Stochastic optimization for large-scale machine learning : variance reduction and acceleration

Andrei Kulunchakov

Université Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE L' UNIVERSITÉ GRENOBLE ALPES

Spécialité : Mathématiques Appliquées

Arrêté ministériel : 25 mai 2016

Présentée par

Andrei KULUNCHAKOV

Thèse dirigée par **Anatoli JUDITSKY**, professeur, Université Grenoble Alpes, et codirigée par **Julien MAIRAL**, chargé de recherche, Inria

préparée au sein du **Laboratoire Jean Kuntzmann** dans **l'École Doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique**

Optimisation stochastique pour l'apprentissage machine à grande échelle: réduction de la variance et accélération

Stochastic optimization for large-scale machine learning: variance reduction and acceleration

Thèse soutenue publiquement le **3 décembre 2020**,
devant le jury composé de :

Monsieur Jalal Fadili
Professeur, ENSI Caen, Rapporteur
Monsieur Anatoli Juditsky
Professeur, Université Grenoble Alpes, Directeur de thèse
Monsieur Guanghui Lan
Associate Professor, Georgia Tech, Rapporteur
Monsieur Julien Mairal
Chargé de Recherche, Inria, Co-Directeur de thèse
Monsieur Jérôme Malick
Directeur de Recherche, CNRS, Président
Madame Asu Ozdaglar
Professor, EECS, Examinateur

Abstract

Many machine learning problems are formulated in terms of minimization of mathematical functions, a task which is then solved by various optimization methods. As more and more data is becoming available, optimization techniques have to address problems with high-dimensional feature spaces and inexact information. Therefore, first-order optimization methods with low per-iteration cost have become predominant in practice and have attained considerable attention in the literature in recent years. A common characteristic of this type of techniques is slow convergence, which motivates a great interest in *acceleration* of existing algorithms in order to save computational resources.

A goal of this thesis is to explore several topics in optimization for high-dimensional stochastic problems. The first task is related to various incremental approaches, which rely on exact gradient information, such as SVRG, SAGA, MISO, SDCA. While the minimization of large limit sums of functions was thoroughly analyzed, we suggest in Chapter 2 a new technique, which allows to consider all these methods in a generic fashion and demonstrate their robustness to possible stochastic perturbations in the gradient information. Our technique is based on extending the concept of estimate sequence introduced originally by Yurii Nesterov in order to accelerate deterministic algorithms. Using the finite-sum structure of the problems, we are able to modify the aforementioned algorithms to take into account stochastic perturbations. At the same time, the framework allows to derive naturally new algorithms with the same guarantees as existing incremental methods. Finally, we propose a new accelerated stochastic gradient descent algorithm and a new accelerated SVRG algorithm that is robust to stochastic noise. This acceleration essentially performs the typical deterministic acceleration in the sense of Nesterov, while preserving the optimal variance convergence.

Next, we address the problem of *generic* acceleration in stochastic optimization, trying to repeat the success of *direct* acceleration performed in Chapter 2. For this task, we generalize in Chapter 3 the multi-stage approach called Catalyst, which was originally aimed to accelerate deterministic methods. In order to apply it to stochastic problems, we improve its flexibility with respect to the choice of surrogate functions minimized at each stage. Finally, given an optimization method with mild convergence guarantees for strongly convex problems, our developed multi-stage procedure accelerates convergence to a noise-dominated region, and then achieves a near-optimal (optimal up to a logarithmic factor) worst-case convergence depending on the noise variance of the gradients. Thus, we successfully address the acceleration of various stochastic methods, including the variance-reduced approaches considered and generalized in Chapter 2. Again, the developed framework bears similarities with the acceleration performed by Yurii Nesterov using the estimate sequences. In this sense, we try to fill the gap between deterministic and stochastic optimization in terms of Nesterov's acceleration. A side contribution of this chapter is a generic analysis that can handle inexact proximal operators, providing new insights about the robustness of stochastic algorithms when the proximal operator cannot be exactly computed.

In Chapter 4, we study properties of non-Euclidean stochastic algorithms applied to the problem of sparse signal recovery. A sparse structure significantly reduces the effects of noise in gradient observations. We propose a new stochastic algorithm,

called SMD-SR, allowing to make better use of this structure. This method is a multi-step procedure which uses the stochastic mirror descent algorithm as a building block over its stages. Essentially, the SMD-SR procedure has two phases of convergence with the linear convergence during the preliminary phase and the optimal asymptotic rate during the asymptotic phase. Comparing to the most effective existing solutions to stochastic sparse optimization, we offer an improvement in several aspects. First, we establish the linear convergence of the initial error (similar to the one of the deterministic gradient descent algorithm, when the full gradient observation $\nabla f(x)$ is available), while showing the optimal robustness to noise. Second, we achieve this rate for a large class of noise models, including sub-Gaussian, Rademacher, multivariate Student distributions and scale mixtures. Finally, these results are obtained under the optimal condition on the level of sparsity which can approach the total number of iterations of the algorithm (up to a logarithmic factor).

Keywords: stochastic optimization, convex optimization, complexity, acceleration, sparse optimization

Résumé

De nombreux problèmes d'apprentissage automatique sont formulés en termes de minimisation de fonctions mathématiques, une tâche qui est ensuite résolue par diverses méthodes d'optimisation. Une réponse aux défis liés au traitement de données massives et hétérogènes fait appel aux techniques d'optimisation pour des problèmes de grande dimension et avec des informations incertaines. C'est pourquoi les méthodes d'optimisation à faible coût par itération, dites de premier ordre, sont devenues un outil numérique principal d'apprentissage automatique, et ont attiré une attention considérable dans la littérature au cours des dernières années. Une caractéristique commune de ce type de technique est une convergence lente, ce qui motive un grand intérêt pour l'*accélération* des algorithmes existants afin d'économiser les ressources de calcul.

Cette thèse vise à explorer divers sujets liés à l'analyse des méthodes de premier ordre appliquées à des problèmes stochastiques de grande dimension. Notre première contribution porte sur divers algorithmes incrémentaux, tels que SVRG, SAGA, MISO, SDCA, qui ont été analysés de manière approfondie pour les problèmes avec des informations de gradient exactes. Nous proposons dans le chapitre 2 une nouvelle technique, qui permet de traiter ces méthodes de manière unifiée et de démontrer leur robustesse à des perturbations stochastiques lors de l'observation des gradients. Notre approche est basée sur une extension du concept de suite d'estimation introduite par Yurii Nesterov pour l'analyse d'algorithmes déterministes accélérés. En utilisant la structure de somme finie des problèmes considérés, nous proposons une modification de ces algorithmes pour tenir compte des perturbations stochastiques. De plus, notre approche permet de concevoir de façon naturelle de nouveaux algorithmes incrémentaux offrant les mêmes garanties que les méthodes existantes tout en étant robustes aux perturbations stochastiques. Enfin, nous proposons un nouvel algorithme de descente de gradient stochastique accéléré et un nouvel algorithme SVRG accéléré robuste au bruit stochastique. Dans le dernier cas il s'agit essentiellement de l'accélération déterministe au sens de Nesterov, qui préserve la convergence optimale des erreurs stochastiques.

Ensuite, nous abordons le problème de l'accélération *générique*, en essayant de répéter le succès de l'accélération directe réalisée au Chapitre 2. Pour cela, nous étendons dans le Chapitre 3 l'approche multi-étapes de Catalyst, qui visait à l'origine l'accélération de méthodes déterministes. Afin de l'appliquer aux problèmes stochastiques, nous le modifions pour le rendre plus flexible par rapport au choix des fonctions auxiliaires minimisées à chaque étape de l'algorithme. Finalement, à partir d'une méthode d'optimisation pour les problèmes fortement convexes, avec des garanties *standard* de convergence, notre procédure commence par accélérer la convergence vers une région dominée par le bruit, pour converger avec une vitesse quasi-optimale ensuite. Cette approche nous permet d'accélérer diverses méthodes stochastiques, y compris les algorithmes à variance réduite décrits et généralisés au Chapitre 2. Là encore, le cadre développé présente des similitudes avec l'analyse d'algorithmes accélérés à l'aide des suites d'estimation proposées par Yurii Nesterov. En ce sens, nous essayons de combler l'écart entre l'optimisation déterministe et stochastique en termes d'accélération de Nesterov. Une autre contribution de ce chapitre est une analyse unifiée d'algorithmes proximaux stochastiques lorsque l'opérateur proximal ne peut pas être calculé de façon exacte.

Au Chapitre 4, nous étudions des propriétés d'algorithmes stochastique non-Euclidiens appliqués au problème d'estimation parcimonieuse. La structure de parcimonie permet de réduire de façon significative les effets du bruit dans les observation du gradient. Nous proposons un nouvel algorithme stochastique, appelé SMD-SR, permettant de faire meilleur usage de cette structure. Là encore, la méthode en question est une routine multi-étapes qui utilise l'algorithme stochastique de descente en miroir comme élément constitutif de ses étapes. Cette procédure comporte deux phases de convergence, dont la convergence linéaire de l'erreur pendant la phase préliminaire, et la convergence à la vitesse asymptotique optimale pendant la phase asymptotique. Par rapport aux solutions existantes les plus efficaces aux problèmes d'optimisation stochastique parcimonieux, nous proposons une amélioration sur plusieurs aspects. Tout d'abord, nous montrons que l'algorithme proposé réduit l'erreur initiale avec une vitesse linéaire (comme un algorithme déterministe de descente de gradient, utilisant l'observation complète du gradient de la fonction-objective), avec un taux de convergence optimal par rapport aux caractéristiques du bruit. Deuxièmement, nous obtenons ce taux pour une grande classe de modèles de bruit, y compris les distributions sous-gaussiennes, de Rademacher, de Student multivariées, etc. Enfin, ces résultats sont obtenus sous la condition optimale sur le niveau de parcimonie qui peut approcher le nombre total d'iterations de l'algorithme (à un facteur logarithmique près).

Mots-clés : optimisation stochastique, optimisation convexe, complexité, accélération, optimisation parcimonieuse

Acknowledgements

This thesis would not have been possible without the exceptional support of my supervisors. I am sincerely grateful to Julien Mairal and Anatoli Juditsky for these years of our fruitful collaboration. Your inspiration and creative energy were one of the key driving forces during my PhD. Your outstanding professionalism, kindness and commitment have helped me to improve myself in many aspects, not solely from the academic point of view. It was a great pleasure to work with you.

I would like also to express my gratitude to Guanghui Lan and Jalal Fadili for their time invested in reviewing this thesis. I sincerely appreciate your participation in my defense, as well as having incredible researchers Jérôme Malick and Asu Ozdaglar in the defense jury. It was a great pleasure to finalize this work in front of such celebrated committee.

My thankfulness goes to all members of the Thoth team for making these three years a great time with their energy, kindness and exploration of wide variety of concepts through endless exciting discussions. Your inspiring creativity created an incredible atmosphere for work and research. I am especially thankful to my friends from Russia, which have been reaching me across the borders all these years, providing me a support, which is hard to describe in words.

And I want to express my heartfelt gratitude to my parents, who are always there for me. Thank you for your unconditional love, warmth and wisdom, which support me all these years. It is difficult to convey how blessed I am to have you both beside me. The success of this work is partially owed to you.

Contents

Chapter 1

Introduction

Over the last decades, a great research interest has been directed towards data science and machine learning problems, such as fraud and risk detection, search engines, recommendation systems, image and speech recognition, reinforcement learning, to name a few. Many of these problems are naturally cast as maximization of a quality criterion over a set of unknown parameters, leading to an increasing demand for *mathematical optimization* routines. Nowadays, the choice of optimization method is one of the key steps in building a machine learning pipeline.

Recent technological advances in data collection and storage raise new challenges in *large-scale* machine learning problems that essentially involve optimization over sets of parameters of particularly large dimension [Bottou et al., 2018]. Specifically, there is a demand for efficient optimization methods that process data samples at low cost from both computational and memory usage points of views, while preserving good theoretical guarantees. This setting is beyond the realm of classical polynomial time optimization methods [d'Aspremont, 2008, Juditsky and Nemirovski, 2011a], so the *first-order methods* take the leading role.

One specific artifact of large-scale setting is working under *uncertainty* because usually one does not have access to precise information about incoming data samples. This concept is often represented by the assumption that the objective function to optimize is *stochastic*, leading to *stochastic optimization* [Nemirovsky and Yudin, 1983]. Operating on problems of this kind, first-order methods can have access only to inexact and (typically) unbiased estimations of gradients. This setting dates back to the pioneering work of [Robbins and Monro, 1951], who established that even under uncertainty such methods converge to the optimal solution in expectation under rather mild conditions imposed on the criterion. Since then a large amount of work was devoted to develop, enhance or analyze different methods dealing with stochasticity [see Nemirovski et al., 2009, Ghadimi and Lan, 2012, Lan, 2012, Bottou et al., 2018] among many others. When dealing with stochasticity, one popular approach is to use the *variance reduction* technique, where gradient estimates are iteratively refined in order to decrease the variance of estimates [Johnson and Zhang,

2013, Defazio et al., 2014a, Allen-Zhu, 2017, Lan and Zhou, 2018a]. In this thesis, we provide a new viewpoint to shed new light on several variance reduction techniques. In addition, we improve their robustness to uncertainty and design new algorithms with better convergence guarantees.

Each optimization method is built over a set of assumptions imposed on problems to be solved. Therefore, it naturally belongs to one or several particular classes of methods unified by the nature of these assumptions. Methods of each class treat problems of particular type, that is, for example, optimization problems with/without uncertainty, with/without sparsity etc. For instance, it is known that some deterministic accelerated methods may be impractical on optimization problems with uncertainty. As another example, methods that automatically adapt to hidden geometrical properties of criteria (like strong convexity) may be beneficial, when an experimenter does not know if a problem possesses such properties. Therefore, because typically the exact type of a problem is unknown in practice and the exhaustive search over methods of different classes is usually intractable in large-scale setting, there exists a strong demand for developing *universal* methods that are applicable to wide classes of optimization problems [Lan, 2012, Allen-Zhu and Yuan, 2016].

Apart from specifics of the large-scale setting, we are also interested in general approaches to *acceleration* of existing algorithms. An example of generic scheme accelerating first-order optimization methods called Catalyst was introduced in [Lin et al., 2015]. It universally treats algorithms from several classes of methods, including those related to variance reduction techniques, and can be seen as an inexact accelerated proximal point algorithm. The method to be accelerated is used iteratively to solve approximately a well-constructed sequence of problems. However, this technique is not applicable to problems endowed with uncertainty. In the thesis, we address this challenge and show how to make Catalyst robust to uncertainty in incoming data.

Another common way to deal with large dimensionality is to exploit the inner structure of a problem. For instance, in machine learning applications, it happens frequently that the quality criterion has many insignificant or negligible parameters. Therefore, it is natural to attempt to recover only a small number of these parameters, while trimming the others. This setting leads to *sparse recovery* optimization problems that have gained a lot of attention in the literature from both practical and theoretical perspectives [Donoho et al., 2000, Bühlmann and Van De Geer, 2011], and demonstrate their merits in a variety of applications [Agarwal et al., 2012b]. Sparsity of solutions may be induced in several ways: by using ℓ_0 penalization [Blumensath and Davies, 2009, Jain et al., 2014, Bhatia et al., 2015] that specifically may be seen as feature selection; by relaxing the problem to ℓ_1-minimization, leading to the well studied Lasso and Dantzig Selector estimators [see Juditsky and Nemirovski, 2011c, and references therein]; or by an intermediate choice of ℓ_p-"norm" with $p \in (0, 1)$ [Foygel Barber and Liu, 2019, Zhao and Luo, 2019] balancing the merits of the first two approaches. In any case, exploiting sparsity always results in better convergence with respect to data uncertainty. In the thesis, we develop a fast first-order method that exploits sparse structure for a wide variety of problems, including sparse regression and low-rank matrix recovery. Moreover, we show that the theoretical guarantees of the resulted method are improved comparing to the best currently known results from [Hazan and Kale, 2010, Juditsky and Nesterov, 2010, Agarwal et al., 2012b,

Ghadimi and Lan, 2013].

In this introduction, we gave a high-level conceptual overview of main directions and challenges taken by the thesis. The contributions of the thesis related both to theoretical and practical sides of our findings are listed in the next section along with a brief explanation.

1.1 Contributions of the thesis

Before we give a detailed overview with precise definitions and explanations of the concepts briefly introduced above, we present the list of the main contributions of the thesis from a high-level perspective, and we cite the corresponding publications. Detailed descriptions for each point will be given in Sections 2.1.1, 3.1.1 and 4.1.1 respectively.

— In Chapter 2, we propose a unified view for a certain class of first-order algorithms of stochastic optimization. This class consists of different incremental approaches with a specific form of the gradient step, including variants of stochastic gradient descent [Robbins and Monro, 1951] and several variance-reduced algorithms. This common viewpoint is developed by extending the concept of *estimate sequences* introduced in [Nesterov, 1983, 2014]. More precisely, using the estimate sequence construction, we interpret the considered methods as procedures that iteratively minimize a surrogate of the objective. Finally, this framework allows us to come up with (i) a unified viewpoint and proof of convergence for all of these methods; (ii) generic strategies to make these algorithms robust to stochastic noise (the aforementioned *uncertainty* in optimization problems), (iii) new accelerated stochastic variance-reduced algorithm with theoretically optimal complexity. While (i) is a rather minor addition, (ii) and (iii) are of particular importance, as for example, robustness to noise and acceleration of variance-reduced methods was only partially analyzed in the literature. This contribution is based on the following publications

  - A. Kulunchakov and J. Mairal. Estimate sequences for variance-reduced stochastic composite optimization. In *Proceedings of the International Conference on Machine Learning (ICML)*, June 2019c
  - A. Kulunchakov and J. Mairal. Estimate sequences for stochastic composite optimization: Variance reduction, acceleration, and robustness to noise. *Journal of Machine Learning Research (JMLR)*, 2020

— In Chapter 3, we introduce several mechanisms to "generically" accelerate different first-order stochastic algorithms of convex and strongly convex optimization. In a nutshell, we extend the Catalyst approach [Lin et al., 2015], originally developed for deterministic problems, to the stochastic setting. The developed generalization of Catalyst includes a generic acceleration of variance-reduced algorithms for the case of stochastic problems, which relies upon the results of Chapter 2. In brief, the applicable stochastic methods (those that can be accelerated) are united by a mild condition of being linearly convergent to a fixed noise-dominated region (all notions are defined in Section 1.3). From a high-level perspective, given

such an algorithm, we accelerate its convergence to a noise-dominated region, and then achieve a near-optimal convergence by using a restart procedure with exponentially increasing mini-batches. Finally, we demonstrate that the developed generic acceleration framework is competitive with the directly accelerated methods studied in Chapter 2. The overall contribution is based on the following paper.

