Jacobi polynomial iteration and the Euler–Poisson–Darboux equation. We now give analogs of Theorems 5.1 and 5.2 for the convergence of the Jacobi polynomial iteration to the heat equation. Let x_n denote the iterates of the Jacobi polynomial iteration (5.10)-(5.13) initialized from $x_0 = \mathbb{1}_0$ and $u(t, y)$ the fundamental solution (5.4) of the EPD equation (5.3).

Assumptions. In this section, we assume that ω is symmetric ($\omega(-v) = \omega(v)$) and aperiodic. While the aperiodicity assumption is clearly necessary for Theorem 5.4 to hold, we do not know if these assumptions are necessary otherwise.

Theorem 5.3 (weak convergence)**.** We have the following weak convergence in the space of signed measures: for all $t > 0$,

$$\sum_{v \in \mathbb{Z}^d} x_{\lfloor t/\varepsilon \rfloor}(v) \delta_{\varepsilon v} \xrightarrow[\varepsilon \to 0]{} u(t,y) \mathrm{d}y \,.$$

Theorem 5.4 (local convergence)**.** Denote $\psi(y) = \prod_{i=1}^d \mathrm{sinc}(y(i))$ where $y(i)$ is the i-th component of $y \in \mathbb{R}^d$ and $\mathrm{sinc}(y) = \frac{\sin(\pi y)}{\pi y}$. Then

$$\sum_{v \in \mathbb{Z}^d} \left(x_n(v) - (u(n,.) * \psi)(v) \right)^2 = o\left(\frac{1}{n^d}\right) \qquad \text{as } n \to \infty.$$

In words, the local convergence holds, provided that we convolve the solution $u(n,.)$ of the EPD equation with the filter ψ. The two theorems are proved in Appendix 5.B.

Remark 5.1. The statements of Theorems 5.3 and 5.4 and their proofs can be easily adapted to study the Jacobi polynomial iterations $x_n = \pi_n^{(\alpha,\beta)}(\omega)$ for other parameters (α, β), as long as $\alpha > d/2 - 1/2$ and $\beta \leqslant \alpha$. In this case, the limiting PDE depends on α. We have the convergence to the fundamental solution (5.15) of the general EPD equation (5.14).

Remark 5.2 (Extension beyond the Jacobi polynomial iteration)**.** Our theorems are stated for the Jacobi polynomial iteration only for a technical reason: the proofs are based on well-known asymptotic properties of the Jacobi polynomials, stated in Proposition 5.3. We conjecture that all other sequences of polynomials with recursion coefficients satisfying (5.9) also satisfy the same properties: this would prove the scaling to the EPD equation for all second-order gossip algorithms satisfying (5.9).

This conjecture is supported by Aptekarev [1993]: he shows that a sequence of orthogonal polynomial must satisfy the Mehler-Heine asymptotics (Proposition 5.3.(1)) provided that the recurrence coefficients of the polynomials satisfy some conditions that resemble (5.9). Interestingly, he explains that the asymptotics of the recurrence coefficients are related to the shape of the orthogonality measure near 1: this links the approaches of Chapter 4 and 5.

5.2.3. Application: sharp rates of the Jacobi polynomial iteration on $\mathbb{Z}^d$. In this section, we apply Theorem 5.4 to obtain sharp rates for the Jacobi polynomial iteration.

Corollary 5.1. Assume that ω is symmetric and aperiodic. Let x_n be the iterates of the Jacobi polynomial iteration (5.10)-(5.13), initialized at $x_0 = \mathbb{1}_0$. Then we have the asymptotic equivalence

$$\sum_{v \in \mathbb{Z}^d} x_n(v)^2 \underset{n \to \infty}{\sim} \frac{1}{(\det Q)^{1/2} |B(0,1)|} \frac{1}{n^d} \,,$$

where $|B(0,1)| = \frac{\pi^{d/2}}{\Gamma(d/2+1)}$ is the volume of the Euclidean unit ball in dimension d.

Compare with Theorem 4.1. Here our theorem applies only to regular lattices (not all graphs of spectral dimension d), but we obtain an asymptotic equivalent, while Theorem 4.1 gives only the exponent in n. In Figure 5.4, we compare the two asymptotic equivalent quantities in the case of the Jacobi polynomial iteration on the triangular lattice. Note that similarly, one could obtain sharp rates for simple gossip from the local central limit Theorem 5.2.

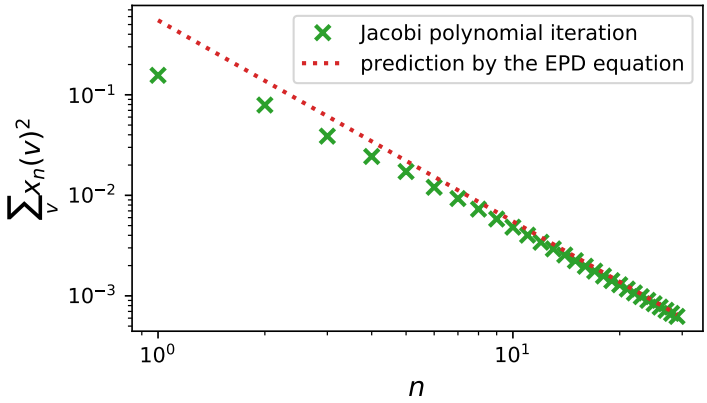


FIGURE 5.4. Comparison between the empirical error $\sum_{v\in\mathbb{Z}^d} x_n(v)^2$ and the asymptotic rate predicted by Corollary 5.1. Here, x_n are the iterates of the Jacobi polynomial iteration on the triangular lattice (5.6). Note that the corollary predicts sharply not only the scaling in n (the asymptotic slope in the logarithmic plot) but also the leading constant (the intercept of the asymptotic line).

Appendix of Chapter 5

5.A. The Euler–Poisson–Darboux (EPD) equation

The EDP equation is the partial differential equation

$$\partial_{tt} u + \frac{2\alpha+1}{t} \partial_t u = \nabla_y \cdot (Q \nabla_y u) \ . \tag{5.14}$$

Posing a rigorous framework for solving this equation is subtle because there is a diverging coefficient $\frac{2\alpha+1}{t}$ as $t \to 0$. Moreover, we see below that fundamental solutions are irregular; they are defined in a weak sense. Thankfully, we do not have to bother with these technical details as our rigorous results only require to know the expression of the fundamental solution of the EPD equation (Proposition 5.1) and its Fourier transform (Proposition 5.2). These expressions are given by Bresters [1973] in the case $Q = \mathrm{Id}$; here, we easily extend the expressions for a general matrix Q.

Proposition 5.1. The fundamental solution of the EPD equation, i.e., the solution initialized from $u(0,.) = \delta_0$, $\partial_t u(0,.) = 0$, is

$$u(t,y) = \frac{\Gamma(\alpha+1)}{\pi^{d/2}\Gamma(\alpha+1-d/2)(\det Q)^{1/2}} \frac{1}{t^{2\alpha}} \left(t^2 - \left\langle y, Q^{-1} y \right\rangle\right)_+^{\alpha-d/2} \ , \tag{5.15}$$

where $(.)_+$ denotes the positive part of a real number.

The case $\alpha = d/2$ is particularly important to this chapter; in this case we recover (5.4) from (5.15).