- A. Kulunchakov and J. Mairal. A generic acceleration framework for stochastic composite optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019

— In Chapter 4, we consider the problem of sparse signal recovery in stochastic optimization setting. Using the stochastic mirror descent as a workhorse, we develop a multi-stage procedure with explicit feature selection between stages. We show that the resulted algorithm enjoys fast linear convergence to a noise-dominated region, where exploitation of sparse structure of the problem allows to show better theoretical guarantees compared to non-sparsity-aware optimization algorithms. At the same time, we maintain the optimal asymptotic rate, shown in [Agarwal et al., 2012b] for sparse regression, while enlarging the set of admissible values of the number s of nonvanishing signal components from $\lesssim \sqrt{N}$ to $\lesssim N$, where N is the number of iterations. We also enhance the reliability of the corresponding solutions by using Median-of-Means like techniques [Nemirovsky and Yudin, 1983, Minsker, 2015]. This contribution is based on the following manuscript.

- A. Juditsky, A. Kulunchakov and H. Tsyntseus. Sparse Recovery by Reduced Variance Stochastic Approximation. *arXiv:2006.06365*, 2020

1.2 Optimization problems in machine learning

In machine learning applications, the behavior of a real system of interest is usually described in terms of some parametrized mathematical model [Hastie et al., 2001, Bottou et al., 2018], chosen by experimenters. They also set up a mathematical function that represents the "goodness of fit"—how good does the model describe the system of interest—with arguments which are the model parameters. One is naturally interested in searching for such values of these parameters that maximize the "goodness", looking thus for the best approximation model. Throughout the thesis, we refer to this quality criterion as *objective function*.

But first, let us give a more formal explanation. Assume that our system generates data samples (z_i) that live in a subset Z of Euclidean space E. Assume also that these data samples (z_i) possess some characteristics represented by labels (y_i) that lie in a subset Y of another Euclidean space E_Y. Specifically, there is an unknown one-to-one mapping $h_* : Z \to Y$ embedded into the system by nature. This mapping is of interest to us and we want to predict $h_*(z)$ for all $z \in Z$.

In order to recover h, or at least to approximate it, experimenters choose a family of prediction functions—candidate models,

$$\mathcal{H} = \{h(z, x) : Z \times X \to Y\} \tag{1.1}$$

—parametrized with x from a subset X of Euclidean space E_X [Bottou et al., 2018]. This set is assumed to contain a parameter vector corresponding to a "satisfying" model. In order to compare different models, we choose a *loss function* $l : Z \times Y \to \mathbb{R}$ that measures goodness of fit at data samples. When $E_Y = \mathbb{R}$, examples of loss function include squared error loss $l(h(z,x),y) = (h(z,x)-y)^2$, squared hinge loss $l(h(z,x),y) = \max\{0,(1-y\cdot h(z,x))^2\}$ and logistic loss $l(h(z,x),y) = \log\left(1+\exp^{-y\cdot h(z,x)}\right)$. In what follows, unless stated otherwise, we always assume that $E = E_X = \mathbb{R}^p$ and $E_Y = \mathbb{R}$. Moreover, if not explicitly stated, $\|\cdot\|$ will denote the ℓ_2 norm. More generally, there are several ways of using a loss function to select a model [Hastie et al., 2001, Bottou et al., 2018].

1.2.1 Expected Risk and Empirical Risk

Assume that data samples (z_i) and labels (y_i) are random variables with probability distribution $P(z,y)$ on $Z \times Y$. Then, the main objective is to minimize the following function

$$F(x) = \int\limits_{Z\times Y} l(h(z,x),y)dP(z,y) = \mathbb{E}_P\left[l(h(z,x),y)\right], \tag{1.2}$$

referred to as *expected risk*. This risk gives a complete characterization of the model fit over the whole set $Z \times Y$ thus being the ideal criterion when choosing a proper approximation model. Unfortunately, for most cases, the multidimensional integral in (1.2) can not be computed even when distribution P is known.

A common approach is to approximate (1.2) with a finite sum

$$F(x) = \frac{1}{n}\sum_{i=1}^{n} l(h(z_i,x),y_i), \tag{1.3}$$

which converges to the expected risk according to the law of large numbers [Dekking et al., 2005]. This criterion (1.3) is called *empirical risk* [Vapnik, 2000].

1.2.2 Regularization

When optimizing the sampled objective (1.3), one may encounter the problem of overfitting [Hastie et al., 2001]. For example, this problem occurs when data samples (z_i) are corrupted by some noise and the selected model excessively fits this noise thus losing its predictive abilities on unseen data. To mitigate this effect, one popular approach consists in imposing a penalty on model complexity—*risk regularization*. As models in (1.1) are parametrized with x, this complexity is also associated with x and is expressed as a penalty function $\psi(x)$ added to (1.3) thus leading to *composite optimization* problems intensively studied in the literature [Burke and Ferris, 1995, Lan, 2012, Li and Pong, 2015, Shi et al., 2015, Ghadimi et al., 2016, Gasnikov and Nesterov, 2018, to name just a few]. The regularized empirical risk looks as follows

$$F(x) = \frac{1}{n}\sum_{i=1}^{n} l(h(z_i,x),y_i) + \psi(x) \triangleq f(x) + \psi(x). \tag{1.4}$$

For instance, ψ may be the ℓ_p-norm that, for a real $p \geq 1$ is defined as

$$\|x\|_p = \left(|x_1|^p + |x_2|^p + \cdots + |x_n|^p\right)^{1/p}. \tag{1.5}$$

A particular example of (1.5) is the ℓ_1-norm, which is very popular in signal processing and machine learning applications due to for its sparsity-inducing properties [Mairal et al., 2014, Nesterov and Nemirovski, 2013, Becker et al., 2011, Beck and Teboulle, 2009]. Another way to define ψ is to set it as the extended-valued indicator function of a specific subset $X_0 \subset X$

$$\psi(x) = \begin{cases} 0, & \text{when } x \in X_0, \\ +\infty, & \text{when } x \in X_0 \end{cases} \tag{1.6}$$

so that the setting (1.4) encompasses constrained problems [Necoara and Patrascu, 2014, Hiriart-Urruty and Lemaréchal, 1996]. A proper choice of $\psi(x)$ is dictated not only by desired properties of geometry of the resulted solutions, but also by "simplicity" of $\psi(x)$. This notion of simplicity will be expressed in details in the next section, when introducing *proximal operators*, that has to be computed efficiently.

In what follows, we denote x^* the optimal solution to the optimization problem, which may be (1.2), (1.3) or (1.4), and also refer to the optimal objective value $F^* = F(x^*)$. Before introducing proximal operators, we focus on the case when $\psi = 0$, so that we minimize $f(x)$, as the success of optimization of (1.4) depends mostly on the properties of $f(x)$, not of the regularization $\psi(x)$. In many situations, the optimization problems stated in (1.4) are unsolvable [Nesterov, 2014]. Therefore, we need to introduce and assume specific geometric properties of $f(x)$.

1.2.3 Geometry of a problem. Convexity.

One of the main branches of vivid developments in optimization tools is convex optimization.

Definition 1.1 (Convex set)**.** *A set $X \subset E$ is convex if $(1-\alpha)x_1 + \alpha x_2$ belongs to X for every $x_1, x_2 \in X$ and any $\alpha \in [0, 1]$.*

Definition 1.2 (Convex function)**.** *A function $f : X \to \mathbb{R}$ defined on a convex set X is called convex if $\forall x_1, x_2 \in X$ and $\forall \alpha \in [0, 1]$ the following condition holds:*

$$f((1-\alpha)x_1 + \alpha x_2) \leq (1-\alpha)f(x_1) + \alpha f(x_2).$$

Convexity plays important role in different applications due to three important implications. The first states that any local minimum of a convex function is a global minimum as well. This property is important as optimization tools in non-convex optimization are prone to get stuck in local minima that can be sub-optimal [Nesterov, 2014]. The second one is the existence of *sub-gradients*, good surrogates of gradient of convex function [Nemirovsky and Yudin, 1983].

Definition 1.3. *If $f : X \to \mathbb{R}$ is a convex function, a vector g is called a subgradient at a point $x \in X$ if for any $y \in X$ one has*

$$f(y) - f(x) \geq \langle g, y - x \rangle .$$

The final property is that while providing characteristics of optimization problems, the property of convexity is flexible enough so that many machine learning tasks may be formulated in terms of optimization of a convex objective, [see Boyd et al., 2004, Dattorro, 2010, Bubeck, 2014, and references therein]. For instance, the examples of popular loss functions described earlier, that is squared error loss, hinge loss and logistic loss, are all convex functions. Of course, if the loss function $l(h(z, x), y)$ is convex for all $(z, y) \in Z \times Y$, then both risks (1.2) and (1.3) are convex functions as well. The penalty function $\psi(x)$ is also typically assumed to be convex, which is the case for the ℓ_p-norm with $p \geq 1$ or the extended-valued penalty function of a convex set (1.6).

Definition 1.2 may be strengthened with introduction of *strong convexity*.

Definition 1.4 (Strongly convex function)**.** *A function $f : X \to \mathbb{R}$ defined on a convex set X is called strongly convex with parameter μ if $\forall x_1, x_2 \in X$ and $\forall \alpha \in [0, 1]$ the following condition holds:*

$$f((1-\alpha)x_1 + \alpha x_2) \leq (1-\alpha)f(x_1) + \alpha f(x_2) - \frac{\mu\alpha(1-\alpha)}{2} \|x_1 - x_2\|_2^2 . \tag{1.7}$$

First, strongly convex functions have unique optima. Second, one may verify by comparing results of [Lan, 2012] with [Ghadimi and Lan, 2013] that the strong convexity property significantly improves convergence of optimization algorithms. [1]

Another important notion, that arises throughout the thesis, is L-smoothness.

Definition 1.5 (L-smooth functions)**.** *Assume that a function $f : X \to \mathbb{R}$ is continuously differentiable everywhere on X [Nemirovsky and Yudin, 1983], then it is called L-smooth if its gradient is Lipschitz continuous with constant L, that is*

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\| .$$

The L-smoothness property allows to prove better theoretical convergence rates as we will see later. This assumption is not strong, as many optimization objectives in machine learning applications are L-smooth. For example, if $f(x)$ is convex and twice continuously differentiable, a sufficient condition would be an existence of an upper bound L for eigenvalues of the Hessian $\nabla^2 f(x)$.

Problem statement Finally, we give a formal description of the general problem that we optimize throughout the thesis.

$$\min_{x \in X} \left\{ F(x) \triangleq f(x) + \psi(x) \right\}, \tag{1.8}$$

where X is a convex set, ψ is a convex, semi-continuous penalty function, not necessarily differentiable, and f is (μ-strongly) convex and L-smooth.

1. This will also be shown in Section 1.3, presenting theoretically optimal convergence rates for convex and strongly convex optimization.

1.2.4 Proximal operators

In order to incorporate the composite setting of (1.4), one can introduce *proximal operator* associated with the penalty $\psi(x)$ and the projection on the convex set X [Moreau, 1962]. Specifically, it is defined as

$$\text{Prox}_{\eta\psi}[u] \triangleq \underset{x \in X}{\text{argmin}} \left\{ \eta\psi(x) + \frac{1}{2}\|x - u\|^2 \right\}. \tag{1.9}$$

The usage of a proximal operator allows to optimize composite $F(x)$ without querying sub-gradient information on $\psi(x)$, when it is simple enough for (1.9) to be computed in closed form or by an efficient fast search routine [Parikh and Boyd, 2014]. This is the case, for example, for quadratic penalties, l_p-norms (with $p \geq 1$) or an extended-valued indicator function of a simple convex set [Agarwal et al., 2012b, Parikh and Boyd, 2014]. Due to the use of (1.9), the optimization methods applied to the pure risks (1.2) and (1.3) are often generalizable to the composite setting with the same convergence guarantees [see Beck and Teboulle, 2009, Lan, 2012, Nesterov, 2013].

1.2.5 Oracle complexity of optimization problems

A classical statement of optimization problem involves the notion of a *black-box* [Nemirovsky and Yudin, 1983, Nesterov, 2014]. Black-box models suggest that one does not possess any information about $f(x)$, but rather accesses it by pieces through queries to *oracles*. The most popular type is a *first-order* oracle that returns the gradient $\nabla f(x)$ at any requested point x.

The black-box model allows to build a meaningful definition of *complexity* of an optimization method being the number of oracle calls requested in order to achieve an ε-accurate solution $\hat{x}$, that is $f(\hat{x}) - f^* \leq \varepsilon$. This is also called *analytical complexity* compared to *arithmetical complexity* defined as the number of arithmetic operations requested by a method [Nesterov, 2014]. Though the latter is a more ideal measure in practice, we stick to analytical complexity, because it allows to obtain a complete theory of convex optimization [Nemirovsky and Yudin, 1983] by providing concrete upper bounds on complexities of methods that could be achieved. Moreover, arithmetic complexity can usually be derived from it [Nesterov, 2014].

In the *stochastic setting*, an oracle can return noisy estimates $g(x)$ of the true gradient $\nabla f(x)$. Typically, in the literature [Nemirovsky and Yudin, 1983, Nemirovski et al., 2009, Bottou et al., 2018], it is assumed that $g(x)$ is unbiased $g(x) = \mathbb{E}[\nabla f(x)]$ along with introduction of the following important characteristic, called *oracle variance*.

Definition 1.6 (Oracle variance)**.** *Assume that there is an oracle that for each $x \in X$ returns an unbiased noisy estimate $g(x)$ of the true gradient $\nabla f(x)$. The variance of this estimate is denoted as σ^2*

$$\forall x \in X \quad \mathbb{E}\|g(x) - \nabla f(x)\|^2 \leq \sigma^2. \tag{1.10}$$

In other words, the value of σ^2 is a measure of the oracle quality. The assumption of boundness of the oracle variance is not a weak one. For example, the expected

deviation $\mathbb{E}\left\|g(x)-\nabla f(x)\right\|^2$ may often be of the same order as $\left\|\nabla f(x)\right\|^2$, and the latter may be unbounded, which is the case, for instance, for the aforementioned squared hinge loss $l(h(z,x),y)=\max\left\{0,(1-y\cdot h(z,x))^2\right\}$. However, some functions, such as logistic loss, have bounded gradients over $\mathbb{R}^p$. Moreover, the minimization may be conducted over a compact condition set X, so that the value of $\mathbb{E}\left\|g(x)-\nabla f(x)\right\|^2$ remains bounded.

1.3 Classical theoretical results on complexities of optimization methods

Before we overview and develop algorithms, we discuss *how fast* these algorithms may be in theory. In terms of analytical complexity, we are mainly concerned with the optimal rates of methods applied to convex and strongly convex objectives. We assume that an oracle in Definition 1.6 is at our disposal and we express the optimal *worst-case* convergence rates. Given an optimal rate, it means that there exists a function of the considered class, which can not be minimized faster than with this prescribed rate. For the rest of the manuscript, the convergence rates given in this section are referred to as *optimal* and we refer to the algorithms that achieve them (either for $\sigma=0$ or $\sigma>0$) as *accelerated*. All other rates are referred to as *sub-optimal*. In what follows, we often represent the results in the $\mathcal{O}(\cdot)$ notation for simplicity hiding the absolute constants. Moreover, we introduce the notation $\tilde{\mathcal{O}}(1)$ that is essentially $\mathcal{O}(\cdot)$ with probably a hidden logarithmic factor in the problem dimension or condition number L/μ.

Convex case. Assume that the smooth part $f(x)$ in the objective $F(x)$ of (1.8) is convex, but not strongly convex. According to [Nemirovsky and Yudin, 1983], for this class of functions the best convergence rate is

$$\mathbb{E}\left[F(x_N)-F^*\right]\leq \underbrace{\mathcal{O}\left(\frac{LR^2}{N^2}\right)}_{\text{bias}}+\underbrace{\mathcal{O}\left(\frac{\sigma R}{\sqrt{N}}\right)}_{\text{variance}}, \tag{1.11}$$

where R is an upper bound on $\left\|x_0-x^*\right\|$ with x_0 being the initial estimate and N is the total number of oracle calls. In this statement, we highlight a classical *bias-variance decomposition* of a convergence rate of a stochastic method [Bach and Moulines, 2013, Dieuleveut et al., 2017]. The "bias" term expresses the "reaction" of an algorithm to the "deterministic" component of random observations. It is associated with the Lipschitz property of the gradient of $f(x)$. The variance term expresses the effectiveness of treatment of noise inherited by the oracle. In what follows, we distinguish *preliminary* and *asymptotic* phases of convergence of a method, specifically for *bias* or *variance* domination in (1.11) respectively.

The deterministic case $\sigma=0$ was successfully solved in the seminal paper [Nesterov, 1983], establishing an accelerated algorithm with the optimal rate. The stochastic case $\sigma>0$, when neglecting bias, was considered in several papers [Nesterov and Vial, 2008, Nemirovski et al., 2009, Xiao, 2010] yielding the optimal asymptotic rate $\mathcal{O}\left(1/\sqrt{N}\right)$ using averaging techniques and decreasing step sizes. The overall universal case (1.11) then was closed in [Lan, 2012], achieving optimal bias and variance.

Convergence rate	$\Rightarrow$	Complexity
$\frac{L\|x_0-x^*\|^2}{N^2}+\frac{\sigma\|x_0-x^*\|}{\sqrt{N}}$		$\|x_0-x^*\|\sqrt{\frac{L}{\varepsilon}}+\frac{\sigma^2\|x_0-x^*\|^2}{\varepsilon^2}$
$\frac{L\|x_0-x^*\|^2}{N}+\frac{\sigma\|x_0-x^*\|}{\sqrt{N}}$		$\frac{L\|x_0-x^*\|^2}{\varepsilon}+\frac{\sigma^2\|x_0-x^*\|^2}{\varepsilon^2}$
$\left(1-\sqrt{\frac{\mu}{L}}\right)^N(F(x_0)-F^*)+\frac{\sigma^2}{\mu N}$		$\sqrt{\frac{L}{\mu}}\log\left(\frac{F(x_N)-F^*}{\varepsilon}\right)+\frac{\sigma^2}{\mu\varepsilon}$
$\left(1-\frac{\mu}{L}\right)^N(F(x_0)-F^*)+\frac{\sigma^2}{\mu N}$		$\frac{L}{\mu}\log\left(\frac{F(x_N)-F^*}{\varepsilon}\right)+\frac{\sigma^2}{\mu\varepsilon}$

Table 1.1 – List of convergence rates to and corresponding complexities in different cases. For brevity of presentation, every expression is given up to absolute constants.

Strongly convex case. Once an optimization problem is strongly convex with $\mu > 0$, the best possible convergence rate of a method applied to minimize it can be significantly improved. According to [Nemirovsky and Yudin, 1983, Agarwal et al., 2012a, Nesterov, 2014], for this class of problems, a method can not be faster than

$$\mathbb{E}\left[F(x_N)-F^*\right] \leq \mathcal{O}(1)\left(1-\sqrt{\frac{\mu}{L}}\right)^N(F(x_0)-F^*)+\mathcal{O}\left(\frac{\sigma^2}{\mu N}\right). \tag{1.12}$$

The deterministic case $\sigma = 0$ was solved in [Nemirovsky and Yudin, 1983]. As in the convex setting, the stochastic case $\sigma > 0$ in the pure asymptotic regime was successfully considered in [Nesterov and Vial, 2008, Nemirovski et al., 2009, Lacoste-Julien et al., 2012] yielding the optimal asymptotic rate $\mathcal{O}(1/N)$ by decreasing step sizes accordingly $\mathcal{O}(1/N)$. In the stochastic setting, the overall universal case (1.12) then was closed in [Ghadimi and Lan, 2013].

Terminology Throughout the thesis, we refer to different notions related to convergence rates established above. Next, for such N that the bias is dominated by the variance part of rate, we say that an algorithm has converged to a *noise-dominated region*. Another important notion here is complexity, given in Section 1.2.5. Complexity of a method is basically derived from its convergence rate, being the number of iterations required to obtain a solution x_N such that $\mathbb{E}\left[F(x_N)-F^*\right] \leq \varepsilon$ for a given $\varepsilon > 0$. While the convergence rate expresses an achieved accuracy via number of iterations, complexity does an inverse task, defining the latter via accuracy, see Table 1.1 for examples.

Main goal In the current thesis, we explore different ways to achieve both (1.11) and (1.12) simultaneously in the general stochastic composite setting. It is important to note that by the term "simultaneously" we mean that, while the parameters of the developed framework may change depending on μ, the same framework nonetheless achieves either (1.11) or (1.12) through rather straightforward modifications.

1.4 Stochastic algorithms and variance reduction

Before we give an overview of the methods developed for stochastic optimization problems with inexact information ($\sigma > 0$), we introduce two key paradigms for solving them, based on two different Monte Carlo sampling techniques.

1.4.1 SA and SAA paradigms

The expected risk (1.2) and the empirical risk (1.3) are minimized differently in the sense of the Monte Carlo sampling technique which is used to access the information. Depending on which risk is minimized— (1.2) or (1.3)—an applied method can belong to the *stochastic approximation* (SA) paradigm or the *sample average approximation* (SAA) paradigm respectively. Let us now describe each of them in details.

SAA methods are essentially two-step deterministic algorithms applied to (1.3). The first step consists in construction of deterministic risk to be optimized. At the second step, once all pairs $\{(z_i, y_i)\}_{i=1}$ are sampled and (1.3) bears no uncertainty, a deterministic optimization algorithm is applied. While numerical complexity of the solution depends on the optimization procedure used at the second step, the statistical properties are completely defined by the sampling step [Candes et al., 2007, Negahban et al., 2012, Zhang et al., 2014]. Nonetheless, different studies [Kleywegt et al., 2002, Shapiro, 2003, Shapiro and Nemirovski, 2005, Linderoth et al., 2006] show that these techniques are quite efficient from both theoretical and practical points of view [Nemirovski et al., 2009].

The first stochastic approximation method appears in the seminal paper [Robbins and Monro, 1951]. SA methods directly target the expected risk (1.2) processing incoming data samples one by one in online manner. At the core of SA methods lies a simple algorithm with iterative updates of low computational cost. This allows to significantly reduce computational burden comparing to SAA, as experimenters do not need to store and reuse the whole sampled pool of (1.3).

For a long time SAA methods were considered superior to SA approaches [Nemirovski et al., 2009], partially because the corresponding optimization routines may use the structure of the problem to solve. Another reason resided in poorly working step size strategy established by classical theory [Chung, 1954, Sacks, 1958]. The first step in improvement of SA approaches was done in [Nemirovsky and Yudin, 1983, Polyak and Juditsky, 1992] where longer step sizes were justified, so that the resulted algorithms are shown to be competitive with classical SAA approaches. Since then, SA algorithms developed rapidly and became a highly dynamic domain with numerous important contributions [Duchi et al., 2011, Lan, 2012, Ghadimi and Lan, 2013, Kingma and Ba, 2014, to name just a very few]. In what follows, we focus on algorithms of SA type. We start with a description of a common example of SA algorithm.

1.4.2 Stochastic gradient descent

The gradient descent (GD) algorithm [Rumelhart et al., 1985, Baldi, 1995, Ruder, 2016] is probably the most studied optimization method in the literature and widely used in practice due to its simplicity. In a nutshell, this method boils down to a simple scheme:

set up an initial estimation x_0, define a step size sequence $(\eta_i)_{i=1}$ and iteratively apply the update rule for $k \geq 1$

$$x_k = x_{k-1} - \eta_k \nabla f(x_{k-1}), \tag{1.13}$$

(we will assume $\psi(x) = 0$ and $X = \mathbb{R}^p$ to start). [2]. The main logic behind (1.13) is that *locally* the antigradient shows the direction of the steepest descent of $f(x)$ [Nesterov, 2014]. The step sizes $(\eta_i)_{i=1}$ control how far the GD goes in its updates. There are many different strategies on the choice of η_k [Duchi et al., 2011, Nesterov, 2014, Ruder, 2016], but in any case, for L-smooth functions, there is a limit value for it. This value is found from minimizing the right side of

$$f(x_k) \leq f(x_{k-1}) + \nabla f(x_{k-1})^\top (x_k - x_{k-1}) + \frac{L}{2}\|x_k - x_{k-1}\|_2^2,$$

yielding finally $\eta_k \leq 1/L$. The inequality holds true for L-smooth functions due to Theorem 2.1.5 in [Nesterov, 2014]. For the rest of the section, we assume that the step size is constant $\eta_i = \eta$ for all i.

In the stochastic setting, we do not have access to the exact gradients $\nabla f(x)$ and dispose of its stochastic estimates $g(x)$ only. Therefore, the relation (1.13) is rewritten as

$$x_k = x_{k-1} - \eta_k g(x_{k-1}) \tag{1.15}$$

being the update of the well-known *stochastic gradient descent* algorithm. This update is preferable in large-scale setting, because computations of $g(x)$ may be performed much faster than that of $\nabla f(x)$, [Zhang, 2004, Bottou et al., 2018]. We consider algorithms with averaging of iteration trajectories [Polyak and Juditsky, 1992], so that the approximate solution at the step N is formed according to

$$\hat{x}_N = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{1.16}$$

Define R as an upper bound on the initial approximation error $\|x_0 - x^*\|$. It is well-known that the following bound holds for $\hat{x}_N$ in the case of a convex and L-smooth objective [Nemirovsky and Yudin, 1983, Nemirovski et al., 2009]

$$\mathbb{E}\left[f(\hat{x}_N) - f^*\right] \leq \frac{R^2}{N\eta} + 2\eta\sigma^2. \tag{1.17}$$

When the number N of performed iterations is known in advance (we refer to this setting as *finite horizon*), the bound (1.17) transforms into

$$\mathbb{E}\left[f(\hat{x}_N) - f^*\right] \leq \frac{LR^2}{N} + \frac{2\sigma R}{\sqrt{N}} \tag{1.18}$$

2. In the case, when $\psi(x) \neq 0$, this update transforms into

$$x_k = \mathrm{Prox}_{\eta_k \psi}\left[x_{k-1} - \eta_k \nabla f(x_{k-1})\right]. \tag{1.14}$$

(by utilizing a proper choice of η).

For μ-strongly convex functions, (1.17) becomes

$$\mathbb{E}\left[f(x_N) - f^*\right] \leq (1 - \eta\mu)^N \left(f(x_0) - f^*\right) + \frac{\eta L \sigma^2}{2\mu}, \tag{1.19}$$

that may be found for example in Theorem 4.6 in [Bottou et al., 2018]. The variance is still bounded, and the convergence of the bias terms becomes faster due to improved curvature of $f(x)$. The variance term in (1.19) may be reduced by a specific restarting procedure (see discussion after Theorem 4.6 in [Bottou et al., 2018]). In a nutshell, this procedure employs exponentially decreasing step sizes, that are reduced each time we decrease the expected risk by a factor 2. Note that there are different other ways to obtain a converging algorithm. For instance, one can consider the following scheme: (i) launch (1.15) with a fixed step size $\eta = 1/L$, (ii) reach a *noise-dominated* region $f(x_N) - f^* \leq \eta L \sigma^2/\mu$, (iii) restart the algorithm with decreasing step sizes $\eta_k \leq \min\{\eta, 2/\mu k\}$. The final rate of the described procedure, according to Theorem 4.7 of [Bottou et al., 2018], is

$$\mathbb{E}\left[f(x_N) - f^*\right] \leq \left(1 - \frac{\mu}{L}\right)^N \left(f(x_0) - f^*\right) + \frac{L}{\mu}\frac{2\sigma^2}{\mu N}. \tag{1.20}$$

This rate is not optimal as the variance depends on excessive factor (L/μ) that could be large in high-dimensional problems.

1.4.3 Stochastic Mirror Descent

To finish consideration of the SGD, we need to introduce its celebrated descendant, called *stochastic mirror descent*. Let the SGD algorithm be applied to an optimization problem with $\psi = 0$ and constrained to a convex set X. Then, its update (1.14) is as follows

$$x_k = \mathrm{proj}_X \left[x_{k-1} - \eta_k \nabla f(x_{k-1})\right],$$

with proj_X being the projection operator on X. This rule can be viewed as the minimization of a quadratically penalized local Taylor expansion of $f(x)$ around x_{k-1}, that is

$$x_k = \operatorname*{argmin}_{x \in X} \left\{ f(x_{k-1}) + \langle \nabla f(x_{k-1}), x - x_{k-1} \rangle + \underbrace{\frac{1}{2\eta_k} \left\| x - x_{k-1} \right\|_2^2}_{\text{penalty}} \right\}. \tag{1.21}$$

One may replace the Euclidean norm with a different regularizer to obtain a different update rule of the algorithm.

Bregman divergence Let E be an Euclidean space and $\vartheta : E \to \mathbb{R}$ be a continuously differentiable convex function which is strongly convex with respect to some norm $|\cdot|$ (that is not necessarily Euclidean), i.e.,

$$\langle \nabla\vartheta(x) - \nabla\vartheta(x'), x - x' \rangle \geq |x - x'|^2, \quad \forall x, x' \in E.$$

Following [Nesterov, 2005, Juditsky and Nemirovski, 2011a], we refer to ϑ as a distance-generating function (for the examples of d.-g. functions for different norms $|\cdot|$ see [Nesterov, 2005, Juditsky and Nemirovski, 2011a]). We define Bregman divergence associated with the d.-g. function ϑ according to

$$V(x,z) = \vartheta(z) - \vartheta(x) - \langle \nabla\vartheta(x), z - x\rangle, \quad \forall\, z, x \in X. \tag{1.22}$$

Being substituted into (1.21), it leads to the update

$$\begin{aligned} x_k &= \operatorname*{argmin}_{z\in X}\left\{ f(x_{k-1}) + \langle \nabla f(x_{k-1}), z - x_{k-1}\rangle + \frac{1}{2\eta_k} V(x_{k-1}, z)\right\} \\ &= \operatorname*{argmin}_{z\in X}\left\{ \langle \nabla f(x_{k-1}), z\rangle + \frac{1}{2\eta_k} V(x_{k-1}, z)\right\} \\ &= \operatorname*{argmin}_{z\in X}\left\{ \left\langle \nabla f(x_{k-1}) - \frac{1}{2\eta_k}\nabla\vartheta(x_{k-1}), z\right\rangle + \frac{1}{2\eta_k}\vartheta(z)\right\}. \end{aligned} \tag{1.23}$$

Finally, the exact gradient $\nabla f(x_{k-1})$ may be replaced by its stochastic observation g_k, so that we arrive to the update rule of the stochastic mirror descent algorithm:

$$\boxed{x_k = \operatorname*{argmin}_{z\in X}\left\{ \langle g_k, z\rangle + \frac{1}{2\eta_k} V(x_{k-1}, z)\right\}}. \tag{1.24}$$

This optimization problem is solved at each iteration of the stochastic mirror descent algorithm. In order to allow for the efficient implementation of the method, this problem should be easy (for example, to admit a closed form solution or may be solved by a simple linear search).

Note that the proximal operator considered here is substantially different from that defined in Section 1.2.4. In Chapters 2 and 3, we use the notion of proximal operator as defined in (1.9), while in Chapter 4 the definition (1.24) is used. It remains to note that the convergence guarantees of the SMD algorithm are of the same type as those of the SGD algorithm in the case of Euclidean norm.

1.4.4 Variance-reduced algorithms

In order to mitigate the impact of noise that comes from an inexact oracle, one may exploit a particular structure of the optimization problem. A typical example is provided by the *finite-sum* structure. A finite-sum optimization problem is as follows

$$\min_{x\in X}\left\{ F(x) \triangleq f(x) + \psi(x) \triangleq \frac{1}{n}\sum_{i=1}^{n} f_i(x) + \psi(x)\right\}. \tag{1.25}$$

The assumptions imposed on X, $f(x)$ and $\psi(x)$ are the same as in (1.8). Let us assume for now that the *deterministic* terms $f_i(x)$, same as $f(x)$, are μ-strongly convex and L-smooth.

Because all $(f_i)_{i=1}^n$ are deterministic, the overall objective $F(x)$ is deterministic as well. Moreover, it is assumed that the oracles returning gradient estimations $\nabla f_i(x)$ for each i and $x \in X$ are exact. This allows as well to calculate the full gradient $\nabla f(x) =$

$(1/n)\sum_{i=1}^n f_i$ and apply any deterministic procedure to (1.25). However, in large-scale setting this strategy would be too expensive from a computational point of view. Therefore, new optimization algorithms were proposed recently which utilize the finite-sum structure, while maintaining computational speed and simplicity of SA approaches.

While the finite-sum setting is obviously a particular case of expectation with a discrete probability distribution, the deterministic nature of $F(x)$ drastically changes the corresponding performance guarantees. The stochastic gradient descent may be applied to the minimization of a finite-sum objective when the random gradient realizations is taken as $g_k = f_{i_k}$, where i_k is chosen randomly from $[1, n]$. While the update (1.15) is computationally cheap, it does not explicitly exploit the finite-sum structure of $f(x)$. Therefore, the stochastic gradient descent, accessing only random unbiased gradient realizations, can not improve on the $\mathcal{O}(1/N)$-rates from (1.12).

Nonetheless, linear convergence rates can be obtained if first-order methods operate not on the estimates $\nabla f_i(x)$, but rather on a refined version of them with a decreasing variance. Such technique has been introduced by [Blatt et al., 2007, Schmidt et al., 2017] under the name of a *variance-reduced* algorithm to build proper unbiased estimates $g_k(x)$ of $\nabla f(x)$ with the variance $\mathbb{E}\left[\|g_k(x) - \nabla f(x)\|^2\right]$ decreasing over the iterations. Since then, different versions of $g_k(x)$ were proposed leading to randomized incremental approaches with linear convergence rates, such as SAG [Schmidt et al., 2017], SAGA [Defazio et al., 2014a], SVRG [Johnson and Zhang, 2013, Xiao and Zhang, 2014], SDCA [Shalev-Shwartz and Zhang, 2016], MISO [Mairal, 2015], Katyusha [Allen-Zhu, 2017], MiG [Zhou et al., 2018], SARAH [Nguyen et al., 2017a], directly accelerated SAGA [Zhou, 2019] or RPDG [Lan and Zhou, 2018a].

The key idea used in building the estimate $g_k(x)$ in variance-reduced algorithms is that given two random variables X and Y, it is possible to define a new variable $Z = X - Y + \mathbb{E}[Y]$ which has the same expectation as X but potentially a lower variance if Y is positively correlated with X. For example, SVRG approach takes a gradient step using the following estimate at a step k

$$g_k = \nabla f_{i_k}(x_{k-1}) - \nabla f_{i_k}(\hat{x}) + \nabla f(\hat{x}),$$

where $\hat{x}$ is an *anchor* point updated every n steps. The full gradient $\nabla f(\hat{x})$ is stored after the update of $\hat{x}$, and the calculation of $\nabla f_{i_k}(\hat{x})$ is cheap, so that the overall update (1.15) is cheap on average as well. The resulted convergence rate of the algorithm applied to (1.25) with exact oracles for $(f_i(x))_{i=1}^n$ is

$$\mathbb{E}[F(x_N) - F^*] \leq \mathcal{O}(1)\left(1 - \frac{\mu}{L}\right)^N (F(x_0) - F^*), \tag{1.26}$$

which is similar to the one obtained in [Defazio et al., 2014a, Shalev-Shwartz and Zhang, 2016, Mairal, 2015, Schmidt et al., 2017], while faster convergence rates were obtained in [Allen-Zhu, 2017, Lan and Zhou, 2018a, Zhou et al., 2018, Zhou, 2019]. To attain (1.26) one needs to conduct several passes over the data, so that N should be larger than n. All these algorithms have about the same cost per-iteration as the stochastic gradient descent method, being $\mathcal{O}(n)$ times lower than of SAA techniques. However, these results holds only for the case when the terms f_i are deterministic.