Proof of Proposition 5.1. In the case $Q = \mathrm{Id}$, the solution is given by Bresters [1973, Equation (3.4)]:

$$u(t,y) = \frac{\Gamma(\alpha+1)}{\pi^{d/2}\Gamma(\alpha+1-d/2)} \frac{1}{t^{2\alpha}} \left(t^2 - \|y\|^2\right)_+^{\alpha-d/2} \ .$$

In the general case, consider $v(t,y) = u(t, Q^{1/2} y)(\det Q)^{1/2}$. Computations give that $v(t,y)$ is the fundamental solution of the EPD equation (5.14) with $Q = \mathrm{Id}$, thus

$$u(t, Q^{1/2} y)(\det Q)^{1/2} = v(t,y) = \frac{\Gamma(\alpha+1)}{\pi^{d/2}\Gamma(\alpha+1-d/2)} \frac{1}{t^{2\alpha}} \left(t^2 - \|y\|^2\right)_+^{\alpha-d/2} \ .$$

This gives the desired formula. □

Proposition 5.2. The Fourier transform in space of the fundamental solution (5.15) is

$$\widehat{u}(t,\xi) = \int_{\mathbb{R}^d} \mathrm{d}y \, e^{i\langle \xi, y\rangle} u(t,y) = 2^\alpha \Gamma(\alpha+1) \left\langle \xi, Q\xi \right\rangle^{-\alpha/2} t^{-\alpha} J_\alpha \left(t \left\langle \xi, Q\xi \right\rangle^{1/2} \right) \ ,$$

where J_α denotes the Bessel function of the first kind of order α [Szegö, 1939, Section 1.71].

Proof. In the case $Q = \mathrm{Id}$, the result is given by Bresters [1973, Equation (3.1)]:

$$\widehat{u}(t,\xi) = 2^\alpha \Gamma(\alpha+1) \|\xi\|^{-\alpha} t^{-\alpha} J_\alpha \left(t\|\xi\|\right) \ .$$

In the general case, $v(t,y) = u(t, Q^{1/2}y)(\det Q)^{1/2}$ is a solution of the EPD equation (5.14) with $Q = \mathrm{Id}$. Moreover,

$$\begin{aligned}\widehat{v}(t,\xi) &= \int_{\mathbb{R}^d} \mathrm{d}y\, e^{i\langle \xi, y\rangle} v(t,y) \\ &= (\det Q)^{1/2} \int_{\mathbb{R}^d} \mathrm{d}y\, e^{i\langle \xi, y\rangle} u(t, Q^{1/2}y)\,.\end{aligned}$$

In the last integral, we change the variable to $x = Q^{1/2}y$. Then $\mathrm{d}x = \det(Q^{1/2})\mathrm{d}y = (\det Q)^{1/2}\mathrm{d}y$.

$$\begin{aligned}\widehat{v}(t,\xi) &= \int_{\mathbb{R}^d} \mathrm{d}x\, e^{i\langle \xi, Q^{-1/2}x\rangle} u(t,x) \\ &= \widehat{u}(t, Q^{-1/2}\xi)\,.\end{aligned}$$

Thus

$$\widehat{u}(t,\xi) = \widehat{v}(t, Q^{1/2}\xi) = 2^\alpha \Gamma(\alpha+1)\, \langle \xi, Q\xi\rangle^{-\alpha/2}\, t^{-\alpha} J_\alpha\left(t\, \langle \xi, Q\xi\rangle^{1/2}\right)\,.$$

□

5.B. Proof of Theorems 5.3 and 5.4

Our proofs use the asymptotic properties of the Jacobi polynomials; we rewrite the iteration (5.10)-(5.13), initialized from $x_0 = \mathbb{1}_0$, as $x_n = \pi_n^{(d/2,0)}(\omega)$. Here, we use the notations of Chapter 4: $\pi_n^{(d/2,0)}$ denotes the Jacobi polynomial, rescaled such that $\pi_n^{(d/2,0)}(1) = 1$, with parameters $\alpha = d/2$, $\beta = 0$. The evaluation of the polynomial $\pi_n^{(d/2,0)}$ in ω is done by taking the convolution as the product.

The proofs below use the following well-known results on Jacobi polynomials.

Proposition 5.3. (1) (Mehler-Heine asymptotic) The Jacobi polynomials satisfy the following asymptotic at the edge of the orthogonality measure

$$\lim_{n\to\infty} \pi_n^{(d/2,0)}\left(1 - \frac{z^2}{2n^2}\right) = 2^{d/2}\Gamma\left(\frac{d}{2}+1\right) z^{-d/2} J_{d/2}(z)\,,$$

where $J_{d/2}$ denotes the Bessel function of the first kind of order $d/2$ [Szegö, 1939, Section 1.71]. The convergence is uniform for z in compact sets.

(2) On the whole support of the orthogonality measure, we have the following bounds: there exists constants $C_1, C_2 > 0$ such that for all $n \geqslant 0$,

$$\left|\pi_n^{(d/2,0)}(\lambda)\right| \leqslant \begin{cases} C_1\, (\arccos|\lambda|)^{-d/2-1/2}\, n^{-d/2-1/2} & \text{if } |\lambda| \leqslant 1 - \frac{1}{n^2}\,, \\ C_2 & \text{otherwise.}\end{cases}$$

Proof. (1) Szegö [1939, Theorem 8.1.1] gives the Mehler-Heine asymptotic for the classical Jacobi polynomials $P_n^{(d/2,0)}$:

$$\lim_{n\to\infty} n^{-d/2} P_n^{(d/2,0)}\left(1 - \frac{z^2}{2n^2}\right) = 2^{d/2} z^{-d/2} J_{d/2}(z)\,,$$

with uniform convergence for z in compact sets. As

$$\pi_n^{(d/2,0)} = \frac{P_n^{(d/2,0)}}{P_n^{(d/2,0)}(1)} = \frac{P_n^{(d/2,0)}}{\binom{n+d/2}{n}}\,, \qquad \binom{n+d/2}{n} \sim \frac{n^{d/2}}{\Gamma\left(\frac{d}{2}+1\right)}\,,$$

we obtain the desired formula.

(2) For $\lambda \geqslant 0$, this is only a reformulation of Proposition 4.15 (that follows straightforwardly from [Szegö, 1939, Theorem 7.32.2]). For $\lambda < 0$, we use the symmetry of the Jacobi polynomials $P_n^{(d/2,0)}(\lambda) = (-1)^n P_n^{(0,d/2)}$ [Szegö, 1939, Equation (4.1.3)] and use again [Szegö, 1939, Theorem 7.32.2].

□

Proof of Theorem 5.3. Denote

$$\mu_{t,\varepsilon} = \sum_{v\in\mathbb{Z}^d} x_{\lfloor t/\varepsilon\rfloor}(v)\delta_{\varepsilon v}\,, \qquad\qquad \mu_t = u(t,y)\mathrm{d}y$$

The proof is based on [Baez-Duarte, 1993, Theorem 2.1], a variant of Lévy's theorem for signed measures: in order to prove the weak convergence $\mu_{t,\varepsilon} \to \mu_t$ as $\varepsilon \to 0$, it is sufficient to check that the family of measures $\mu_{t,\varepsilon}$, $\varepsilon > 0$ is tight, bounded in total variation, that we have the pointwise convergence of the Fourier transform $\widehat{\mu}_{t,\varepsilon} \to \widehat{\mu}_t$ almost everywhere. These three conditions are checked below.

Tightness of $\mu_{t,\varepsilon}$, $\varepsilon > 0$. ω has a finite support, thus there exists $R > 0$ such that the support of ω is included in $B(0,R)$. Then for all $n \geqslant 0$, the support of $w^{*n} = w * \cdots * w$ (with n terms) is included in $B(0,nR)$. The vector x_n is a linear combination of the w^{*l} for $l \leqslant n$, thus is also included in $B(0,nR)$. Finally, when rescaling by ε, the support of $\mu_{t,\varepsilon} = \sum_{v\in\mathbb{Z}^d} x_{\lfloor t/\varepsilon\rfloor}(v)\delta_{\varepsilon v}$ is included in $B(0,\varepsilon\lfloor t/\varepsilon\rfloor R) \subset B(0,tR)$. The latter set is independent of ε, thus the family of measures $\mu_{t,\varepsilon}$, $\varepsilon > 0$ is tight.