1.4.5 Variance-reduced algorithms with perturbations

In this thesis, our objective is to provide a unified view of stochastic optimization algorithms, with and without variance reduction, but we are especially interested in improving their *robustness* to random perturbations. Specifically, we consider objectives with an explicit finite-sum structure when only *inexact* estimates of the gradients $\nabla f_i(x)$ are available. In other words, there are n oracles $\tilde{f}_i$ of the general type (1.6) for each of the terms $(f_i)_i^n$ corrupted by random perturbations $(\rho_i)_i^n$, and the smooth part f from (1.8) may be written as

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) \quad \text{with} \quad f_i(x) = \mathbb{E}_{\rho_i}\left[\tilde{f}_i(x, \rho_i)\right]. \tag{1.27}$$

The described setting may occur in various applications. For instance, the perturbations ρ_i may be injected during training in order to achieve better generalization on the test data [Srivastava et al., 2014], perform stable feature selection [Meinshausen and Bühlmann, 2010], improve the model robustness [Zheng et al., 2016], or for privacy-aware learning [Wainwright et al., 2012].

Such problems can not be addressed by algorithms of deterministic optimization, and most of the aforementioned variance-reduction methods do not apply anymore. The standard approach to address this problem is to ignore the finite-sum structure and use variants of the SGD algorithms which were already seen to be sub-optimal. The reason for this is that the variance of the gradient estimate for (1.27) decomposes into two parts $\sigma^2 = \sigma_s^2 + \tilde{\sigma}^2$, where σ_s^2 is due to the random sampling of the index i_k and $\tilde{\sigma}^2 \ll \sigma_s$ is due to the random data perturbation in the terms $(f_i)_{i=1}^n$. The SGD algorithm preserves the constant variance that depends on σ_s^2. And while variance-reduced algorithms manage to git rid of it when $\tilde{\sigma}^2 = 0$, generalization to the case when $\tilde{\sigma}^2 > 0$ is still an open question for most of them.

In this thesis, our objective is to achieve *robustness* to the noise $\tilde{\sigma}^2 > 0$, while preserving a linearly convergent term, that is obtaining the following convergence guarantees

$$\mathbb{E}\left[F(x_N) - F^*\right] \leq \mathcal{O}(1)\left(1 - \frac{\mu}{L}\right)^N (F(x_0) - F^*) + \mathcal{O}(1)\frac{\tilde{\sigma}^2}{\mu N}, \tag{1.28}$$

with the variance that is typically much smaller than $\sigma^2/(\mu N)$. This is not a contradiction with the theoretically optimal result stated in (1.12), as the assumptions about the oracles are different. We adapt several incremental algorithms, including variance-reduced methods such as SVRG, SAGA, SDCA or MISO, to stochastic optimization setting by providing a unified common convergence proof for them. The main novelty is the modification to the aforementioned algorithms that makes them robust to stochastic perturbations.

Note that the SAGA and SVRG methods were adapted for this purpose by [Hofmann et al., 2015], although the resulting algorithms have non-zero asymptotic error. The MISO method was adapted by [Bietti and Mairal, 2017] at the cost of a memory overhead of $\mathcal{O}(np)$, whereas other variants of SAGA and SVRG were proposed by [Zheng and Kwok, 2018] for linear models in machine learning. While non-uniform sampling strategies for incremental methods are now classical [Xiao and Zhang, 2014, Schmidt et al., 2015], the robustness to stochastic perturbations has not been studied for all these methods and the

aforementioned approaches have significant limitations. At the same time, they may be useful, when the terms $\tilde{f}_i$ have different Lipschitz smoothness constants. In this case, for example, a non-uniform strategy allows to express the convergence rate in terms of average Lipschitz smoothness constant $\bar{L} = (1/n)\sum_{i=1}^{n} L_i$, and not in $L = \max_i (L_i)$, which is larger. In this thesis, we adapt these strategies to work in the generalized stochastic variance-reduced algorithms.

1.5 Accelerated methods

Optimal optimization methods—those having the optimal bias and variance convergence—are of particular interest for practitioners. Let us start with consideration of fast optimization methods for the deterministic setting.

A heavy-ball algorithm with the optimal convergence rate for the deterministic case was introduced in [Polyak, 1964] for minimization of smooth strongly convex functions. Next, using the ideas behind the conjugate gradient method, authors of [Nemirovsky and Yudin, 1983] developed an algorithm with the optimal convergence rate for L-smooth functions. Finally, [Nesterov, 1983] introduced the optimal algorithm of smooth optimization, now known as the *accelerated gradient* method. Essentially, this method had the same worst-case complexity with [Nemirovsky and Yudin, 1983]. While the ideas behind the algorithm of [Nemirovsky and Yudin, 1983] are simple, the geometric interpretation of the accelerated gradient method is still a question for community leading to several attempts to explain it [Allen-Zhu and Orecchia, 2014, Bubeck et al., 2015, Drusvyatskiy et al., 2018]. This algorithm was successfully generalized to composite problems in [Tseng, 2008, Beck and Teboulle, 2009, Nesterov, 2013]. Finally, [Lan, 2012] generalized the Nesterov's method to a composite optimization setting of minimizing objectives with smooth stochastic and non-smooth components. Estimate sequences have already been used to analyze stochastic optimization algorithms [Devolder et al., 2011, Lin et al., 2014, Lu and Xiao, 2015]. In this thesis, one contribution is development of a unified framework utilizing estimate sequences for derivation and analysis of accelerated methods in the stochastic setting.

For the case of smooth strongly-convex stochastic optimization, the optimal method, attaining the rate (1.12), was described in [Ghadimi and Lan, 2013] using a multi-stage procedure built over the algorithm from [Ghadimi and Lan, 2012]. The idea of the multi-stage procedure is simple. Given some *base* method $\mathcal{M}$, we iteratively apply it to the problem of interest. Each stage is a launch of $\mathcal{M}$ (for a certain number of iterates) initialized from the last obtained solution. This scheme was successfully used to build an optimal algorithm for stochastic strongly convex optimization.

A different multi-stage scheme called Catalyst was developed in [Lin et al., 2015] for minimization of deterministic objectives. The Catalyst approach covers convex and strongly convex finite-sum optimization problems, each along with their composite settings. At the core of Catalyst lies the assumption that a base method $\mathcal{M}$ enjoys a sub-optimal linear convergence when minimizing strongly convex objectives. Then, Catalyst builds a sequence of ℓ_2-penalized, well-conditioned sub-problems, each of which is solved efficiently by $\mathcal{M}$ up to a given accuracy. After each stage the extrapolation step is applied in order to prepare the initialization for the next stage. The described multi-stage procedure

is shown to derive meta-algorithms with near-optimal convergence rates for many *base* approaches, such as gradient descent, block coordinate descent, SAG, SAGA, SDCA, SVRG, Finito/MISO, mentioned earlier. [3] In this thesis, our objective is to generalize the Catalyst approach to the stochastic case and achieve near-optimal convergence rates for a large class of stochastic optimization methods.

1.6 Optimization methods for sparse recovery

Another way to mitigate the impact of noise that comes from an inexact oracle, is to exploit a particular structure of the problem. One example of such structure is *sparsity*. For instance, a data generation model is called s-sparse, if the driving parameter vector x^* has at most s non-zero components. Although, we generally aim at solving sparse stochastic optimization problems of the general form (1.8), in this section, we focus on sparse linear regression and low-rank matrix recovery in order to simplify the overview. For sparse linear regression models, the labels are generated according to

$$y = \langle \phi, x^* \rangle + \sigma\xi, \tag{1.29}$$

where $\phi \in \mathbb{R}^p$ and $\xi \in \mathbb{R}$ are typically i.i.d. random variables, and x^* is s-sparse. The low-rank matrix recovery model arises straightforwardly from (1.29), that is

$$y = \langle \Phi, X^* \rangle + \sigma\xi, \tag{1.30}$$

where $\langle \cdot, \cdot \rangle$ is the Frobenius inner product and Φ is a matrix of regressors, and X^* has at most s non-zero eigenvalues. In this section and then in Chapter 4, we explicitly distinguish $\|\cdot\|_2^2$-norm from other norms (like ℓ_p-norms, matrix norms) that are also exploited.

Given a sparse linear regression model (1.29), one may try to recover x^* by using the least squared loss function and stating an optimization problem. For example, (1.2) gives rise to the following stochastic optimization problem

$$\min_{x \in X} \left\{ f_{\mathrm{SA}}(x) = \frac{1}{2}\mathbb{E}\left\| y_1 - \phi_1^\top x \right\|_2^2 \right\}, \tag{1.31}$$

where the minimization may be conducted, for instance, over the space X of s-sparse vectors. Note that observations y_i and ϕ_i provide us directly with unbiased estimates $g_i = \phi_i \left(\phi_i^\top x - y_i \right)$ of the true gradient $\nabla f_{\mathrm{SA}}(x)$. Therefore, it is possible to address (1.31) with methods of SA type.

Another type of loss is given by sample average approximation (SAA). In this case, we sample N observations (ϕ_i, y_i) and define the following quantities $f_i(x) = \left\| y_i - \phi_i^\top x \right\|_2^2$, $y = (y_1, \ldots, y_N)^\top$, and D—the matrix with columns $\{\phi_1, \phi_2, \ldots, \phi_N\}$. Then, the optimization objective becomes

$$f_{\mathrm{SAA}}(x) = \frac{1}{N}\sum_{i=1}^N f_i(x) = \frac{1}{2N}\left\| y - D^\top x \right\|_2^2, \tag{1.32}$$

where the matrix $D \in \mathbb{R}^{p \times N}$ is referred to as *sensing* or *design* matrix.

3. A convergence rate is called near-optimal if it is optimal up to factors which are logarithmic in the condition number or dimension of the problem.

1.6.1 Literature overview

Sparse recovery optimization problems have gained a lot of attention in the literature from both practical and theoretical perspectives [Donoho et al., 2000, Bühlmann and Van De Geer, 2011]. In order to recover sparse solutions we have to change somehow the general problem statement and explicitly aim at exploiting its particular structure. In this introduction, we overview several ways to induce sparsity that are particularly important to us.

ℓ_1-minimization In order to exploit merits of convex optimization, one can use a relaxation of the ℓ_0 pseudonorm to its closest convex surrogate—the ℓ_1-norm (or a nuclear norm in the case of matrix regressors Φ). For the sparse linear regression, this approach leads to the celebrated Lasso estimator

$$\min_{x \in \mathbb{R}^p} \left\{ \frac{1}{2} \|y - \Phi x\|^2 + \lambda \|x\|_1 \right\}. \tag{1.33}$$

and Dantzig Selector, see [Candes et al., 2006, 2007, Bickel et al., 2009, Van De Geer and Bühlmann, 2009, Candès et al., 2009, Candes and Plan, 2011a, Raskutti et al., 2010, Fazel et al., 2008, Candès and Recht, 2009, Candes and Plan, 2011b, Juditsky and Nemirovski, 2011b, Negahban and Wainwright, 2011, Koltchinskii et al., 2011, Rudelson and Zhou, 2012, Dalalyan and Thompson, 2019].

Iterative thresholding In order to enforce sparsity of solutions, one can directly use the ℓ_0 pseudonorm either in the form of penalty or constraints. This essentially leads to non-convex optimization problems. A standard technique to treat such problems is *iterative hard thresholding* [Blumensath and Davies, 2009, Jain et al., 2014, Bhatia et al., 2015, Liu et al., 2019, 2020]. At each iteration, this technique takes a gradient step and then applies subsequent sparsification of the resulted estimate. In other words, these algorithms apply the following general update rule

$$\begin{aligned} &\text{Gradient step:} \quad x'_k = x_{k-1} - \eta_k \nabla f_{\text{SAA}}(x_{k-1}), \\ &\text{Sparsification:} \quad x_k = x'_{k-1} \text{ with } s \text{ components largest in magnitudes left.} \end{aligned} \tag{1.34}$$

[Foygel Barber and Liu, 2019, Zhao and Luo, 2019] replace hard thresholding with a less conservative operator in order to improve convergence (see also references in [Zhao and Luo, 2019]). When applied to sparse linear regression, the convergence bounds for considered algorithms are typically of the following form and hold with high probability

$$f_{\text{SAA}}(x_t) - f^*_{\text{SAA}} \leq \mathcal{O}(1) \left(1 - \frac{\mu}{L}\right)^t \frac{L}{2} \|x_0 - x^*\|^2 + \mathcal{O}(1) \left(\frac{L}{\mu}\right)^\gamma \frac{\sigma^2 s \log(p)}{N}, \tag{1.35}$$

where N is the number of columns in the sensing matrix D, and γ is a positive value. Thus, in terms of prediction error, they obtain a linear convergence to some inevitable statistical error, which is at least $\mathcal{O}((L/\mu)s\sigma^2/n)$ with n being the number of rows in D. This statistical error $\mathcal{O}((L/\mu)s\sigma^2/n)$ coincides with classical bounds for Lasso problem [Bickel et al., 2009, Zhang et al., 2014].

The approaches based on iterative thresholding are SAA techniques that optimize (1.32). They operate on a sensing matrix and require full gradient computation at each iteration. This makes them computationally expensive in large-scale setting unless $s \ll p$.

Sparse recovery by stochastic approximation In our thesis, we focus on first-order methods of SA type aimed to solve (1.31) in large-scale setting. Therefore, we consider algorithms with convergence bounds which are "essentially independent" (logarithmic, at most) in the problem dimension p. This requirement rules out the use of standard Euclidean stochastic approximation algorithms. Indeed, typical bounds for the expected risk $\mathbb{E}\left[f(x_N) - f^*\right]$ in SA contains the term proportional to $\sigma^2\mathbb{E}\left[\|\phi_1\|_2^2\right]$ and thus proportional to p in the case of dense regressors with $\mathbb{E}\left[\|\phi_1\|_2^2\right] = \mathcal{O}(p)$. For example, [Bietti and Mairal, 2017] succeed in derivation of a dimension-free bound for the variance convergence at the cost of assuming that the regressors ϕ_1 are bounded in the ℓ_2-norm. Therefore, unless the regressors ϕ are sparse (or possess a special structure, e.g., when ϕ_i are low rank matrices in the case of low rank matrix recovery), standard stochastic approximation techniques have accuracy bounds which are proportional to p [Nguyen et al., 2017b]. In other words, our interest is in non-Euclidean stochastic approximation procedures, such as the stochastic mirror descent algorithm.

In particular, different forms of the SMD were extensively used in ℓ_1 minimization for deterministic and stochastic objectives. Authors of [Srebro et al., 2010] establish the following asymptotic bound for risks

$$\mathbb{E}\left[f(x_N) - f^*\right] \leq \mathcal{O}\left(\sigma\sqrt{s\log(d)/N}\right) \tag{1.36}$$

in the sparse generation model. The obtained rate of convergence is often referred to as "slow rate", besides this, the considered algorithm does not yield the linear bias convergence like in (1.35). A similar rate was obtained in [Shalev-Shwartz and Tewari, 2011] for the problem (1.33), that is

$$\mathbb{E}\left[f(x_N) - f^*\right] \leq \mathcal{O}\left(\eta s\sqrt{\log(d)/N}\right), \tag{1.37}$$

where η is the magnitude of the gradient of the objective.

Local Strong Convexity In order to improve on slow rates of [Srebro et al., 2010, Shalev-Shwartz and Tewari, 2011], one may use strong or uniform convexity of the problem if there is one [Juditsky and Nemirovski, 2011a, Ghadimi and Lan, 2013, Juditsky and Nesterov, 2014]. However, such assumptions do not generally hold in the problems such as sparse linear regression problem. More generally, strong convexity of the objective associated with smoothness is a feature of the Euclidean setup. For instance, the conditioning of a smooth objective (the ratio of the Lipschitz constant of the gradient to the constant of strong convexity) when measured with respect to the ℓ_1-norm cannot be less than p [Juditsky and Nesterov, 2014]. Therefore, this notion is always replaced by various "local" conditions [Negahban et al., 2012, Xiao and Zhang, 2013, Nguyen et al., 2014, Loh and Wainwright, 2015, Foygel Barber and Liu, 2019, Murata and Suzuki, 2018].

In particular, strong convexity argument is replaced with Restricted (or Local) Strong Convexity condition.

Definition 1.7 (Restricted Strong Convexity)**.** *A function $f : X \to \mathbb{R}$ defined on a convex set X satisfies an R-restricted form of strong convexity with constant $\mu = \mu(R)$ if the following bound holds*

$$f(x_1) \geq f(x_2) + \langle \nabla f(x_2), x_1 - x_2 \rangle + \frac{\mu}{2} \|x_2 - x_1\|_2^2, \tag{1.38}$$

for any $x_1, x_2 \in X$ with $\|x_1\|_1 \leq R$ and $\|x_2\|_1 \leq R$.

Condition (1.38) is equivalent to the standard definition of strong convexity (1.7) when $R = \infty$. Definition 1.7 may slightly change from paper to paper, but the meaning remains the same: strong convexity is confined to some specific subset of the feasible set, and the constant μ becomes a function of the size of this set.