Boundedness of $\mu_{t,\varepsilon}$, $\varepsilon > 0$. Note that $\mu_{t,\varepsilon}(\mathbb{R}^d) = 1$, but as $\mu_{t,\varepsilon}$ is a signed measure, we need to show that the total mass $\|\mu_{t,\varepsilon}\| = |\mu_{t,\varepsilon}|(\mathbb{R}^d)$ of the total variation $|\mu_{t,\varepsilon}|$ is bounded independently of ε. By Hölder's inequality,

$$\|\mu_{t,\varepsilon}\| = \|x_{\lfloor t/\varepsilon\rfloor}\|_{l^1(\mathbb{Z}^d)} \leqslant |\operatorname{Supp} x_{\lfloor t/\varepsilon\rfloor}|^{1/2}\|x_{\lfloor t/\varepsilon\rfloor}\|^{1/2}_{l^2(\mathbb{Z}^d)}\,, \tag{5.16}$$

where $|\operatorname{Supp} x_{\lfloor t/\varepsilon\rfloor}|$ denotes the cardinal of the support of $x_{\lfloor t/\varepsilon\rfloor}$. As this support is included in $B(0,\lfloor t/\varepsilon\rfloor R)$, its cardinal can be bounded by the number of integer points in $B(0,\lfloor t/\varepsilon\rfloor R)$. This is dominated by ε^{-d} as $\varepsilon \to 0$. Thus

$$|\operatorname{Supp} x_{\lfloor t/\varepsilon\rfloor}| = O(\varepsilon^{-d})\,.$$

We now bound the second term in (5.16), namely the norm $\|x_{\lfloor t/\varepsilon\rfloor}\|_{l^2(\mathbb{Z}^d)}$. By Plancherel identity,

$$\|x_n\|^2_{\ell^2(\mathbb{Z}^d)} = \left\|\pi_n^{(d/2,0)}(\omega)\right\|^2_{\ell^2(\mathbb{Z}^d)} = \frac{1}{(2\pi)^d}\left\|\widehat{\pi_n^{(d/2,0)}(\omega)}\right\|^2_{L^2([-\pi,\pi]^d)}\,.$$

We now use that the Fourier transform of a convolution is the product of the Fourier transforms, thus $\widehat{\pi_n^{(d/2,0)}(\omega)} = \pi_n^{(d/2,0)}(\widehat{\omega})$. We obtain

$$\|x_n\|^2_{\ell^2(\mathbb{Z}^d)} = \frac{1}{(2\pi)^d}\left\|\pi_n^{(d/2,0)}(\widehat{\omega})\right\|^2_{L^2([-\pi,\pi]^d)} = \frac{1}{(2\pi)^d}\int_{[-\pi,\pi]^d}\mathrm{d}\xi\left|\pi_n^{(d/2,0)}(\widehat{\omega}(\xi))\right|^2\,.$$

Here, as ω is symmetric, $\widehat{\omega}(\xi)$ is real. We can use the bounds of Proposition 5.3.(2). We need to estimate $\widehat{\omega}(\xi)$. We use the following lemma.

Lemma 5.1. As ω is aperiodic, there exists $\lambda > 0$ such that

$$|\widehat{\omega}(\xi)| \leqslant 1 - \lambda\|\xi\|^2\,, \qquad \xi \in [-\pi,\pi]^d\,.$$

This lemma is simple and proved by Curien [2020, Section 7.1]. We now return to our estimate of $\left|\pi_n^{(d/2,0)}(\widehat{\omega}(\xi))\right|^2$.

- If $\|\xi\| \geqslant \frac{1}{\sqrt{\lambda}n}$, we have $|\widehat{\omega}(\xi)| \leqslant 1 - \lambda\|\xi\|^2 \leqslant 1 - \frac{1}{n^2}$. Thus by Proposition 5.3.(2),
$$\left|\pi_n^{(d/2,0)}(\widehat{\omega}(\xi))\right| \leqslant C_1 \left(\arccos|\widehat{\omega}(\xi)|\right)^{-d/2-1/2} n^{-d/2-1/2}$$
$$\leqslant C_1 \left(\arccos\left(1-\lambda\|\xi\|^2\right)\right)^{-d/2-1/2} n^{-d/2-1/2}.$$
- If $\|\xi\| < \frac{1}{\sqrt{\lambda}n}$, we can only say $\left|\pi_n^{(d/2,0)}(\widehat{\omega}(\xi))\right| \leqslant C_2$.

Thus
$$\|x_n\|^2_{\ell^2(\mathbb{Z}^d)} = \frac{1}{(2\pi)^d}\int_{[-\pi,\pi]^d} \mathrm{d}\xi \left|\pi_n^{(d/2,0)}(\widehat{\omega}(\xi))\right|^2$$
$$\leqslant C_3 n^{-d-1}\int_{\{\|\xi\|\geqslant 1/(\sqrt{\lambda}n)\}} \mathrm{d}\xi \left(\arccos\left(1-\lambda\|\xi\|^2\right)\right)^{-d-1} + C_4\int_{\{\|\xi\|<1/(\sqrt{\lambda}n)\}}\mathrm{d}\xi$$
where we use the notation C_i to denote constants independent of n. We use a spherical change of variables in the first integral:
$$\|x_n\|^2_{\ell^2(\mathbb{Z}^d)} \leqslant C_5 n^{-d-1}\int_{1/(\sqrt{\lambda}n)}^{\sqrt{d}\pi} \mathrm{d}r\, r^{d-1}\left(\arccos\left(1-\lambda r^2\right)\right)^{-d-1} + C_6 n^{-d}. \tag{5.17}$$
As $r\to 0$, $\arccos(1-\lambda r^2) \sim \sqrt{2\lambda}r$ and thus
$$r^{d-1}\left(\arccos\left(1-\lambda r^2\right)\right)^{-d-1} \sim \sqrt{2}\lambda^{-d/2-1/2} r^{-2}.$$
Thus $r^{d-1}\left(\arccos\left(1-\lambda r^2\right)\right)^{-d-1}$ is not integrable at 0. We thus have, as $n\to\infty$,
$$\int_{1/(\sqrt{\lambda}n)}^{\sqrt{d}\pi} \mathrm{d}r\, r^{d-1}\left(\arccos\left(1-\lambda r^2\right)\right)^{-d-1} \sim C_7\int_{1/(\sqrt{\lambda}n)}^{\sqrt{d}\pi}\mathrm{d}r\, r^{-2} \sim C_8 n.$$
Putting back in (5.17), we obtain $\|x_n\|^2_{\ell^2(\mathbb{Z}^d)} = O(n^{-d})$. Finally, getting back to (5.16), we obtain as $\varepsilon\to 0$
$$\|\mu_{t,\varepsilon}\| \leqslant |\operatorname{Supp} x_{\lfloor t/\varepsilon\rfloor}|^{1/2}\|x_{\lfloor t/\varepsilon\rfloor}\|^{1/2}_{l^2(\mathbb{Z}^d)} = O(\varepsilon^{-d/2})O\left(\left\lfloor\frac{t}{\varepsilon}\right\rfloor^{-d/2}\right) = O(1).$$
This shows that the family of measures $\mu_{t,\varepsilon}$, $\varepsilon>0$ is bounded in total variation.