Multi-stage procedures The multi-stage procedures mentioned in Section 1.5 are also utilized in sparse recovery. In order to improve the convergence from $\mathcal{O}\left(s^{1 \text{ or } 1/2}/\sqrt{N}\right)$ to $\mathcal{O}\left(s\sigma^2/N\right)$ for (locally) strongly convex objectives, the authors in [Agarwal et al., 2012b, Steinhardt et al., 2014, Sedghi et al., 2014] propose an approach similar in spirit to multi-stage procedures of [Hazan and Kale, 2010, Juditsky and Nesterov, 2010, Ghadimi and Lan, 2013]. In this approach, one iteratively forms a sequence of ℓ_1 regularized objectives

$$\left(\min_{x \in \Omega_i} \left\{ f_i(x) \triangleq f(x) + \lambda_i \|x\|_1 \right\} \right)_{i=1}^{t}$$

with decreasing penalties $(\lambda_i)_{i \geq 1}$ and some constraint sets $(\Omega_i)_{i \geq 1}$ with updated proximal centers and decreasing radii. At each stage of the method, the new objective $F_i(x)$ is minimized using the stochastic dual averaging algorithm [Nesterov, 2009]. Comparing to the hard thresholding approaches, [Agarwal et al., 2012b] do not use sparsification steps and do not try to recover the solution, which is *exactly* sparse. As mentioned above, the authors also assume the restricted versions of strong convexity along with smoothness of the objective and sub-Gaussian stochastic gradients. At the end, they present asymptotic convergence guarantees, which resume to

$$\mathbb{E}\left[\|x_N - x^*\|_2^2\right] \leq \mathcal{O}\left(\frac{sB^2\sigma^2}{\mu^2 N}\right) \tag{1.39}$$

for the case when the regressors ϕ_i are bounded, $\|\phi\|_\infty \leq B$; this rate is known to be unimprovable in a certain setting [Raskutti et al., 2009]. Notice that this result appears only in one section of [Agarwal et al., 2012b], while the rest of the paper uses the following definition of the variance

$$\forall x \in X \quad \mathbb{E}\|g(x) - \nabla f(x)\|_\infty^2 \leq \hat{\sigma}^2 \gg \sigma^2, \tag{1.40}$$

where $g(x)$ is a stochastic gradient. This is a more standard definition of the noise variance in the literature on stochastic approximation, including the aforementioned works on

multi-stage procedures. However, in general, the value of $\hat{\sigma}$ is proportional to the diameter of the feasible set and may be much larger than σ^2 (this will be explained in more details in Section 4.1 of Chapter 4).

It is important to note that the admissible values of s in [Agarwal et al., 2012b] are bounded with $\mathcal{O}\left(\mu\sqrt{N/\log p}\right)$. Indeed, their method requires to perform at least $s^2 \log p/\mu^2$ iterations per stage, implying that the method in question can be used only if the number of nonvanishing entries in the parameter vector does not exceed $\mathcal{O}\left(\mu\sqrt{N/\log p}\right)$. However, the corresponding limit is $\mathcal{O}\left(N\mu/\log p\right)$ for Lasso [Raskutti et al., 2010] and iterative thresholding procedures [Barber and Ha, 2018, Foygel Barber and Liu, 2019]). Therefore, there is a gap for improvement of the condition $s \leq \mathcal{O}\left(\mu\sqrt{N/\log p}\right)$ to become $s \leq \mathcal{O}\left(N\mu/\log p\right)$.

Chapter 2

Estimate Sequences for Stochastic Optimization

In Section 1.4.4 we overviewed various algorithms based on variance reduction technique applied to deterministic objectives with finite-sum structure. The success of these algorithms is partly related to the possibility of keeping the step size constant while still being able to decrease the variance of the gradient estimations "on the fly". As we saw in Section 1.4.4, there is a limitation of variance-reduced algorithms to objectives with a finite-sum structure when the information about the terms $f_i(x)$ is perturbed by stochastic noise. Another open question concerns the acceleration of these methods when the objective is strongly convex.

In this chapter, we address these questions one by one. At the core of our framework lies the concept of estimate sequences introduced by [Nesterov, 2014]. For a long time, the acceleration via this framework was known to be unstable in the stochastic case. By extending this concept to minimization of stochastic objectives, we propose a unified view of first-order methods for *stochastic* convex composite optimization. More precisely, we interpret a large class of incremental approaches as procedures that iteratively minimize some surrogate of the objective. As a result, we cover the stochastic gradient descent algorithm and several incremental variance-reduced approaches like SAGA, SVRG, MISO, Finito and SDCA. This point of view has several advantages: (i) we provide a simple generic proof of convergence for all of the aforementioned methods; (ii) we naturally obtain new algorithms with the same guarantees as the existing incremental methods; (iii) we derive generic strategies to make these algorithms robust to stochastic noise, which is useful when data is corrupted by small random perturbations. Finally, we propose a new accelerated stochastic gradient descent algorithm and an accelerated SVRG algorithm with the optimal complexity that is robust to stochastic noise.

This chapter is based on the following publications:

- A. Kulunchakov and J. Mairal. Estimate sequences for variance-reduced stochastic composite optimization. In *Proceedings of the International Conference on Machine Learning (ICML)*, June 2019c
- A. Kulunchakov and J. Mairal. Estimate sequences for stochastic composite optimization: Variance reduction, acceleration, and robustness to noise. *Journal of Machine Learning Research (JMLR)*, 2020

2.1 Introduction

In order to be precise, we restate the exact optimization problem that we aim to solve for each chapter throughout the thesis. In this chapter, we minimize objective functions of the following form

$$\min_{x\in\mathbb{R}^p}\left\{F(x)\triangleq f(x)+\psi(x)\triangleq\frac{1}{n}\sum_{i=1}^{n}f_i(x)+\psi(x)\right\}, \tag{2.1}$$

where f is convex and L-smooth, and we call μ its strong convexity modulus with respect to the Euclidean norm (if such μ is positive). The function ψ is convex lower semi-continuous and is not assumed to be necessarily differentiable, with possible practical examples of this penalty function given in Section 1.2.2. Each term f_i is convex and L_i-smooth with $L_i \geq \mu$. Moreover, $(f_i)_{i=1}^n$ are assumed to be given in the form of expectation $f_i = \mathbb{E}_{\rho_i}\left[\tilde{f}_i(x,\rho_i)\right]$, so that one accesses only inexact gradient information for each of them unless the variance of the gradient estimate is zero. This statement (2.1) gives a general setting which embodies both minimization of the expected risk when $n = 1$ and and the classical empirical risk when ρ_i is deterministic. Therefore, the optimal convergence rate for solving (2.1) can not be better than (1.12) for strongly convex problems and (1.11) for non-strongly convex ones.

Case with exact gradients Assume that we are in a high-dimensional setting when either dimensionality of the problem p or the number of terms n is large. Then, the optimization methods querying the full gradient of $\nabla f(x)$ are not applicable and one standard approach is to operate on the gradients of the terms $\nabla f_i(x)$ instead. First, we focus on the deterministic case so that we have access to the true gradient estimations $\nabla f_i(x)$ for each i. In this case, the stochastic gradient descent algorithm with step size proportional to $1/\mu N$ enjoys the asymptotic convergence rate $\mathcal{O}(1/N)$ when $n = 1$. Even though this pessimistic result applies to the general stochastic case, linear convergence rates are obtained for the deterministic finite-sum setting in [Schmidt et al., 2017] and plethora of other incremental variants of the SGD algorithm. For example, these methods include SAG [Schmidt et al., 2017], SAGA [Defazio et al., 2014a], SVRG [Johnson and Zhang, 2013, Xiao and Zhang, 2014], SDCA [Shalev-Shwartz and Zhang, 2016], MISO [Mairal, 2015], Katyusha [Allen-Zhu, 2017], MiG [Zhou et al., 2018], SARAH [Nguyen et al., 2017a], directly accelerated SAGA [Zhou, 2019] or RPDG [Lan and Zhou, 2018a]. All these algorithms admit the same per-iteration cost as the stochastic gradient descent method, since they access only a single (or two) gradients $\nabla f_i(x)$ at each iteration on average.

Moreover, sometimes they may achieve lower computational complexity than that of accelerated gradient descent methods [Nesterov, 1983, 2013, 2014, Beck and Teboulle, 2009] in expectation, by exploiting the specific structure of the objective function. In this chapter, one contribution is a framework that provides a unified view on many of the aforementioned methods.

Case with inexact gradients One can weaken the assumption of being deterministic imposed on the terms f_i, and suggest that they are given in the form of (unknown) expectation $f_i = \mathbb{E}_{\rho_i}\left[\tilde{f}_i(x, \rho_i)\right]$. As a result, one has access only to noisy estimates of the true gradients $\nabla f_i(x)$ for each i. As we saw in Section 1.4.5, this setting is common for several applications.

Our objective is to investigate *robustness* to this inexactness in the gradient information for the aforementioned variance-reduced approaches SVRG/SAGA/SDCA/MISO. The original versions of these methods do not apply to this setting anymore, at least from theoretical point of view, so that the standard approach to address (2.1) was to ignore the finite-sum structure and use SGD or one of its variants.

In order to establish robustness, we adopt the concept of estimate sequences introduced in [Nesterov, 2014], which consists of building iteratively a quadratic model of the objective. Typically, estimate sequences are used to analyze the convergence of existing algorithms or to design new ones, in particular with acceleration. Our construction is however slightly different than the original since it is based on stochastic estimates of the gradients. We note that estimate sequences have been used before for stochastic optimization [Devolder et al., 2011, Lu and Xiao, 2015, Lin et al., 2014], but not for the same generic purpose as ours.

Additionally, we analyze the aforementioned approaches under a non-uniform sampling strategy $Q = \{q_1, \dots, q_n\}$ where q_i is the probability of drawing example i at each iteration, the strategy that is quite common in literature. Typically, when the gradients ∇f_i have different Lipschitz constants L_i, the uniform distribution Q of the sampling yields complexities that depend on $L_Q = \max_i L_i$, whereas a non-uniform Q may yield a smaller quantity $L_Q = \frac{1}{n}\sum_i L_i$. This gain comes at a price of a larger variance convergence, which is essentially multiplied by $\rho_Q = 1/(n \min q_i) \geq 1$. Whereas non-uniform sampling strategies for incremental methods are now classical [Xiao and Zhang, 2014, Schmidt et al., 2015], the robustness to stochastic perturbations has not been studied for all these methods and existing approaches such as [Hofmann et al., 2015, Bietti and Mairal, 2017, Zheng and Kwok, 2018] have various limitations as discussed earlier in Section 1.4.4.

Finally, when making the aforementioned algorithms robust to stochastic perturbations, we aim at obtaining the following worst-case iteration complexity for solving (2.1) up to accuracy ε

$$\mathcal{O}\left(\left(n + \frac{L_Q}{\mu}\right)\log\left(\frac{F(x_0) - F^*}{\varepsilon}\right)\right) + \mathcal{O}\left(\frac{\rho_Q \tilde{\sigma}^2}{\mu\varepsilon}\right), \tag{2.2}$$

where $L_Q = \max_i L_i/(q_i n)$ and $\rho_Q = 1/(n \min q_i) \geq 1$. The term on the left corresponds to the complexity of the variance-reduction methods in deterministic setting, and $\mathcal{O}\left(\tilde{\sigma}^2/\mu\varepsilon\right)$ is the desired sublinear rate of convergence for stochastic finite-sums optimization problems

when the gradient estimates for each f_i have the variance bounded with $\tilde{\sigma}^2$.[1] Our next contribution is a derivation of *accelerated* stochastic algorithms. In this direction, we first show that our construction of estimate sequences naturally leads to an accelerated stochastic gradient method with the iteration complexity for μ-strongly convex optimization problems

$$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{F(x_0)-F^*}{\varepsilon}\right)\right)+\mathcal{O}\left(\frac{\sigma^2}{\mu\varepsilon}\right),$$

which was also achieved by [Ghadimi and Lan, 2013, Cohen et al., 2018, Aybat et al., 2019]. When the objective is convex, but not strongly convex, we provide a sublinear convergence rate for a finite horizon setting. Given a budget of K iterations, our algorithm returns an iterate x_K such that

$$\mathbb{E}\left[F(x_K)-F^*\right]\leq\frac{2L\left\|x_0-x^*\right\|^2}{(K+1)^2}+\sigma\sqrt{\frac{8\left\|x_0-x^*\right\|^2}{K+1}}. \tag{2.3}$$

This convergence rate is optimal for stochastic first-order convex optimization [Lan, 2012]. Second, we address the acceleration of variance-reduced algorithms and design a new accelerated algorithm for minimization of large limit sums of stochastic functions based on the SVRG gradient estimator, with complexity, for μ-strongly convex functions,

$$\mathcal{O}\left(\left(n+\sqrt{n\frac{L_Q}{\mu}}\right)\log\left(\frac{F(x_0)-F^*}{\varepsilon}\right)\right)+\mathcal{O}\left(\frac{\rho_Q\tilde{\sigma}^2}{\mu\varepsilon}\right), \tag{2.4}$$

where the term on the left is the classical optimal complexity for deterministic finite sums optimization, which has been well studied when $\tilde{\sigma}^2=0$ [Arjevani and Shamir, 2016, Allen-Zhu, 2017, Zhou et al., 2018, Lan and Zhou, 2018a, Zhou, 2019, Kovalev et al., 2020]. To the best of our knowledge, our algorithm is nevertheless the first to achieve such a complexity when $\tilde{\sigma}^2>0$. Most related to our work, the general case $\tilde{\sigma}^2>0$ was considered by [Lan and Zhou, 2018b] in the context of distributed optimization, with an approach that was shown to be optimal in terms of communication rounds. Yet, when applied in the same context as ours (in a non-distributed setting), the complexity they achieve is sub-optimal. Specifically, their dependence in $\tilde{\sigma}^2$ involves an additional logarithmic factor $\mathcal{O}\left(\log\left(1/\mu\varepsilon\right)\right)$ and the deterministic part is sublinear in $\mathcal{O}\left(1/\varepsilon\right)$.

When the problem is convex but not strongly convex, given a budget of K greater than $\mathcal{O}\left(n\log(n)\right)$, our algorithm returns a solution x_K such that

$$\mathbb{E}\left[F(x_K)-F^*\right]\leq\frac{18nL_Q\left\|x_0-x^*\right\|^2}{(K+1)^2}+9\tilde{\sigma}\left\|x_0-x^*\right\|\sqrt{\frac{\rho_Q}{K+1}}, \tag{2.5}$$

where the term on the right is potentially better than the same in (2.3) when $\tilde{\sigma}\ll\sigma$. When the objective is deterministic ($\tilde{\sigma}=0$), the term (2.5) yields the complexity $\mathcal{O}\left(\sqrt{nL_Q}/\sqrt{\varepsilon}\right)$, which is potentially better than the $\mathcal{O}\left(n\sqrt{L}/\sqrt{\varepsilon}\right)$ complexity of accelerated gradient descent, unless L is significantly smaller than L_Q.

Now let us finally summarize all contributions of this chapter in a compact list.

1. The more accurate expression for $\tilde{\sigma}^2$ will be given in (2.18) of Section 2.3.2.

2.1.1 Contributions of Chapter 2

- We revisit many incremental optimization algorithms, like SGD, SVRG, SAGA, SDCA, MISO and Finito, and provide a unified convergence proof for these methods. In addition, we show that they can be modified and become adaptive to the strong convexity constant μ, which may be important in applications where μ is hard to estimate.
- We improve these methods by making them robust to stochastic noise in the terms f_i. For strongly convex problems, we develop approaches with the following worst-case iteration complexity for minimizing (2.1) up to accuracy ε

$$\mathcal{O}\left(\left(n+\frac{L_Q}{\mu}\right)\log\left(\frac{F(x_0)-F^*}{\varepsilon}\right)\right)+\mathcal{O}\left(\frac{\rho_Q\tilde{\sigma}^2}{\mu\varepsilon}\right),$$

  where $L_Q=\max_i L_i/(q_i n)$ and $\rho_Q=1/(n\min q_i)\geq 1$.
- We show that our construction of estimate sequence naturally leads to an accelerated stochastic gradient method with the following complexity for μ-strongly convex objectives

$$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{F(x_0)-F^*}{\varepsilon}\right)\right)+\mathcal{O}\left(\frac{\sigma^2}{\mu\varepsilon}\right).$$

  When the objective is convex, but not strongly convex, we achieve an optimal sublinear convergence rate for a finite horizon setting

$$\mathbb{E}\left[F(x_K)-F^*\right]\leq\frac{2L\left\|x_0-x^*\right\|^2}{(K+1)^2}+\sigma\sqrt{\frac{8\left\|x_0-x^*\right\|^2}{K+1}}.$$

- We design a new accelerated algorithm for minimization of finite sums based on the SVRG gradient estimator, with complexity, for μ-strongly convex functions,

$$\mathcal{O}\left(\left(n+\sqrt{n\frac{L_Q}{\mu}}\right)\log\left(\frac{F(x_0)-F^*}{\varepsilon}\right)\right)+\mathcal{O}\left(\frac{\rho_Q\tilde{\sigma}^2}{\mu\varepsilon}\right),$$

  To the best of our knowledge, our algorithm is the first to achieve such a complexity when $\tilde{\sigma}^2>0$.
  When the problem is convex but not strongly convex, given a budget of K greater than $\mathcal{O}\left(n\log(n)\right)$, the algorithm returns a solution x_K such that

$$\mathbb{E}\left[F(x_K)-F^*\right]\leq\frac{18nL_Q\left\|x_0-x^*\right\|^2}{(K+1)^2}+9\tilde{\sigma}\left\|x_0-x^*\right\|\sqrt{\frac{\rho_Q}{K+1}}.$$

The rest of the chapter is organized as follows. Section 2.2 introduces the proposed framework based on stochastic estimate sequences; Section 2.3 presents a convergence analysis and Section 2.4 introduces accelerated stochastic optimization algorithms; Section 2.5 contains various experiments demonstrating the effectiveness of the proposed approaches, and Section 2.6 concludes the chapter.

2.2 Framework Based on Stochastic Estimate Sequences

In this section, we present two generic stochastic optimization algorithms to address the problem (2.1). Both algorithms are related to estimate sequence framework of [Nesterov, 2014] and could be seen as its generalization to the stochastic setting. After the algorithms are introduced, we will show their relation to variance-reduction methods.