Pointwise convergence of the Fourier transform.
$$\widehat{\mu}_{t,\varepsilon}(\xi) = \int_{\mathbb{R}^d}\mathrm{d}\mu_{t,\varepsilon}(y)e^{i\langle\xi,y\rangle} = \sum_{v\in\mathbb{Z}^d} x_{\lfloor t/\varepsilon\rfloor}(v)e^{i\langle\xi,\varepsilon v\rangle} = \widehat{x}_{\lfloor t/\varepsilon\rfloor}(\varepsilon\xi) = \pi^{(d/2,0)}_{\lfloor t/\varepsilon\rfloor}\left(\widehat{\omega}(\varepsilon\xi)\right).$$
As $\varepsilon\to 0$,
$$\widehat{\omega}(\varepsilon\xi) = 1 - \frac{\varepsilon^2}{2}\langle\xi, Q\xi\rangle + o(\varepsilon^2).$$
We now apply Proposition 5.3.(1):
$$\widehat{\mu}_{t,\varepsilon}(\xi) = \pi^{(d/2,0)}_{\lfloor t/\varepsilon\rfloor}\left(1-\frac{\varepsilon^2}{2}\left(\langle\xi,Q\xi\rangle + o(1)\right)\right) = \pi^{(d/2,0)}_{\lfloor t/\varepsilon\rfloor}\left(1-\frac{t^2\langle\xi,Q\xi\rangle + o(1)}{2\lfloor t/\varepsilon\rfloor^2}\right)$$
$$\xrightarrow[\varepsilon\to 0]{} 2^{d/2}\Gamma\left(\frac{d}{2}+1\right)\langle\xi,Q\xi\rangle^{-d/4}t^{-d/2}J_{d/2}\left(t\langle\xi,Q\xi\rangle^{1/2}\right) = \widehat{u}(t,\xi) = \widehat{\mu}_t(\xi),$$
where we used the formula for $\widehat{u}(t,\xi)$ from Proposition 5.2. This finishes the proof of the weak convergence. □

Proof of Theorem 5.4. This proof is similar to the one of Theorem 5.3. By Plancherel's formula,

$$\sum_{v\in\mathbb{Z}^d}\left(x_n(v)-\left(u(n,.)*\psi\right)(v)\right)^2=\|x_n-u(n,.)*\psi\|^2_{\ell^2(\mathbb{Z}^d)}=\left\|\widehat{x}_n-\widehat{u(n,.)*\psi}\right\|^2_{L^2([-\pi,\pi]^d)} \tag{5.18}$$

In this last expression, we take the Fourier transform of $u(n,.)*\psi$ as a function of $v\in\mathbb{Z}^d$. However, to decompose the computation, let us first compute the Fourier transform of $y\in\mathbb{R}^d\mapsto(u(n,.)*\psi)(y)$. The Fourier transform of this convolution is the product of the Fourier transforms, and ψ is chosen specifically so that its Fourier transform is $\widehat{\psi}(\xi)=\mathbb{1}_{\{\xi\in[-\pi,\pi]^d\}}$. As a consequence, the Fourier transform of $y\in\mathbb{R}^d\mapsto(u(n,.)*\psi)(y)$ is $\xi\in\mathbb{R}^d\mapsto\widehat{u}(t,\xi)\mathbb{1}_{\{\xi\in[-\pi,\pi]^d\}}$.

We now discretize this function and seek the Fourier transform of $v\in\mathbb{Z}^d\mapsto(u(n,.)*\psi)(v)$. The Fourier transform of the discretization is the periodization of the Fourier transform, thus the Fourier transform of $v\in\mathbb{Z}^d\mapsto(u(n,.)*\psi)(v)$ is $\xi\in[-\pi,\pi]^d\mapsto\widehat{u}(n,\xi)$.

We obtain

$$\left\|\widehat{x}_n-\widehat{u(n,.)*\psi}\right\|^2_{L^2([-\pi,\pi]^d)}=\int_{[-\pi,\pi]^d}\mathrm{d}\xi\left(\pi_n^{(d/2,0)}\left(\widehat{\omega}(\xi)\right)-\widehat{u}(n,\xi)\right)^2.$$

We make the change of variables $\zeta=n\xi$:

$$\left\|\widehat{x}_n-\widehat{u(n,.)*\psi}\right\|^2_{L^2([-\pi,\pi]^d)}=n^{-d}\int_{\mathbb{R}^d}\mathrm{d}\zeta\left(\pi_n^{(d/2,0)}\left(\widehat{\omega}\left(\frac{\zeta}{n}\right)\right)-\widehat{u}\left(n,\frac{\zeta}{n}\right)\right)^2\mathbb{1}_{\{\zeta\in[-n\pi,n\pi]^d\}}. \tag{5.19}$$

Fix $\zeta\in\mathbb{R}^d$. Using the Mehler-Heine asymptotic (Proposition 5.3.(1)), we prove that

$$\pi_n^{(d/2,0)}\left(\widehat{\omega}\left(\frac{\zeta}{n}\right)\right)-\widehat{u}\left(n,\frac{\zeta}{n}\right)\xrightarrow[n\to\infty]{}0.$$

We do not repeat the computations because they are similar to the pointwise convergence in the proof of Theorem 5.3. This proves that the integrand of (5.19) converges pointwise to 0. We want to apply the dominated convergence theorem to conclude, and thus seek a domination of

$$\left(\pi_n^{(d/2,0)}\left(\widehat{\omega}\left(\frac{\zeta}{n}\right)\right)-\widehat{u}\left(n,\frac{\zeta}{n}\right)\right)^2\mathbb{1}_{\{\zeta\in[-n\pi,n\pi]^d\}} \tag{5.20}$$

$$\leqslant 2\pi_n^{(d/2,0)}\left(\widehat{\omega}\left(\frac{\zeta}{n}\right)\right)^2\mathbb{1}_{\{\zeta\in[-n\pi,n\pi]^d\}}+2\widehat{u}\left(n,\frac{\zeta}{n}\right)^2. \tag{5.21}$$

By scale invariance of the EPD equation (or, more simply, from Proposition 5.2), $\widehat{u}\left(n,\frac{\zeta}{n}\right)=\widehat{u}\left(1,\zeta\right)$. Further, by Plancherel's theorem,

$$\int_{\mathbb{R}^d}\mathrm{d}\zeta\,\widehat{u}(1,\zeta)^2=(2\pi)^d\int_{\mathbb{R}^d}\mathrm{d}y\,u(1,y)^2<\infty,$$

thus the second term of (5.21) is independent of n and integrable. We now need to find a domination for the first term. Here, the reasoning is similar to the boundedness of $\mu_{t,\varepsilon}$ in the proof of Theorem 5.3.

- If $\|\zeta\|\geqslant\frac{1}{\sqrt{\lambda}}$, by Lemma 5.1, $\left|\widehat{\omega}\left(\frac{\zeta}{n}\right)\right|\leqslant 1-\frac{1}{n^2}$, thus by Proposition 5.3.(2),

$$\begin{aligned}\pi_n^{(d/2,0)}\left(\widehat{\omega}\left(\frac{\zeta}{n}\right)\right)^2&\leqslant C_1^2\left(\arccos\left|\widehat{\omega}\left(\frac{\zeta}{n}\right)\right|\right)^{-d-1}n^{-d-1}\\&\leqslant C_1^2\left(\arccos\left(1-\lambda\frac{\|\zeta\|^2}{n^2}\right)\right)^{-d-1}n^{-d-1}.\end{aligned}$$

There exists $C_9 > 0$ such that $\arccos(1-z) \geqslant C_9\sqrt{z}$. Thus

$$\pi_n^{(d/2,0)}\left(\widehat{\omega}\left(\frac{\zeta}{n}\right)\right)^2 \leqslant C_{10}\|\zeta\|^{-d-1}\,.$$

- If $\|\zeta\| < \frac{1}{\sqrt{\lambda}}$, then

$$\pi_n^{(d/2,0)}\left(\widehat{\omega}\left(\frac{\zeta}{n}\right)\right)^2 \leqslant C_2^2\,.$$

We thus define the domination

$$g(\zeta) = \begin{cases} C_{10}\|\zeta\|^{-d-1} & \text{if } \|\zeta\| \geqslant \frac{1}{\sqrt{\lambda}}, \\ C_2^2 & \text{if } \|\zeta\| < \frac{1}{\sqrt{\lambda}}. \end{cases}$$

This domination is integrable on $\mathbb{R}^d$; this concludes the theorem.