2.2.1 A Classical Iteration Revisited

Consider the stochastic gradient descent that performs the following updates:

$$x_k \leftarrow \mathrm{Prox}_{\eta_k \psi}\left[x_{k-1} - \eta_k g_k\right] \quad \text{with} \quad \mathbb{E}\left[g_k | \mathcal{F}_{k-1}\right] = \nabla f(x_{k-1}), \tag{A}$$

where $\mathcal{F}_{k-1}$ is the filtration representing all information up to iteration $k-1$, g_k is an unbiased estimate of the gradient $\nabla f(x_{k-1})$, $\eta_k > 0$ is a step size, and $\mathrm{Prox}_{\eta\psi}[.]$ is the proximal operator of Section 1.2.4 defined for any scalar $\eta > 0$ as the unique solution of

$$\mathrm{Prox}_{\eta\psi}[u] = \operatorname*{argmin}_{x \in \mathbb{R}^p} \left\{ \eta\psi(x) + \frac{1}{2}\left\|x - u\right\|^2 \right\}. \tag{2.6}$$

The iteration (A) of stochastic gradient descent is generic and encompasses many existing algorithms, which basically differ in the way of g_k is constructed. These algorithms will be reviewed later. Key to our analysis, we are interested in a simple interpretation of this update rule as iterative minimization of strongly convex surrogate functions.

Interpretation with stochastic estimate sequence. Consider now the function

$$d_0(x) = d_0^* + \frac{\gamma_0}{2}\|x - x_0\|^2, \tag{2.7}$$

with $\gamma_0 \geq \mu$ and d_0^* is a scalar value that is left unspecified at the moment. Then, it is easy to show that x_k in (A) minimizes the following quadratic function d_k defined for $k \geq 1$ as

$$\begin{aligned} d_k(x) = (1 - \delta_k)\, d_{k-1}(x) + \\ \delta_k \left(f(x_{k-1}) + g_k^\top (x - x_{k-1}) + \frac{\mu}{2}\left\|x - x_{k-1}\right\|^2 + \psi(x_k) + \psi'(x_k)^\top (x - x_k) \right), \end{aligned} \tag{2.8}$$

where δ_k, γ_k satisfy the system of equations

$$\delta_k = \eta_k \gamma_k \quad \text{and} \quad \gamma_k = (1 - \delta_k)\,\gamma_{k-1} + \mu\delta_k, \tag{2.9}$$

and

$$\psi'(x_k) = \frac{1}{\eta_k}(x_{k-1} - x_k) - g_k.$$

We note that $\psi'(x_k)$ is a subgradient in $\partial\psi(x_k)$, see Definition 1.3. By simply using the definition of the proximal operator (2.6) and considering first-order optimality conditions,

we indeed have that $0 \in x_k - x_{k-1} + \eta_k g_k + \eta_k \partial\psi(x_k)$ and x_k coincides with the minimizer of d_k. This allows us to write d_k in the generic form

$$d_k(x) = d_k^* + \frac{\gamma_k}{2} \|x - x_k\|^2 \quad \text{for all } k \geq 0.$$

The construction (2.8) is akin to that of estimate sequences introduced by Nesterov [2014], which are typically used for designing accelerated gradient-based optimization algorithms, as we have overviewed in Section 1.5. In this section, we are not interested in acceleration for now, but instead we focus on stochastic optimization and variance reduction. One of the main properties of estimate sequences that we will use is their ability do behave asymptotically as a lower bound of the objective function near the optimum, like was shown in Lemma 2.2.1 of Nesterov [2014] for the deterministic case $\sigma = 0$. Indeed, we have

$$\mathbb{E}\left[d_k(x^*)\right] \leq (1 - \delta_k)\, \mathbb{E}\left[d_{k-1}(x^*)\right] + \delta_k F^* \leq \Gamma_k d_0(x^*) + (1 - \Gamma_k)\, F^*, \tag{2.10}$$

where $\Gamma_k = \prod_{t=1}^k (1 - \delta_t)$ and $F^* = F(x^*)$. The first inequality comes from a strong convexity argument since $\mathbb{E}\left[g_k^\top (x^* - x_{k-1}) | \mathcal{F}_{k-1}\right] = \nabla f(x_{k-1})^\top (x^* - x_{k-1})$, and the second inequality is obtained by unrolling the relation obtained between $\mathbb{E}\left[d_k(x^*)\right]$ and $\mathbb{E}\left[d_{k-1}(x^*)\right]$. When Γ_k converges to zero, the contribution of the initial surrogate d_0 disappears and $\mathbb{E}\left[d_k(x^*)\right]$ behaves as a lower bound of F^*.

Relation with existing algorithms. Now we can overview the body of algorithms covered by the iteration (A). First, it encompasses such approaches as ISTA (proximal gradient descent), which uses the exact gradient $g_k = \nabla f(x_{k-1})$ leading to deterministic iterates $(x_k)_{k \geq 0}$ [Beck and Teboulle, 2009, Nesterov, 2013] or proximal variants of the stochastic gradient descent method to deal with a composite objective [see Lan, 2012, for instance]. Second, and of particular interest for us, the variance-reduced stochastic optimization approaches SVRG [Xiao and Zhang, 2014] and SAGA [Defazio et al., 2014a] also follow the iteration (A) with a specific unbiased gradient estimator g_k whose variance decreases over time. Specifically, the basic form of these estimators is

$$g_k = \nabla f_{i_k}(x_{k-1}) - z_{k-1}^{i_k} + \bar{z}_{k-1} \quad \text{with} \quad \bar{z}_{k-1} = \frac{1}{n} \sum_{i=1}^n z_{k-1}^i, \tag{2.11}$$

where i_k is an index chosen uniformly in $\{1, \ldots, n\}$ at random, and each auxiliary variable z_k^i is equal to the gradient $\nabla f_i(\tilde{x}_k^i)$, where $\tilde{x}_k^i$ is one of the previous iterates. The motivation is that given two random variables X and Y, it is possible to define a new variable $Z = X - Y + \mathbb{E}\left[Y\right]$ which has the same expectation as X but potentially a lower variance if Y is positively correlated with X. SVRG and SAGA are two different approaches to build such positively correlated variables. SVRG uses the same anchor point $\tilde{x}_k^i = \tilde{x}_k$ for all i, where $\tilde{x}_k$ is updated every m iterations. Typically, the memory cost of SVRG is that of storing the variable $\tilde{x}_k$ and the gradient $\bar{z}_k = \nabla f(\tilde{x}_k)$, which is thus $\mathcal{O}\left(p\right)$. On the other hand, SAGA updates only $z_k^{i_k} = \nabla f_{i_k}(x_{k-1})$ at iteration k, such that $z_k^i = z_{k-1}^i$ if $i \neq i_k$. Thus, SAGA requires storing n gradients. While in general the overhead cost in memory is of order $\mathcal{O}\left(np\right)$, it may be reduced to $\mathcal{O}\left(n\right)$ when dealing with linear models in machine

learning [see Defazio et al., 2014a]. Note that variants with non-uniform sampling of the indices i_k have been proposed by Xiao and Zhang [2014], Schmidt et al. [2015].

In order to make our proofs consistent for all considered incremental methods, we analyze a variant of SVRG with a randomized gradient updating schedule [Hofmann et al., 2015]. Remarkably, this variant was recently used in a concurrent work [Kovalev et al., 2020] to get the accelerated rate when $\tilde{\sigma}^2 = 0$.

2.2.2 A Less Classical Iteration with a Different Estimate Sequence

In the previous section, we have interpreted the classical iteration (A) as the iterative minimization of the stochastic surrogate (2.8). Here, we show that a slightly different construction leads to a new algorithm. To obtain a lower bound, we have indeed used basic properties of the proximal operator to obtain a subgradient $\psi'(x_k)$ and we have exploited the following convexity inequality $\psi(x) \geq \psi(x_k) + \psi'(x_k)^\top (x - x_k)$. Another natural choice to build a lower bound consists then of using directly $\psi(x)$ instead of $\psi(x_k) + \psi'(x_k)^\top (x - x_k)$, leading to the construction

$$d_k(x) = (1 - \delta_k)\, d_{k-1}(x) + \delta_k \left(f(x_{k-1}) + g_k^\top (x - x_{k-1}) + \frac{\mu}{2} \|x - x_{k-1}\|^2 + \psi(x) \right), \tag{2.12}$$

where x_{k-1} is assumed to be the minimizer of the composite function d_{k-1}, δ_k is defined as in Section 2.2.1, and x_k is a minimizer of d_k. To initialize the recursion, we define then d_0 as

$$d_0(x) = c_0 + \frac{\gamma_0}{2} \|x - \bar{x}_0\|^2 + \psi(x) \geq d_0^* + \frac{\gamma_0}{2} \|x - x_0\|^2,$$

with $x_0 = \mathrm{Prox}_{\psi/\gamma_0}[\bar{x}_0]$ is the minimizer of d_0 and $d_0^* = d_0(x_0) = c_0 + \frac{\gamma_0}{2} \|x_0 - \bar{x}_0\|^2 + \psi(x_0)$ is the minimum value of d_0; c_0 is left unspecified since it does not affect the algorithm. Typically, one may choose $\bar{x}_0$ to be a minimizer of ψ such that $x_0 = \bar{x}_0$. Unlike in the previous section, the surrogates d_k are not quadratic, but they remain γ_k-strongly convex. It is also easy to check that the relation (2.10) still holds.

The corresponding algorithm. It is also relatively easy to show that the iterative minimization of the stochastic lower bounds (2.12) leads to the following iterations

$$\bar{x}_k \leftarrow (1 - \mu\eta_k)\bar{x}_{k-1} + \mu\eta_k x_{k-1} - \eta_k g_k \quad \text{and} \quad x_k = \mathrm{Prox}_{\psi/\gamma_k}[\bar{x}_k] \tag{B}$$

where again g_k is an unbiased gradient estimator $\mathbb{E}[g_k | \mathcal{F}_{k-1}] = \nabla f(x_{k-1})$. As we will see, the convergence analysis for algorithm (A) also holds for algorithm (B) such that both variants enjoy similar theoretical properties. In one case, the function $\psi(x)$ appears explicitly, whereas a lower bound $\psi(x_k) + \psi'(x_k)^\top (x - x_k)$ is used in the other case. The introduction of the variable $\bar{x}_k$ allows us to write the surrogates d_k in the canonical form

$$d_k(x) = c_k + \frac{\gamma_k}{2} \|x - \bar{x}_k\|^2 + \psi(x) \geq d_k^* + \frac{\gamma_k}{2} \|x - x_k\|^2,$$

where c_k is constant and the inequality on the right is due to the strong convexity of d_k.

Relation to existing approaches. The approach (B) is related to several optimization methods. When the objective is a deterministic finite sum, it is possible to relate the update (B) to the MISO [Mairal, 2015], and Finito [Defazio et al., 2014b] algorithms, even though they were derived from a significantly different point of view. This is also the case of a primal variant of SDCA [Shalev-Shwartz, 2016]. For instance, SDCA is a dual coordinate ascent approach, whereas MISO and Finito are explicitly derived from the iterative surrogate minimization we adopt in this chapter. As the links between (B) and these previous approaches are not obvious at first sight, we detail them in Appendix 2.B.

2.2.3 Gradient Estimators and Algorithms

In this chapter, we consider the iterations (A) and (B) with the following gradient estimators.

— **exact gradient** with $g_k = \nabla f(x_{k-1})$, when the problem is deterministic and we have access to the full gradient;

— **stochastic gradient**, when we simply assume that g_k has bounded variance. Typically, when $f(x) = \mathbb{E}_\xi[\tilde{f}(x,\xi)]$, a data point ξ_k is drawn at iteration k and $g_k = \tilde{\nabla} f(x, \xi_k)$.

— **random-SVRG**: for finite sums, we consider a variant of the SVRG gradient estimator with non-uniform sampling and a random update of the anchor point $\tilde{x}_{k-1}$, proposed originally by Hofmann et al. [2015]. Specifically, g_k is also an unbiased estimator of $\nabla f(x_{k-1})$, defined as

$$g_k = \frac{1}{q_{i_k} n} \left(\tilde{\nabla} f_{i_k}(x_{k-1}) - z_{k-1}^{i_k} \right) + \bar{z}_{k-1}, \tag{2.13}$$

where i_k is sampled from a distribution $Q = \{q_1, \dots, q_n\}$ and $\tilde{\nabla}$ denotes that the gradient is perturbed by a zero-mean noise variable with variance $\tilde{\sigma}^2$. More precisely, if $f_i(x) = \mathbb{E}_\rho[\tilde{f}_i(x,\rho)]$ for all i, where ρ is a stochastic perturbation, instead of accessing $\nabla f_{i_k}(x_{k-1})$, we draw a perturbation ρ_k and observe

$$\tilde{\nabla} f_{i_k}(x_{k-1}) = \nabla \tilde{f}_{i_k}(x_{k-1}, \rho_k) = \nabla f_{i_k}(x_{k-1}) + \underbrace{\nabla \tilde{f}_{i_k}(x_{k-1}, \rho_k) - \nabla f_{i_k}(x_{k-1})}_{\zeta_k},$$

where the perturbation ζ_k has zero mean given $\mathcal{F}_{k-1}$ and its variance is bounded by $\tilde{\sigma}^2$. When there is no perturbation, we simply have $\tilde{\nabla} = \nabla$ and $\zeta_k = 0$.
Then, the variables z_k^i and $\bar{z}_k$ also correspond to possibly noisy estimates of the gradients:

$$z_k^i = \tilde{\nabla} f_i(\tilde{x}_k) \quad \text{and} \quad \bar{z}_k = \frac{1}{n} \sum_{i=1}^n z_k^i,$$

where $\tilde{x}_k$ is an anchor point that is updated on average every n iterations. Whereas the classical SVRG approach [Xiao and Zhang, 2014] updates $\tilde{x}_k$ on a fixed schedule, we perform random updates: with probability $1/n$, we choose $\tilde{x}_k = x_k$ and recompute $\bar{z}_k = \tilde{\nabla} f(\tilde{x}_k)$; otherwise $\tilde{x}_k$ is kept unchanged. In comparison with the fixed schedule, the analysis with the random one is simplified and can be unified

Algorithm 2.1 Variant (A) with random-SVRG estimator

Input: x_0 in $\mathbb{R}^p$ (initial point); K (number of iterations); $(\eta_k)_{k\geq 0}$ (step sizes); $\gamma_0 \geq \mu$ (if averaging);
Initialization: $\tilde{x}_0 = \hat{x}_0 = x_0$; $\bar{z}_0 = \frac{1}{n}\sum_{i=1}^n \tilde{\nabla} f_i(\tilde{x}_0)$;
for $k = 1, \ldots, K$ **do**
    Sample i_k according to the distribution $Q = \{q_1, \ldots, q_n\}$;
    Compute the gradient estimator, possibly corrupted by random perturbations:

$$g_k = \frac{1}{q_{i_k} n}\left(\tilde{\nabla} f_{i_k}(x_{k-1}) - \tilde{\nabla} f_{i_k}(\tilde{x}_{k-1})\right) + \bar{z}_{k-1};$$

    Obtain the new iterate

$$x_k \leftarrow \mathrm{Prox}_{\eta_k \psi}\left[x_{k-1} - \eta_k g_k\right];$$

    With probability $1/n$,

$$\tilde{x}_k = x_k \quad \text{and} \quad \bar{z}_k = \frac{1}{n}\sum_{i=1}^n \tilde{\nabla} f_i(\tilde{x}_k);$$

    Otherwise, with probability $1 - 1/n$, keep $\tilde{x}_k = \tilde{x}_{k-1}$ and $\bar{z}_k = \bar{z}_{k-1}$;
    Optional: Use the online averaging strategy using δ_k obtained from (2.9):

$$\hat{x}_k = (1-\tau_k)\hat{x}_{k-1} + \tau_k x_k \quad \text{with} \quad \tau_k = \min\left(\delta_k, \frac{1}{5n}\right);$$

end for
Output: x_K or $\hat{x}_K$ if averaging.

with that of SAGA/SDCA or MISO. The use of this estimator with iteration (A) is illustrated in Algorithm 2.1. It is then easy to modify it to use variant (B) instead. In terms of memory, the random-SVRG gradient estimator requires to store an anchor point $\tilde{x}_{k-1}$ and the average gradients $\bar{z}_{k-1}$. The variables z_k^i do not need to be stored; only the n random seeds to produce the perturbations are kept into memory, which allows us to compute $z_{k-1}^{i_k} = \tilde{\nabla} f_{i_k}(\tilde{x}_{k-1})$ at iteration k, with the same perturbation for index i_k that was used to compute $\bar{z}_{k-1} = \frac{1}{n}\sum_{i=1}^n z_{k-1}^i$ when the anchor point was last updated. The overall cost is thus $\mathcal{O}(n+p)$.

— **SAGA**: The estimator has a form similar to (2.13) but with a different choice of variables z_k^i. Unlike SVRG that stores an anchor point $\tilde{x}_k$, the SAGA estimator requires storing and incrementally updating the n auxiliary variables z_k^i for $i = 1, \ldots, n$, while maintaining the relation $\bar{z}_k = \frac{1}{n}\sum_{i=1}^n z_k^i$. We consider variants such that each time a gradient $\nabla f_i(x)$ is computed, it is corrupted by a zero-mean random perturbation with variance $\tilde{\sigma}^2$. The procedure is described in Algorithm 2.2 for variant (A) when using uniform sampling. When $\beta = 0$, we recover indeed the original SAGA algorithm, whereas the choice $\beta > 0$ corresponds to a more general estimator that we will discuss next.

The case with non-uniform sampling is slightly different and is described in Algorithm 2.3; it requires an additional index j_k for updating a variable $z_k^{j_k}$. The reason for that is to remove a difficulty in the convergence proof, a strategy also adopted by Schmidt et al. [2015] for a variant of SAGA with non-uniform sampling.