□

5.C. Proof of Corollary 5.1

Note that $\sum_{v\in\mathbb{Z}^d} x_n(v)^2 = \|x_n\|^2_{l^2(\mathbb{Z}^d)}$ and by Theorem 5.4,

$$\Big|\|x_n\|_{l^2(\mathbb{Z}^d)} - \|u(n,.)*\psi\|_{l^2(\mathbb{Z}^d)}\Big| \leqslant \|x_n - u(n,.)*\psi\|_{l^2(\mathbb{Z}^d)} = o(n^{-d/2})\,.$$

It is thus sufficient to prove that

$$\|u(n,.)*\psi\|^2_{l^2(\mathbb{Z}^d)} \sim \frac{1}{(\det Q)^{1/2}|B(0,1)|}\frac{1}{n^d}\,.$$

In the proof of Theorem 5.4, we explain that the Fourier transform of $v \in \mathbb{Z}^d \mapsto (u(n,.)*\psi)(v)$ is $\xi \in [-\pi,\pi]^d \mapsto \widehat{u}(n,\xi)$. Thus by Plancherel's theorem,

$$\|u(n,.)*\psi\|^2_{l^2(\mathbb{Z}^d)} = \frac{1}{(2\pi)^d}\,\|\widehat{u}(n,.)\|^2_{L^2([-\pi,\pi]^d)} = \frac{1}{(2\pi)^d}\int_{[-\pi,\pi]^d} \mathrm{d}\xi\,\widehat{u}(n,\xi)^2\,.$$

We make the change of variables $\zeta = \xi/n$. Note that by scale invariance of the EPD equation (or, more simply, from Proposition 5.2), $\widehat{u}\left(n,\frac{\zeta}{n}\right) = \widehat{u}\,(1,\zeta)$. Thus

$$\|u(n,.)*\psi\|^2_{l^2(\mathbb{Z}^d)} = \frac{1}{(2\pi)^d n^d}\int_{[-n\pi,n\pi]^d} \mathrm{d}\zeta\,\widehat{u}(1,\zeta)^2 = \frac{1}{(2\pi)^d n^d}\left(\int_{\mathbb{R}^d} \mathrm{d}\zeta\,\widehat{u}(1,\zeta)^2 + o(1)\right)\,. \tag{5.22}$$

We use again Plancherel's theorem and then (5.4):

$$\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \mathrm{d}\zeta\,\widehat{u}(1,\zeta)^2 = \int_{\mathbb{R}^d} \mathrm{d}y\,u(1,y)^2 = \left(\frac{\Gamma(d/2+1)}{\pi^{d/2}(\det Q)^{1/2}}\right)^2\left|\left\{y\middle|\left\langle y, Q^{-1}y\right\rangle \leqslant 1\right\}\right|\,.$$

As the volume of the d-dimensional unit ball is $|B(0,1)| = \frac{\pi^{d/2}}{\Gamma(d/2+1)}$, the volume of the ellipsoid $|\{y|\langle y, Q^{-1}y\rangle \leqslant 1\}|$ is $\frac{\pi^{d/2}(\det Q)^{1/2}}{\Gamma(d/2+1)}$, thus

$$\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \mathrm{d}\zeta\,\widehat{u}(1,\zeta)^2 = \frac{\Gamma(d/2+1)}{\pi^{d/2}(\det Q)^{1/2}} = \frac{1}{(\det Q)^{1/2}|B(0,1)|}\,.$$

Substituting in (5.22), this concludes the proof.

Conclusion and Research Directions

We have learned to think of optimization and gossip algorithms in parallel. We believe that both fields benefit from the perspective of the other one. As a younger field, gossip algorithms benefit from the transposition of ideas from optimization. Conversely, diffusions and more generally gossip algorithms are visual and thus can be intuited more easily (Chapter 5). Moreover, the continuous time of gossip algorithm brought an elegant perspective on Nesterov acceleration (Chapter 3).

We pushed the parallel in various settings: stochastic or deterministic, parametric or non-parametric. We tried to make the equivalence rigorous, i.e., to prove results simultaneously in both fields at once (Chapter 2). The difference between continuous and discrete time became an annoying limitation to the parallel. However, it revealed to be not a technical detail but symptomatic of the challenge caused by asynchrony in gossip algorithm: this reflection led to Chapter 3.

Overall, we hope that this parallel can be used to inspire and simplify research in these fields. A natural playground could be distributed optimization, as it mixes single-machine optimization and gossip algorithms [Assran et al., 2020, Nedic et al., 2010]. Seeing distributed optimization as classical optimization on a lifted function inspires algorithms and simplifies proofs, see [Scaman et al., 2017] for instance. It remains to see how far this technique can be pushed.

In this thesis, we focused on two specific assumptions that we believe to be particularly relevant and that we want to encourage.

First, we often assumed that our stochastic gradients are noiseless, meaning without any additive noise and only with pure multiplicative noise. This assumptions is made more and more often as it occurs when minimizing the training error of overparameterized models: there exists a predictor that perfectly fits the training data [Bartlett et al., 2021]. However, it is also possible that there exists a predictor that achieves a zero generalization risk, not only a zero training error. This is the case, for instance, in the cat-vs-dog problem where humans achieve almost zero error; this can be seen as a function interpolation problem from the noiseless observation of its values at random points (see Section 1.4.3 for a lengthy discussion). In this case, there is no need for the model to be overparameterized to have noiseless gradients. This motivation for noiseless gradients is reasonable, but rarely expressed (see [Wojtowytsch, 2021] for an exception). We also gave a more unusual reason to make the noiseless assumption: it is natural for the application to gossip algorithms.

Second, we gave a special focus on the non-parametric theory; in our case, this corresponds to source and capacity conditions, or a spectral dimension assumption. These problem description are more suited to the large-scale problems that appear in modern computer science: large dimension, large networks. Of course, non-parametric statistics have existed for a long time, but the spectral dimension was new to gossip algorithm. We encourage to use more this description for large networks.

Under both assumptions, even the convergence rates of stochastic gradient descent were unknown (Chapter 2). Thus many questions are open, notably starting with the question of acceleration in the same setting. We provide only a partial answer in Section 3.4.2, this is restricted to the regularity $\alpha = 1$, and we leave the question open for other regularities.

More generally, among the numerous questions that could be asked, let us identify two motivating directions.

Statistical optimality of first-order methods. In computer science, a large number of works seek an algorithm with low complexity that achieves the information theoretic bound, i.e., the performance of the best of all algorithms, including the most time-consuming ones. If this is possible, we say that the algorithm with low complexity is optimal. Often, complexity is taken into account in a crude way, for instance by restricting to polynomial-time algorithms. However, because of the large scale of modern statistical problems, a more realistic requirement is to restrict to first-order methods, with a single-pass on the data (or few passes) [Bottou and Le Cun, 2005]. Thus: can single-pass first order methods achieve optimality?

There exists numerous works studying this question; for instance Dieuleveut and Bach [2016] show that averaged regularized stochastic gradient descent can achieve optimality in the noisy case. However, Chapter 2 shows that stochastic gradient descent is suboptimal for function interpolation. This raises the question of accelerating stochastic gradient descent up to optimality in the noiseless case. We warn that the optimal rates in the noiseless case should be faster that in the noisy case; achieving optimality could thus be more demanding.

Going non-linear. This thesis has largely focused on linear iterations (in the initialization), that correspond to least-squares problems or quadratic objectives (with the exception of the first half of Chapter 3). This enabled a heavy use of the covariance operator, or Hessian, and its diagonalization. To bridge the gap with practice, we need to study the case of non-quadratic losses, general convex objectives, or even some non-convex objectives. However, we must also warn that the world out there is vast and wild. Statistical learning with non-convex neural networks is notoriously hard to analyze [Bartlett et al., 2021].

On the gossip side, we can also replace the local averaging at the meeting of two agents by a potentially non-linear update rule. As reviewed by Aldous [2013], one can obtain a wide range of models: pandemic processes, voter models, token processes, etc. In each case, the question of the long-term behavior can be asked; however it gets much harder to state results in a general setting. Ad hoc results are built for specific geometries and specific update rules.

Similarly, we expect the exploration of statistical learning for non-convex objectives to progress through small steps. The enthusiastic research will find a limitless source of excitement in exploring this wide bestiary of research problems.

Bibliography

D. Aldous. Interacting particle systems as stochastic social dynamics. *Bernoulli*, 19(4):1122–1149, 2013.

D. Aldous and J. A. Fill. Reversible markov chains and random walks on graphs, 2002. Unfinished monograph, recompiled 2014, available at `http://www.stat.berkeley.edu/$\sim$aldous/RWG/book.html`.

D. Aldous and D. Lanoue. A lecture on the averaging process. *Probability Surveys*, 9:90–102, 2012.

Z. Allen-Zhu and L. Orecchia. Linear Coupling: An Ultimate Unification of Gradient and Mirror Descent. In *Proceedings of the 8th Innovations in Theoretical Computer Science*, ITCS '17, 2017.

S. Andres, M. T. Barlow, J.-D. Deuschel, and B. M. Hambly. Invariance principle for the random conductance model. *Probability Theory and Related Fields*, 156(3-4):535–580, 2013.

A. I. Aptekarev. Asymptotics of orthogonal polynomials in a neighborhood of the endpoints of the interval of orthogonality. *Sbornik: Mathematics*, 76(1):35, 1993.

M. Arioli and J. Scott. Chebyshev acceleration of iterative refinement. *Numerical Algorithms*, 66 (3):591–608, 2014.

Y. Arjevani, S. Shalev-Shwartz, and O. Shamir. On lower and upper bounds in smooth and strongly convex optimization. *Journal of Machine Learning Research*, 17(126):1–51, 2016.

S. Armstrong and P. Dario. Elliptic regularity and quantitative homogenization on percolation clusters. *Communications on Pure and Applied Mathematics*, 71(9):1717–1849, 2018.

S. Armstrong, T. Kuusi, and J.-C. Mourrat. *Quantitative stochastic homogenization and large-scale regularity*, volume 352. Springer, 2019.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

M. Assran, A. Aytekin, H. R. Feyzmahdavian, M. Johansson, and M. G. Rabbat. Advances in asynchronous parallel and distributed optimization. *Proceedings of the IEEE*, 108(11):2013–2031, 2020.

H. Attouch, Z. Chbani, J. Peypouquet, and P. Redont. Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity. *Mathematical Programming*, 168(1):123–175, 2018.

H. Attouch, Z. Chbani, and H. Riahi. Rate of convergence of the Nesterov accelerated gradient method in the subcritical case $\alpha \leqslant 3$. *ESAIM: Control, Optimisation and Calculus of Variations*, 25:2, 2019.

K. Avrachenkov, L. Cottatellucci, and M. Hamidouche. Eigenvalues and spectral dimension of random geometric graphs in thermodynamic regime. In *International Conference on Complex Networks and Their Applications*, pages 965–975. Springer, 2019.

O. Axelsson. *Iterative solution methods*. Cambridge university press, 1996.

N. S. Aybat, A. Fallah, M. Gurbuzbalaban, and A. Ozdaglar. Robust accelerated gradient methods for smooth strongly convex functions. *SIAM Journal on Optimization*, 30(1):717–751, 2020.

F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Advances in Neural Information Processing Systems*, pages 773–781, 2013.

L. Baez-Duarte. Central limit theorem for complex measures. *Journal of Theoretical Probability*, 6 (1):33–56, 1993.

P. Barooah and J. P. Hespanha. Estimation from relative measurements: Electrical analogy and large graphs. *IEEE Transactions on Signal Processing*, 56(6):2181–2193, 2008.

P. Bartlett, A. Montanari, and A. Rakhlin. Deep learning: a statistical viewpoint. *arXiv preprint arXiv:2103.09177*, 2021.

B. Bauer, L. Devroye, M. Kohler, A. Krzyżak, and H. Walk. Nonparametric estimation of a function from noiseless observations at random points. *Journal of Multivariate Analysis*, 160:93–104, 2017.

L. Becchetti, A. Clementi, P. Manurangsi, E. Natale, F. Pasquale, P. Raghavendra, and L. Trevisan. Average whenever you meet: Opportunistic protocols for community detection. In *26th Annual European Symposium on Algorithms (ESA 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

L. Becchetti, A. Clementi, E. Natale, F. Pasquale, and L. Trevisan. Find your place: Simple distributed algorithms for community detection. *SIAM Journal on Computing*, 49(4):821–864, 2020.

R. Berthier, F. Bach, and P. Gaillard. Tight nonparametric convergence rates for stochastic gradient descent under the noiseless linear model. In *Advances in Neural Information Processing Systems*, volume 33, pages 2576–2586, 2020.

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

M. Betancourt, M. Jordan, and A. Wilson. On symplectic optimization. *arXiv preprint arXiv:1802.03653*, 2018.

P. Billingsley. *Probability and measure*. John Wiley & Sons, 2008.

M. Biskup et al. Recent progress on the random conductance model. *Probability Surveys*, 8:294–373, 2011.

B. Bordelon and C. Pehlevan. Learning curves for sgd on structured features. *arXiv preprint arXiv:2106.02713*, 2021.

L. Bottou and Y. Le Cun. On-line learning for very large data sets. *Applied Stochastic Models in Business and Industry*, 21(2):137–151, 2005.

L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah. Randomized gossip algorithms. *IEEE Transactions on Information Theory*, 52(6):2508–2530, 2006.

P. Braca, S. Marano, and V. Matta. Enforcing consensus while monitoring the environment in wireless sensor networks. *IEEE Transactions on Signal Processing*, 56(7):3375–3380, 2008.

D. Bresters. On the equation of euler–poisson–darboux. *SIAM Journal on Mathematical Analysis*, 4(1):31–41, 1973.

S. Bubeck, Y. T. Lee, and M. Singh. A geometric alternative to Nesterov's accelerated gradient descent. *arXiv preprint arXiv:1506.08187*, 2015.

M. Cao, D. A. Spielman, and E. M. Yeh. Accelerated gossip algorithms for distributed computation. In *44th Annual Allerton Conference on Communication, Control, and Computation*, pages 952–959, 2006.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

V. Cevher and B. C. Vũ. On the linear convergence of the stochastic gradient method with constant step-size. *Optimization Letters*, 13(5):1177–1187, 2019.

J. Chen and A. H. Sayed. Diffusion adaptation strategies for distributed optimization and learning over networks. *IEEE Transactions on Signal Processing*, 60(8):4289–4305, 2012.

R. P. Chhabra. Non-newtonian fluids: an introduction. In *Rheology of complex fluids*, pages 3–34. Springer, 2010.

L. Chizat and F. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/file/a1afc58c6ca9540d057299ec3016d726-Paper.pdf.

F. Chung. *Spectral Graph Theory*. Number 92 in CBMS Regional Conference Series in Mathematics. American Mathematical Soc., 1997.

M. Cohen, J. Diakonikolas, and L. Orecchia. On acceleration with noise-corrupted gradients. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1019–1028. PMLR, 2018.

F. Cucker and D. X. Zhou. *Learning theory: an approximation theory viewpoint*, volume 24. Cambridge University Press, 2007.

H. Cui, B. Loureiro, F. Krzakala, and L. Zdeborová. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. *arXiv preprint arXiv:2105.15004*, 2021.

N. Curien. Random walks and graphs, 2020. lecture notes, available at https://www.imo.universite-paris-saclay.fr/~curien/enseignement.html.