— **SDCA/MISO**: To put SAGA, MISO and SDCA under the same umbrella, we introduce a lower bound β on the strong convexity constant μ, and a correcting term involving β that appears only when the sampling distribution Q is not uniform:

$$g_k = \frac{1}{q_{i_k} n}\left(\tilde{\nabla} f_{i_k}(x_{k-1}) - z_{k-1}^{i_k}\right) + \bar{z}_{k-1} + \beta\left(1 - \frac{1}{q_{i_k} n}\right) x_{k-1}. \tag{2.14}$$

It is then possible to show that when Q is uniform and under the big data condition $L/\mu \leq n$ (used for instanced by Mairal 2015, Defazio et al. 2014b, Schmidt et al. 2017) and with $\beta = \mu$, variant (B) combined with the estimator (2.14) yields the MISO algorithm, which performs similar updates as a primal variant of SDCA [Shalev-Shwartz, 2016]. These links are highlighted in Appendix 2.B.
The motivation for introducing the parameter β in $[0, \mu]$ comes from empirical risk minimization problems, where the functions f_i may have the form $f_i(x) = \phi(a_i^\top x) + \frac{\beta}{2}\|x\|^2$, where a_i in $\mathbb{R}^p$ is a data point; then, β is a lower bound on the strong convexity modulus μ, and $\nabla f_i(x) - \beta x$ is proportional to a_i and can be stored with a single additional scalar value, assuming a_i is already in memory.

Algorithm 2.2 Variant (A) with SAGA/SDCA/MISO estimator and uniform sampling

Input: x_0 in $\mathbb{R}^p$; K (number of iterations); $(\eta_k)_{k\geq 0}$ (step sizes); $\beta \in [0, \mu]$; if averaging, $\gamma_0 \geq \mu$.
Initialization: $z_0^i = \tilde{\nabla} f_i(x_0) - \beta x_0$ for all $i = 1, \ldots, n$ and $\bar{z}_0 = \frac{1}{n}\sum_{i=1}^n z_0^i$.
for $k = 1, \ldots, K$ **do**
  Sample i_k in $\{1, \ldots, n\}$ according to the uniform distribution;
  Compute the gradient estimator, possibly corrupted by random perturbations:

$$g_k = \tilde{\nabla} f_{i_k}(x_{k-1}) - z_{k-1}^{i_k} + \bar{z}_{k-1};$$

  Obtain the new iterate

$$x_k \leftarrow \mathrm{Prox}_{\eta_k \psi}\left[x_{k-1} - \eta_k g_k\right];$$

  Update the auxiliary variables

$$z_k^{i_k} = \tilde{\nabla} f_{i_k}(x_{k-1}) - \beta x_{k-1} \quad \text{and} \quad z_k^i = z_{k-1}^i \quad \text{for all} \quad i \neq i_k;$$

  Update the average variable $\bar{z}_k = \bar{z}_{k-1} + \frac{1}{n}(z_k^{j_k} - z_{k-1}^{j_k})$.
  Optional: Use the same averaging strategy as in Algorithm 2.1.
end for
Output: x_K or $\hat{x}_K$ (if averaging).

Algorithm 2.3 Variant (A) with SAGA/SDCA/MISO estimator and non-uniform sampling

Input: x_0 in $\mathbb{R}^p$; K (number of iterations); $(\eta_k)_{k\geq 0}$ (step sizes); $\beta \in [0, \mu]$; if averaging, $\gamma_0 \geq \mu$.
Initialization: $z_0^i = \tilde{\nabla} f_i(x_0) - \beta x_0$ for all $i = 1, \ldots, n$ and $\bar{z}_0 = \frac{1}{n}\sum_{i=1}^n z_0^i$.
for $k = 1, \ldots, K$ **do**
    Sample i_k according to the distribution $Q = \{q_1, \ldots, q_n\}$;
    Compute the gradient estimator, possibly corrupted by random perturbations:

$$g_k = \frac{1}{q_{i_k} n}\left(\tilde{\nabla} f_{i_k}(x_{k-1}) - z_{k-1}^{i_k}\right) + \bar{z}_{k-1} + \beta\left(1 - \frac{1}{q_{i_k} n}\right) x_{k-1};$$

    Obtain the new iterate

$$x_k \leftarrow \mathrm{Prox}_{\eta_k \psi}\left[x_{k-1} - \eta_k g_k\right];$$

    Draw j_k from the uniform distribution in $\{1, \ldots, n\}$;
    Update the auxiliary variables

$$z_k^{j_k} = \tilde{\nabla} f_{j_k}(x_k) - \beta x_k \quad \text{and} \quad z_k^j = z_{k-1}^j \quad \text{for all} \quad j \neq j_k;$$

    Update the average variable $\bar{z}_k = \bar{z}_{k-1} + \frac{1}{n}(z_k^{j_k} - z_{k-1}^{j_k})$.
    Optional: Use the same averaging strategy as in Algorithm 2.1.
end for
Output: x_K or $\hat{x}_K$ (if averaging).

Summary of the new features. As we combine different types of iterations and gradient estimators, we recover both known and new algorithms. Specifically, we obtain the following new features:

— **robustness to noise**: we introduce mechanisms to deal with stochastic perturbations and make all these previous approaches robust to noise.
— **adaptivity to the strong convexity when** $\tilde{\sigma} = 0$: Algorithms 2.1, 2.2, and 2.3 without averaging do not require knowing the strong convexity constant μ (it may only need a lower-bound β, which is often trivial to obtain).
— **new variants:** Whereas SVRG/SAGA were originally developed with the iterations (A) and MISO in the context of (B), we show that these gradient estimators are both compatible with (A) and (B), leading to new algorithms with similar guarantees.

2.3 Convergence Analysis and Robustness

We now present the convergence analysis for iterations (A) or (B). In Section 2.3.1, we present a generic convergence result. Then, in Section 2.3.2, we present specific results for the variance-reduction approaches in including strategies to make them robust to

stochastic noise. Acceleration is discussed in the next section.

2.3.1 Generic Convergence Result Without Variance Reduction

Key to our complexity results, the following proposition gives a first relation between the quantity $F(x_k)$, the surrogate d_k, d_{k-1} and the variance of the gradient estimates.

Proposition 2.1 (**Key relation**). *For either variant (A) or (B), when using the construction of d_k from Sections 2.2.1 or 2.2.2, respectively, and assuming $\eta_k \leq 1/L$, we have for all $k \geq 1$,*

$$\delta_k(\mathbb{E}\left[F(x_k)\right] - F^*) + \mathbb{E}\left[d_k(x^*) - d_k^*\right] \leq (1-\delta_k)\,\mathbb{E}\left[d_{k-1}(x^*) - d_{k-1}^*\right] + \eta_k\delta_k\omega_k^2, \quad (2.15)$$

where F^* is the minimum of F, x^* is one of its minimizers, and $\omega_k^2 = \mathbb{E}\left[\|g_k - \nabla f(x_{k-1})\|^2\right]$.

Proof. We first consider the variant (A) and later show how to modify the convergence proofs to accommodate the variant (B).

$$\begin{aligned}
d_k^* = d_k(x_k) &= (1-\delta_k)\,d_{k-1}(x_k) + \delta_k\left(f(x_{k-1}) + g_k^\top(x_k - x_{k-1}) + \frac{\mu}{2}\|x_k - x_{k-1}\|^2 + \psi(x_k)\right)\\
&\geq (1-\delta_k)\,d_{k-1}^* + \frac{\gamma_k}{2}\|x_k - x_{k-1}\|^2 + \delta_k\left(f(x_{k-1}) + g_k^\top(x_k - x_{k-1}) + \psi(x_k)\right)\\
&\geq (1-\delta_k)\,d_{k-1}^* + \delta_k\left(f(x_{k-1}) + g_k^\top(x_k - x_{k-1}) + \frac{L}{2}\|x_k - x_{k-1}\|^2 + \psi(x_k)\right)\\
&\geq (1-\delta_k)\,d_{k-1}^* + \delta_k F(x_k) + \delta_k(g_k - \nabla f(x_{k-1}))^\top(x_k - x_{k-1}),
\end{aligned}$$

where the first inequality comes from Lemma 2.6—it is in fact an equality when considering Algorithm (A)—and the second inequality simply uses the assumption $\eta_k \leq 1/L$, which yields $\delta_k = \gamma_k\eta_k \leq \gamma_k/L$. Finally, the last inequality uses a classical upper-bound for L-smooth functions presented in Lemma 2.4. Then, after taking expectations,

$$\begin{aligned}
\mathbb{E}[d_k^*] &\geq (1-\delta_k)\,\mathbb{E}\left[d_{k-1}^*\right] + \delta_k\mathbb{E}\left[F(x_k)\right] + \delta_k\mathbb{E}\left[(g_k - \nabla f(x_{k-1}))^\top(x_k - x_{k-1})\right]\\
&= (1-\delta_k)\,\mathbb{E}\left[d_{k-1}^*\right] + \delta_k\mathbb{E}\left[F(x_k)\right] + \delta_k\mathbb{E}\left[(g_k - \nabla f(x_{k-1}))^\top x_k\right]\\
&= (1-\delta_k)\,\mathbb{E}\left[d_{k-1}^*\right] + \delta_k\mathbb{E}\left[F(x_k)\right] + \delta_k\mathbb{E}\left[(g_k - \nabla f(x_{k-1}))^\top(x_k - w_{k-1})\right],
\end{aligned}$$

where we have defined the following quantity

$$w_{k-1} = \mathrm{Prox}_{\eta_k\psi}\left[x_{k-1} - \eta_k\nabla f(x_{k-1})\right].$$

In the previous relations, we have used twice the fact that $\mathbb{E}\left[(g_k - \nabla f(x_{k-1}))^\top y|\mathcal{F}_{k-1}\right] = 0$, for all deterministic variable y given x_{k-1}, such as $y = x_{k-1}$ or $y = w_{k-1}$. We may now use the non-expansiveness property of the proximal operator [Moreau, 1965] to control the quantity $\|x_k - w_{k-1}\|$, which gives us

$$\begin{aligned}
\mathbb{E}[d_k^*] &\geq (1-\delta_k)\,\mathbb{E}\left[d_{k-1}^*\right] + \delta_k\mathbb{E}\left[F(x_k)\right] - \delta_k\mathbb{E}\left[\|g_k - \nabla f(x_{k-1})\|\,\|x_k - w_{k-1}\|\right]\\
&\geq (1-\delta_k)\,\mathbb{E}\left[d_{k-1}^*\right] + \delta_k\mathbb{E}\left[F(x_k)\right] - \delta_k\eta_k\mathbb{E}\left[\|g_k - \nabla f(x_{k-1})\|^2\right]\\
&= (1-\delta_k)\,\mathbb{E}\left[d_{k-1}^*\right] + \delta_k\mathbb{E}\left[F(x_k)\right] - \delta_k\eta_k\omega_k^2.
\end{aligned}$$

This relation can now be combined with (2.10) when $z = x^*$, and we obtain (2.15). It is also easy to see that the proof also works with variant (B). The convergence analysis is identical, except that we take w_{k-1} to be

$$w_{k-1} = \mathrm{Prox}_{\psi/\gamma_k}\left[(1-\mu\eta_k)\,\bar{x}_{k-1} + \mu\eta_k x_{k-1} - \eta_k \nabla f(x_{k-1})\right],$$

and the same result follows. □

Then, without making further assumption on ω_k, we have the following general convergence result, which is a direct consequence of the averaging Lemma 2.12, inspired by Ghadimi and Lan [2012], and presented in Appendix 2.A.3:

Theorem 2.1 (**General convergence result**). *Under the same assumptions as in Proposition 2.1, we have for all $k \geq 1$, and either variant (A) or (B),*

$$\mathbb{E}\delta_k\left(F(x_k) - F^*\right) + d_k(x^*) - d_k^* \leq \Gamma_k\left(d_0(x^*) - d_0^* + \sum_{t=1}^{k}\frac{\delta_t\eta_t\omega_t^2}{\Gamma_t}\right), \tag{2.16}$$

where $\Gamma_k = \prod_{t=1}^{k}(1-\delta_t)$. Then, by using the averaging strategy $\hat{x}_k = (1-\delta_k)\,\hat{x}_{k-1} + \delta_k x_k$ of Lemma 2.12, for any point $\hat{x}_0$ (possibly equal to x_0), we have

$$\mathbb{E}\left[\delta_k\left(F(\hat{x}_k) - F^*\right) + d_k(x^*) - d_k^*\right] \leq \Gamma_k\left(F(\hat{x}_0) - F^* + d_0(x^*) - d_0^* + \sum_{t=1}^{k}\frac{\delta_t\eta_t\omega_t^2}{\Gamma_t}\right). \tag{2.17}$$

Theorem 2.1 allows us to recover convergence rates for various algorithms. Note that the effect of the averaging strategy is to remove the factor δ_k in front of $F(x_k) - F^*$ on the left part of (2.16), thus improving the convergence rate by a factor $1/\delta_k$. Regarding the quantity $d_0(x^*) - d_0^*$, we have the following relations

- For variant (A), $d_0(x^*) - d_0^* = (\gamma_0/2)\,\|x^* - x_0\|^2$;
- For variant (B), this quantity may be larger and we may simply say that $d_0(x^*) - d_0^* = \frac{\gamma_0}{2}\,\|x^* - x_0\|^2 + \psi(x^*) - \psi(x_0) - \psi'(x_0)^\top(x_0 - x^*)$ for variant (B), where $\psi'(x_0) = \gamma_0(x_0 - \bar{x}_0)$ is a subgradient in $\partial\psi(x_0)$. Note that if $\bar{x}_0$ is chosen to be a minimizer of ψ, then $d_0(x^*) - d_0^* = (\gamma_0/2)\,\|x^* - x_0\|^2 + \psi(x^*) - \psi(x_0)$.

In the next section, we will focus on variance reduction mechanisms, which are able to improve the previous convergence rates by better exploiting the structure of the objective. By controlling the variance ω_k of the corresponding gradient estimators, we will apply Theorem 2.1 to obtain convergence rates. Before that, we remark that it is relatively straightforward to use this theorem to recover complexity results for proximal SGD, both for the usual variant (A) or the new one (B). Since these results are classical, we present them in Appendix 2.C. As a sanity check, we note that we recover the optimal noise-dependency [see Nemirovski et al., 2009], both for strongly convex cases, or when $\mu = 0$.

2.3.2 Faster Convergence with Variance Reduction

Stochastic variance-reduced gradient descent algorithms rely on gradient estimates whose variance decreases as fast as the objective function value. Here, we provide a unified proof of convergence for our variants of SVRG, SAGA, and MISO, and we show how to make them robust to stochastic perturbations. Specifically, we consider the minimization of a finite sum of functions as in (1.27), but, as explained in Section 2.2, each observation of the gradient $\nabla f_i(x)$ is corrupted by a random noise variable. The next proposition extends a proof for SVRG [Xiao and Zhang, 2014] to stochastic perturbations, and characterizes the variance of g_k.

As we now consider finite sums, we introduce again the quantity $\tilde{\sigma}^2$, which is an upper-bound on the noise variance due to stochastic perturbations for all x in $\mathbb{R}^p$ and for i in $\{1, \dots, n\}$:

$$\mathbb{E}\left[\left\|\tilde{\nabla} f_i(x) - \nabla f_i(x)\right\|^2\right] \leq \tilde{\sigma}_i^2 \;\text{ such that }\; \tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{q_i n}\tilde{\sigma}_i^2, \tag{2.18}$$

where the expectation is with respect to the gradient perturbation, and $Q = \{q_1, \dots, q_n\}$ is the sampling distribution. As having the variance to be bounded across the domain of x may be a strong assumption, even though classical, we also introduce the quantity

$$\tilde{\sigma}_{i,*}^2 = \mathbb{E}\left[\left\|\tilde{\nabla} f_i(x^*) - \nabla f_i(x^*)\right\|^2\right] \;\text{ with a related quantity }\; \tilde{\sigma}_*^2 = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{q_i n}\tilde{\sigma}_{i,*}^2, \tag{2.19}$$

where x^* is a solution of the optimization problem. As we will show, in this section, our complexity results for unaccelerated methods when $\mu > 0$ under the bounded variance assumption $\tilde{\sigma}^2 < +\infty$ will also hold when simply assuming $\tilde{\sigma}_*^2 < +\infty$ at the cost of slightly degrading the complexity by constant factors. The next proposition provides an upper-bound on the variance of gradient estimators g_k, which we have introduced earlier, as a first step to use Theorem 2.1.

Proposition 2.2 (**Generic variance reduction with non-uniform sampling**). *Consider (2.1) when f is a finite sum of functions $f = \frac{1}{n}\sum_{i=1}^{n} f_i$ where each f_i is convex and L_i-smooth with $L_i \geq \mu$. Then, the gradient estimates g_k of the random-SVRG and MISO/SAGA/SDCA strategies defined in Section 2.2.3 satisfy*

$$\mathbb{E}\left[\|g_k - \nabla f(x_{k-1})\|^2\right] \leq 4L_Q \mathbb{E}\left[F(x_{k-1}) - F^*\right] + \frac{2}{n}\mathbb{E}\left[\sum_{i=1}^{n} \frac{1}{nq_i}\|u_{k-1}^i - u_*^i\|^2\right] + 3\rho_Q \tilde{\sigma}^2, \tag{2.20}$$

where $L_Q = \max_i L_i/(q_i n)$, $\rho_Q = 1/(n \min_i q_i)$, and for all i and k, u_k^i is equal to z_k^i without noise—that is

$$\begin{aligned} u_k^i &= \nabla f_i(\tilde{x}_k) \quad \textit{for random-SVRG} \\ u_k^{j_k} &= \nabla f_{j_k}(x_k) - \beta x_k \quad \textit{and} \quad u_k^j = u_{k-1}^j \quad \textit{if} \quad j \neq j_k \quad \textit{for SAGA/MISO/SDCA}, \end{aligned}$$

and $u_*^i = \nabla f_i(x^*) - \beta x^*$ (with $\beta = 0$ for random-SVRG).