G. Darboux. *Leçons sur la théorie générale des surfaces*. 1896.

A. d'Aspremont, D. Scieur, and A. Taylor. Acceleration methods, 2021.

M. Davis. Piecewise-deterministic markov processes: a general class of non-diffusion stochastic models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(3):353–376, 1984.

M. H. Davis. *Markov models & optimization*. Routledge, 2018.

B. Delyon and A. Juditsky. On the computation of wavelet coefficients. *Journal of Approximation Theory*, 88(1):47–79, 1997.

O. Devolder. Stochastic first order methods in smooth convex optimization. Technical report, CORE, 2011.

J. Diakonikolas and L. Orecchia. The approximate duality gap technique: A unified theory of first-order methods. *SIAM Journal on Optimization*, 29(1):660–689, 2019.

R. Diekmann, A. Frommer, and B. Monien. Efficient schemes for nearest neighbor load balancing. *Parallel Computing*, 25(7):789–812, 1999.

A. Dieuleveut and F. Bach. Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics*, 44(4):1363–1399, 2016.

A. Dieuleveut, N. Flammarion, and F. Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *The Journal of Machine Learning Research*, 18(1):3520–3570, 2017.

A. Dimakis, A. Sarwate, and M. Wainwright. Geographic gossip: Efficient averaging for sensor networks. *IEEE Transactions on Signal Processing*, 56(3):1205–1216, 2008. ISSN 1053-587X. doi: 10.1109/tsp.2007.908946. URL http://dx.doi.org/10.1109/TSP.2007.908946.

A. Dimakis, S. Kar, J. Moura, M. Rabbat, and A. Scaglione. Gossip algorithms for distributed signal processing. *Proceedings of the IEEE*, 98(11):1847–1864, 2010.

J. Duchi, A. Agarwal, and M. J. Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic control*, 57(3):592–606, 2012.

B. Durhuus. Hausdorff and spectral dimension of infinite random graphs. *Acta Phys. Polon.*, 40: 3509–3532, 2009.

W. Ellens, F. Spieksma, P. Van Mieghem, A. Jamakovic, and R. Kooij. Effective graph resistance. *Linear algebra and its applications*, 435(10):2491–2506, 2011.

L. Euler. Institutiones calculi integralis, vol iii, Petropoli. 1770.

L. Evans. Partial differential equations. *Graduate studies in mathematics*, 19(2), 1998.

M. Even, H. Hendrikx, and L. Massoulié. Asynchrony and acceleration in gossip algorithms. *arXiv preprint arXiv:2011.02379*, 2020.
B. Fischer. *Polynomial based iteration methods for symmetric linear systems*. Springer, 1996.
S. Fischer and I. Steinwart. Sobolev norm learning rates for regularized least-squares algorithm. *arXiv preprint arXiv:1702.07254*, 2017.
N. Flammarion and F. Bach. From averaging to acceleration, there is only a step-size. In *Conference on Learning Theory*, pages 658–695. PMLR, 2015.
J. Friedman. A proof of Alon's second eigenvalue conjecture. In *Proceedings of the thirty-fifth annual ACM Symposium on Theory of computing*, pages 720–724. ACM, 2003.
W. Gautschi. *Orthogonal polynomials: computation and approximation*. Oxford University Press on Demand, 2004.
B. V. Gnedenko. On a local limit theorem of the theory of probability. *Uspekhi Matematicheskikh Nauk*, 3(3):187–194, 1948.
G. Golub and C. Van Loan. *Matrix computations*, volume 3. JHU press, 2013.
G. Grimmett. What is percolation? In *Percolation*, pages 1–31. Springer, 1999.
L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.
Hemis62. Illustration licensed under CC BY-SA 4.0. *https://commons.wikimedia.org/wiki/File:Torus_graph.png*.
H. Hendrikx, F. Bach, and L. Massoulié. Accelerated decentralized optimization with local updates for smooth and strongly convex objectives. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 897–906. PMLR, 2019.
T. Hofmann, B. Schölkopf, and A. J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, pages 1171–1220, 2008.
C. Hu, W. Pan, and J. Kwok. Accelerated gradient methods for stochastic optimization and online learning. In *Advances in Neural Information Processing Systems*, volume 22, pages 781–789, 2009.
N. Ikeda and S. Watanabe. *Stochastic differential equations and diffusion processes*. Elsevier, 2014.
J. Jacod and A. Shiryaev. *Limit theorems for stochastic processes*, volume 288. Springer Science & Business Media, 2013.
P. Jain, S. M. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford. Accelerating stochastic gradient descent for least squares regression. In *Conference On Learning Theory*, pages 545–604, 2018.
M. Jordan and T. Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349 (6245):255–260, 2015.
A. Juditsky and Y. Nesterov. Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization. *Stochastic Systems*, 4(1):44–80, 2014.
A. Juditsky, A. Kulunchakov, and H. Tsyntseus. Sparse recovery by reduced variance stochastic approximation. *arXiv preprint arXiv:2006.06365*, 2020.
K.-S. Jun, A. Cutkosky, and F. Orabona. Kernel truncated randomized ridge regression: Optimal rates and low noise acceleration. In *Advances in Neural Information Processing Systems*, pages 15332–15341, 2019.
M. Kac. A stochastic model related to the telegrapher's equation. *The Rocky Mountain Journal of Mathematics*, 4(3):497–509, 1974.
M. Kearns and U. Vazirani. *An introduction to computational learning theory*. MIT press, 1994.
D. Kim and J. Fessler. Optimized first-order methods for smooth convex minimization. *Mathematical programming*, 159(1):81–107, 2016.
M. Kohler and A. Krzyżak. Optimal global rates of convergence for interpolation problems with random design. *Statistics & Probability Letters*, 83(8):1871–1879, 2013.

N. Korda, B. Szörényi, and L. Shuai. Distributed clustering of linear bandits in peer to peer networks. In *International Conference on Machine Learning*, volume 48, pages 1301–1309, 2016.

W. Krichene, A. Bayen, and P. Bartlett. Accelerated mirror descent in continuous and discrete time. *Advances in Neural Information Processing Systems*, 28:2845–2853, 2015.

G. Lan. An optimal method for stochastic composite optimization. *Math. Program.*, 133(1-2, Ser. A):365–397, 2012.

P. Landgren, V. Srivastava, and N. E. Leonard. On distributed cooperative decision-making in multiarmed bandits. In *2016 European Control Conference (ECC)*, pages 243–248. IEEE, 2016.

J. Latz. Analysis of stochastic gradient descent in continuous time. *Statistics and Computing*, 31 (4):1–25, 2021.

G. Lawler and V. Limic. *Random walk: a modern introduction*, volume 123. Cambridge University Press, 2010.

J.-F. Le Gall. *Brownian Motion, Martingales, and Stochastic Calculus*, volume 274. Springer, 2016.

A. M. Legendre. *Nouvelles méthodes pour la détermination des orbites des comètes: avec un supplément contenant divers perfectionnemens de ces méthodes et leur application aux deux comètes de 1805*. Courcier, 1806.

J. Lin and V. Cevher. Optimal convergence for distributed learning with stochastic gradient methods and spectral-regularization algorithms. *arXiv preprint arXiv:1801.07226*, 2018.

T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

J. Liu, B. Anderson, M. Cao, and S. Morse. Analysis of accelerated gossip algorithms. *Automatica*, 49(4):873–883, 2013.

N. Loizou, M. Rabbat, and P. Richtárik. Provably accelerated randomized gossip algorithms. In *ICASSP 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 7505–7509. IEEE, 2019.

S. Ma, R. Bassily, and M. Belkin. The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning. In *Proceedings of the 35th International Conference on Machine Learning*, pages 3325–3334, 2018. URL `http://proceedings.mlr.press/v80/ma18a.html`.

P. Mathieu and E. Remy. Isoperimetry and heat kernel decay on percolation clusters. *The Annals of Probability*, 32(1A):100–128, 2004.

H. McKean. Some mathematical coincidences. *https://as.nyu.edu/content/dam/nyu-as/asSilverDialogues/documents/McKean,%20Henry-Silver%20Dialogues.pdf*, 2003.

M. Mezard and A. Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

C. Moallemi and B. V. Roy. Consensus propagation. In *Advances on Neural Information Processing Systems*, pages 899–906. MIT Press, 2005.

B. Mohar and W. Woess. A survey on spectra of infinite graphs. *Bulletin of the London Mathematical Society*, 21(3):209–234, 1989.

G. Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

E. Montijano, J. I. Montijano, and C. Sagues. Chebyshev polynomials in distributed consensus applications. *IEEE Transactions on Signal Processing*, 61(3):693–706, 2012.

G. Moore. Excerpts from a conversation with Gordon Moore: Moore's law. *https://web.archive.org/web/20121029060050/http://download.intel.com/museum/Moores_Law/Video-Transcripts/Excepts_A_Conversation_with_Gordon_Moore.pdf*, 2005.

N. Mücke, G. Neu, and L. Rosasco. Beating SGD saturation with tail-averaging and minibatching. In *Advances in Neural Information Processing Systems*, pages 12568–12577, 2019.

M. Muehlebach and M. Jordan. A dynamical systems perspective on Nesterov acceleration. In *International Conference on Machine Learning*, pages 4656–4662. PMLR, 2019.

P. Müller and P. Stollmann. Spectral asymptotics of the laplacian on supercritical bond-percolation graphs. *Journal of Functional Analysis*, 252(1):233–246, 2007.

A. Nedic and A. Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.

A. Nedic, A. Ozdaglar, and P. Parrilo. Constrained consensus and optimization in multi-agent networks. *IEEE Transactions on Automatic Control*, 55(4):922–938, 2010.

A. S. Nemirovskij and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience, 1983.

A. S. Nemirovsky. On optimality of krylov's information when solving linear operator equations. *Journal of Complexity*, 7(2):121–130, 1991.

A. S. Nemirovsky. Information-based complexity of linear operator equations. *Journal of Complexity*, 8(2):153–175, 1992.

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: A Basic Course*, volume 87. Springer Science & Business Media, 2003.

Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

P. Nevai. *Orthogonal polynomials*, volume 23. American Mathematical Soc., 1979.

P. Nevai. Géza Freud, orthogonal polynomials and Christoffel functions. a case study. *Journal of Approximation Theory*, 48(1):3–167, 1986.

F. W. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark. *NIST handbook of mathematical functions*. Cambridge University Press, 2010.

M. Penrose. *Random geometric graphs*. Number 5. Oxford University Press, 2003.

G. Peyré and M. Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

L. Pillaud-Vivien, A. Rudi, and F. Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. In *Advances in Neural Information Processing Systems*, pages 8114–8124, 2018.

S. D. Poisson. *Mémoire sur l'intégration des équations linéaires aux différences partielles*. J. de l'Ecole Polytechnique, 1823.

B. Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

B. Polyak. Introduction to optimization. *Optimization Software, Inc, New York*, 1987.

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

P. Rebeschini and S. Tatikonda. Accelerated consensus via min-sum splitting. In *Advances on Neural Information Processing Systems*, 2017.

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

L. Rosasco and S. Villa. Learning with incremental iterative regularization. In *Advances in Neural Information Processing Systems*, pages 1630–1638, 2015.

V. Roulet and A. d'Aspremont. Sharpness, restart, and acceleration. *SIAM Journal on Optimization*, 30(1):262–289, 2020.

J. M. Sanz-Serna and K. Zygalakis. The connections between Lyapunov functions for some optimization algorithms and differential equations. *arXiv preprint arXiv:2009.00673*, 2020.

S. Sardellitti, M. Giona, and S. Barbarossa. Fast distributed average consensus algorithms based on advection-diffusion processes. *IEEE Transactions on Signal Processing*, 58(2):826–842, 2010.

A. Sayed. Adaptation, learning, and optimization over networks. *Foundations and Trends® in Machine Learning*, 7(4-5):311–801, 2014.

K. Scaman, F. Bach, S. Bubeck, Y. T. Lee, and L. Massoulié. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *International Conference on Machine Learning*, 2017.

M. Schmidt and N. Le Roux. Fast convergence of stochastic gradient descent under a strong growth condition. *arXiv preprint arXiv:1308.6370*, 2013.

D. Shah. Gossip algorithms. *Foundations and Trends® in Networking*, 3(1):1–125, 2009.

B. Shi, S. Du, M. Jordan, and W. Su. Understanding the acceleration phenomenon via high-resolution differential equations. *arXiv preprint arXiv:1810.08907*, 2018.

B. Shi, S. Du, W. Su, and M. Jordan. Acceleration via symplectic discretization of high-resolution differential equations. In *Advances in Neural Information Processing Systems*, volume 32, pages 5744–5752, 2019.

S. Sodin. Random matrices, nonbacktracking walks, and orthogonal polynomials. *Journal of Mathematical Physics*, 48(12):123503, 2007.

W. Su, S. Boyd, and E. Candes. A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. *Advances in neural information processing systems*, 27:2510–2518, 2014.

T. Sun, Y. Sun, Y. Xu, and W. Yin. Markov chain block coordinate descent. *Computational Optimization and Applications*, 75(1):35–61, 2020.

S. Sundhar Ram, A. Nedić, and V. Veeravalli. Distributed stochastic subgradient projection algorithms for convex optimization. *Journal of optimization theory and applications*, 147(3):516–545, 2010.

G. Szegö. *Orthogonal polynomials*, volume 23. American Mathematical Soc., 1939.

B. Szorenyi, R. Busa-Fekete, I. Hegedus, R. Ormándi, M. Jelasity, and B. Kégl. Gossip-based distributed stochastic bandit algorithms. In *International Conference on Machine Learning*, pages 19–27, 2013.

P. Tarrès and Y. Yao. Online learning as stochastic approximation of regularization paths: Optimality and almost-sure convergence. *IEEE Transactions on Information Theory*, 60(9):5716–5735, 2014.

P. Tseng and S. Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117(1):387–423, 2009.

A. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Science & Business Media, 2008.

A. Varre, L. Pillaud-Vivien, and N. Flammarion. Last iterate convergence of sgd for least-squares in the interpolation regime. *arXiv preprint arXiv:2102.03183*, 2021.

S. Vaswani, F. Bach, and M. Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1195–1204. PMLR, 2019.

U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

G. Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, 1990. doi: 10.1137/1.9781611970128. URL `https://epubs.siam.org/doi/abs/10.1137/1.9781611970128`.

H. Wendland. *Scattered Data Approximation*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2004. doi: 10.1017/CBO9780511617539.

A. Wibisono, A. Wilson, and M. I. Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113(47):E7351–E7358, 2016.

A. Wilson, B. Recht, and M. Jordan. A Lyapunov analysis of momentum methods in optimization. *arXiv preprint arXiv:1611.02635*, 2016.

W. Woess. *Random walks on infinite graphs and groups*, volume 138. Cambridge University Press, 2000.

S. Wojtowytsch. Stochastic gradient descent with noise of machine learning type. part i: Discrete time analysis. *arXiv preprint arXiv:2105.01650*, 2021.

S. Wright. Coordinate descent algorithms. *Math. Program.*, 151(1, Ser. B):3–34, 2015.

L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *J. Mach. Learn. Res.*, 11:2543–2596, 2010.

Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

Y. Ying and M. Pontil. Online gradient descent learning algorithms. *Foundations of Computational Mathematics*, 8(5):561–596, 2008.

L. Zdeborová and F. Krzakala. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.

J. Zhang, A. Mokhtari, S. Sra, and A. Jadbabaie. Direct Runge-Kutta discretization achieves acceleration. In *Advances in Neural Information Processing Systems*, volume 31, pages 3900–3909, 2018.