If we additionally assume that each function f_i may be written as $f_i(x) = \mathbb{E}_\xi\left[\tilde{f}_i(x,\xi)\right]$ where $\tilde{f}_i(.,\xi)$ is L_i-smooth with $L_i \geq \mu$ for all ξ, then

$$\mathbb{E}\left[\|g_k - \nabla f(x_{k-1})\|^2\right] \leq 16 L_Q \mathbb{E}\left[F(x_{k-1}) - F^*\right] + \frac{2}{n}\mathbb{E}\left[\sum_{i=1}^n \frac{1}{nq_i}\left\|u_{k-1}^i - u_*^i\right\|^2\right] + 6\rho_Q \tilde{\sigma}_*^2. \tag{2.21}$$

In particular, choosing the uniform distribution $q_i = 1/n$ gives $L_Q = \max_i L_i$; choosing $q_i = L_i / \sum_j L_j$ gives $L_Q = (1/n)\sum_i L_i$, which may be significantly smaller than the maximum Lipschitz constant. We note that non-uniform sampling can significantly improve the dependency of the bound to the Lipschitz constants since the average $(1/n)\sum_i L_i$ may be significantly smaller than the maximum $\max_i L_i$, but it may worsen the dependency with the variance $\tilde{\sigma}^2$ since $\rho_Q > 1$ unless Q is the uniform distribution. The proof of the proposition is given in Appendix 2.D.1.

For simplicity, we will present our complexity results in terms of $\tilde{\sigma}^2$. However, when the conditions for (2.21) are satisfied, it is easy to adapt all results of this section to replace $\tilde{\sigma}^2$ by $\tilde{\sigma}_*^2$, by paying a small price in terms of constant factors. Note that this substitution will not work for accelerated algorithms in the next section. The general convergence result is given next; it applies to both variants (A) and (B).

Proposition 2.3 (**Lyapunov function for variance-reduced algorithms**). *Consider the same setting as Proposition 2.2. For either variant (A) or (B) with the random-SVRG or SAGA/SDCA/MISO gradient estimators defined in Section 2.2.3, when using the construction of d_k from Sections 2.2.1 or 2.2.2, respectively, and assuming $\gamma_0 \geq \mu$ and $(\eta_k)_{k\geq 0}$ is non-increasing with $\eta_k \leq \frac{1}{12 L_Q}$, we have for all $k \geq 1$,*

$$\frac{\delta_k}{6}\mathbb{E}\left[F(x_k) - F^*\right] + T_k \leq (1-\tau_k) T_{k-1} + 3\rho_Q \eta_k \delta_k \tilde{\sigma}^2 \quad \text{with} \quad \tau_k = \min\left(\delta_k, \frac{1}{5n}\right), \tag{2.22}$$

where

$$T_k = 5 L_Q \eta_k \delta_k \mathbb{E}\left[F(x_k) - F^*\right] + \mathbb{E}\left[d_k(x^*) - d_k^*\right] + \frac{5\eta_k\delta_k}{2}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \frac{1}{q_i n}\left\|u_k^i - u_*^i\right\|^2\right].$$

The proof of the previous proposition is given in Appendix 2.D.2. From the Lyapunov function, we obtain a general convergence result for the variance-reduced stochastic algorithms.

Theorem 2.2 (**Convergence of variance-reduced algorithms**). *Consider the same setting as Proposition 2.3, which applies to both variants (A) and (B). Then, by using the averaging strategy of Lemma 2.12 with any point $\hat{x}_0$,*

$$\mathbb{E}\left[F(\hat{x}_k) - F^* + \frac{6\tau_k}{\delta_k} T_k\right] \leq \Theta_k\left(F(\hat{x}_0) - F^* + \frac{6\tau_k}{\delta_k} T_0 + \frac{18\rho_Q \tau_k \tilde{\sigma}^2}{\delta_k}\sum_{t=1}^k \frac{\eta_t \delta_t}{\Theta_t}\right), \tag{2.23}$$

where $\Theta_k = \prod_{t=1}^k (1-\tau_t)$. Note that we also have

$$T_0 \leq 10 L_Q \eta_0 \delta_0 \left(F(x_0) - F^*\right) + d_0(x^*) - d_0^*. \tag{2.24}$$

The proof is given in Appendix 2.D.3. From this generic convergence theorem, we now study particular cases. The first corollary studies the strongly-convex case with constant step size.

Corollary 2.3 (**Variance-reduction, $\mu > 0$, constant step size independent of μ**). *Consider the same setting as in Theorem 2.2, where f is μ-strongly convex, $\gamma_0 = \mu$, and $\eta_k = \frac{1}{12L_Q}$. Then, for any point $\hat{x}_0$,*

$$\mathbb{E}\left[F(\hat{x}_k) - F^* + \alpha T_k\right] \leq \Theta_k \left(F(\hat{x}_0) - F^* + \alpha T_0\right) + \frac{3\rho_Q \tilde{\sigma}^2}{2L_Q} \tag{2.25}$$

with $\tau = \min\left(\frac{\mu}{12L_Q}, \frac{1}{5n}\right)$, $\Theta_k = (1-\tau)^k$, and $\alpha = 6\min\left(1, \frac{12L_Q}{5\mu n}\right)$. Note that $T_k \geq \frac{\mu}{2}\left\|x_k - x^*\right\|^2$ and for Algorithm (A), we also have $T_0 \leq (13/12)\left(F(x_0) - F^*\right)$.

The proof is given in Appendix 2.D.4. This corollary shows that the algorithm achieves a linear convergence rate to a noise-dominated region and produces converging iterates $(x_k)_{k\geq 0}$ that do not require to know the strong convexity constant μ. It shows that all estimators we consider can become *adaptive* to μ. Note that the non-uniform strategy slightly degrades the dependency in $\tilde{\sigma}^2$: indeed, $L_Q/\rho_Q = \max_{i=1} L_i$ if Q is uniform, but if $q_i = \max_i L_i / \sum_j L_j$, we have instead $L_Q/\rho_Q = \min_{i=1} L_i$. The next corollary shows that a slightly better noise dependency can be achieved when the step sizes rely on μ.

Corollary 2.4 (**Variance-reduction, $\mu > 0$, μ-dependent constant step size**). *Consider the same setting as Theorem 2.2, where f is μ-strongly convex, $\gamma_0 = \mu$, and $\eta_k = \eta = \min\left(\frac{1}{12L_Q}, \frac{1}{5\mu n}\right)$. Then, for all $\hat{x}_0$,*

$$\mathbb{E}\left[F(\hat{x}_k) - F^* + 6T_k\right] \leq \Theta_k \left(F(\hat{x}_0) - F^* + 6T_0\right) + 18\rho_Q \eta \tilde{\sigma}^2. \tag{2.26}$$

The proof follows similar steps as the proof of Corollary 2.3, after noting that we have $\delta_k = \tau_k$ for all k for this particular choice of step size. We are now in shape to study a converging algorithm.

Corollary 2.5 (**Variance-reduction, $\mu > 0$, decreasing step sizes**). *Consider the same setting as Theorem 2.2, where f is μ-strongly convex and target an accuracy $\varepsilon \leq 24\rho_Q \eta \tilde{\sigma}^2$, with $\eta = \min\left(\frac{1}{12L_Q}, \frac{1}{5\mu n}\right)$. Then, we use the constant step size strategy of Corollary 2.4 with $\hat{x}_0 = x_0$, and stop the optimization when we find points $\hat{x}_k$ and x_k such that $\mathbb{E}\left[F(\hat{x}_k) - F^* + 6T_k\right] \leq 24\rho_Q \eta \tilde{\sigma}^2$. Then, we restart the optimization procedure with decreasing step sizes $\eta_k = \min\left(\frac{1}{12L_Q}, \frac{1}{5\mu n}, \frac{2}{\mu(k+2)}\right)$ and generate a new sequence $(\hat{x}'_k)_{k\geq 0}$. The resulting number of gradient evaluations to achieve $\mathbb{E}\left[F(\hat{x}'_k) - F^*\right] \leq \varepsilon$ is upper bounded by*

$$\mathcal{O}\left(\left(n + \frac{L_Q}{\mu}\right)\log\left(\frac{F(x_0) - F^* + d_0(x^*) - d_0^*}{\varepsilon}\right)\right) + \mathcal{O}\left(\frac{\rho_Q \tilde{\sigma}^2}{\mu\varepsilon}.\right)$$

Note that $d_0(x^*) - d_0^* \leq F(x_0) - F^*$ for variant (A).

The proof is given in Appendix 2.D.5 and shows that variance-reduction algorithms may exhibit an optimal dependency on the noise level $\tilde{\sigma}^2$ when the objective is strongly convex. Next, we analyze the complexity of variant (A) when $\mu = 0$. Note that it is possible to conduct a similar analysis for variant (B), which exhibits a slightly worse complexity (as the corresponding quantity $d_0(x^*) - d_0^*$ is larger).

Corollary 2.6 (**Convergence of variance-reduced algorithms with constant step size,** $\mu = 0$)**.** *Consider the same setting as Theorem 2.2, where f is convex and proceed in two steps. First, run one iteration of (A) with step size $\frac{1}{12L_Q}$ with the gradient estimator $(1/n)\sum_{i=1}^n \tilde{\nabla} f_i(x_0)$. Second, use the resulting point to initialize the variant (A) with the random-SVRG or SAGA/SDCA/MISO gradient estimators, with a constant step size $\eta \leq \frac{1}{12L_Q}$, $\gamma_0 = 1/\eta$, for a total of $K \geq 5n\log(5n)$ iterations. Then,*

$$\mathbb{E}\left[F(\hat{x}_K) - F^*\right] \leq \frac{9n}{\eta(K+1)}\left\|x_0 - x^*\right\|^2 + 36\eta\tilde{\sigma}^2\rho_Q.$$

If in addition we choose $\eta = \min\left(\frac{1}{12L_Q}, \frac{\|x_0 - x^*\|}{2\tilde{\sigma}}\sqrt{\frac{n}{\rho_Q(K+1)}}\right)$.

$$\mathbb{E}\left[F(\hat{x}_K) - F^*\right] \leq \frac{108nL_Q}{(K+1)}\left\|x_0 - x^*\right\|^2 + 36\tilde{\sigma}\left\|x_0 - x^*\right\|\sqrt{\frac{\rho_Q n}{K+1}}. \tag{2.27}$$

The proof is provided in Appendix 2.D.6. The second part of the corollary is not a practical result since the optimal step size depends on unknown quantities such as $\tilde{\sigma}^2$, but it allows us to highlight the best possible dependence between the budget of iterations K, the initial point x_0, and the noise $\tilde{\sigma}^2$. We will show in the next section that acceleration is useful to improve the previous complexity.

2.4 Accelerated Stochastic Algorithms

We now consider the following iteration, involving an extrapolation sequence $(y_k)_{k\geq 1}$, which is a classical mechanism from accelerated first-order algorithms [Beck and Teboulle, 2009, Nesterov, 2013]. Given a sequence of step sizes $(\eta_k)_{k\geq 0}$ with $\eta_k \leq 1/L$ for all $k \geq 0$, and some parameter $\gamma_0 \geq \mu$, we consider the sequences $(\delta_k)_{k\geq 0}$ and $(\gamma_k)_{k\geq 0}$ that satisfy

$$\begin{aligned}\delta_k &= \sqrt{\eta_k\gamma_k} \quad \text{for all } k \geq 0\\ \gamma_k &= (1-\delta_k)\,\gamma_{k-1} + \delta_k\mu \quad \text{for all } k \geq 1.\end{aligned}$$

Then, for $k \geq 1$, we consider the iteration

$$\boxed{\begin{aligned} x_k &= \mathrm{Prox}_{\eta_k\psi}\left[y_{k-1} - \eta_k g_k\right] \quad \text{with} \quad \mathbb{E}\left[g_k|\mathcal{F}_{k-1}\right] = \nabla f(y_{k-1})\\ y_k &= x_k + \beta_k(x_k - x_{k-1}) \quad \text{with} \quad \beta_k = \frac{\delta_k\left(1-\delta_k\right)\eta_{k+1}}{\eta_k\delta_{k+1} + \eta_{k+1}\delta_k^2}, \end{aligned}} \tag{C}$$

where with constant step size $\eta_k = 1/L$, we recover a classical extrapolation parameter of accelerated gradient based methods [Nesterov, 2014]. Traditionally, estimate sequences are used to analyze the convergence of accelerated algorithms. We show in this section how to proceed for stochastic composite optimization and later, we show how to directly accelerate the random-SVRG approach we have introduced. Note that Algorithm (C) resembles the approaches introduced by Hu et al. [2009], Ghadimi and Lan [2012] but is simpler since our approach involves a single extrapolation step.

2.4.1 Convergence Analysis Without Variance Reduction

Consider then the stochastic estimate sequence for $k \geq 1$

$$d_k(x) = (1 - \delta_k)\, d_{k-1}(x) + \delta_k l_k(x),$$

with d_0 defined as in (2.7) and

$$l_k(x) = f(y_{k-1}) + g_k^\top (x - y_{k-1}) + \frac{\mu}{2} \|x - y_{k-1}\|^2 + \psi(x_k) + \psi'(x_k)^\top (x - x_k), \qquad (2.28)$$

and $\psi'(x_k) = \frac{1}{\eta_k}(y_{k-1} - x_k) - g_k$ is in $\partial\psi(x_k)$ by definition of the proximal operator. As in Section 2.2, $d_k(x^*)$ asymptotically becomes a lower bound on F^* since (2.10) remains satisfied. This time, the iterate x_k does not minimize d_k, and we denote by v_k instead its minimizer, allowing us to write d_k in the canonical form

$$d_k(x) = d_k^* + \frac{\gamma_k}{2} \|x - v_k\|^2 .$$

The first lemma highlights classical relations between the iterates $(x_k)_{k\geq 0}$, $(y_k)_{k\geq 0}$ and the minimizers of the estimate sequences d_k, which also appears in [Nesterov, 2014, p. 78] for constant step sizes η_k. The proof is given in Appendix 2.D.7.

Lemma 2.1 (Relations between y_k, x_k and d_k). *The sequences $(x_k)_{k\geq 0}$ and $(y_k)_{k\geq 0}$ produced by Algorithm (C) satisfy for all $k \geq 0$, with $v_0 = y_0 = x_0$,*

$$y_k = (1 - \theta_k)x_k + \theta_k v_k \quad \textit{with} \quad \theta_k = \frac{\delta_k \gamma_k}{\gamma_k + \delta_{k+1}\mu}.$$

Then, the next lemma is key to prove the convergence of Algorithm (C). Its proof is given in Appendix 2.D.8.

Lemma 2.2 (Key lemma for stochastic estimate sequences with acceleration). *Assuming $(x_k)_{k\geq 0}$ and $(y_k)_{k\geq 0}$ are given by Algorithm (C). Then, for all $k \geq 1$,*

$$\mathbb{E}\left[F(x_k)\right] \leq \mathbb{E}\left[l_k(y_{k-1})\right] + \left(\frac{L\eta_k^2}{2} - \eta_k\right) \mathbb{E}\left[\|\tilde{g}_k\|^2\right] + \eta_k \omega_k^2,$$

with $\omega_k^2 = \mathbb{E}\left[\|\nabla f(y_{k-1}) - g_k\|^2\right]$ and $\tilde{g}_k = g_k + \psi'(x_k)$.

Finally, we obtain the following convergence result.

Theorem 2.7 (**Convergence of the accelerated stochastic optimization algorithm**). *Under the assumptions of Lemma 2.1, we have for all* $k \geq 1$,

$$\mathbb{E}\left[F(x_k) - F^* + \frac{\gamma_k}{2}\left\|v_k - x^*\right\|^2\right] \leq \Gamma_k \left(F(x_0) - F^* + \frac{\gamma_0}{2}\left\|x_0 - x^*\right\|^2 + \sum_{t=1}^{k} \frac{\eta_t \omega_t^2}{\Gamma_t}\right),$$

where, as before, $\Gamma_t = \sum_{i=1}^{t}(1-\delta_i)$.

Proof. First, the minimizer v_k of the quadratic surrogate d_k may be written as

$$v_k = \frac{(1-\delta_k)\gamma_{k-1}}{\gamma_k} v_{k-1} + \frac{\mu\delta_k}{\gamma_k} y_{k-1} - \frac{\delta_k}{\gamma_k}\tilde{g}_k = y_{k-1} + \frac{(1-\delta_k)\gamma_{k-1}}{\gamma_k}(v_{k-1} - y_{k-1}) - \frac{\delta_k}{\gamma_k}\tilde{g}_k.$$

Then, we characterize the quantity d_k^*:

$$\begin{aligned}
d_k^* &= d_k(y_{k-1}) - \frac{\gamma_k}{2}\left\|v_k - y_{k-1}\right\|^2 \\
&= (1-\delta_k)\, d_{k-1}(y_{k-1}) + \delta_k l_k(y_{k-1}) - \frac{\gamma_k}{2}\left\|v_k - y_{k-1}\right\|^2 \\
&= (1-\delta_k)\left(d_{k-1}^* + \frac{\gamma_{k-1}}{2}\left\|v_{k-1} - y_{k-1}\right\|^2\right) + \delta_k l_k(y_{k-1}) - \frac{\gamma_k}{2}\left\|v_k - y_{k-1}\right\|^2 \\
&= (1-\delta_k)\, d_{k-1}^* + \left(\frac{\gamma_{k-1}(1-\delta_k)(\gamma_k - (1-\delta_k)\gamma_{k-1})}{2\gamma_k}\right)\left\|v_{k-1} - y_{k-1}\right\|^2 + \delta_k l_k(y_{k-1}) \\
&\qquad - \frac{\delta_k^2}{2\gamma_k}\left\|\tilde{g}_k\right\|^2 + \frac{\delta_k(1-\delta_k)\gamma_{k-1}}{\gamma_k}\tilde{g}_k^\top(v_{k-1} - y_{k-1}) \\
&\geq (1-\delta_k)\, d_{k-1}^* + \delta_k l_k(y_{k-1}) - \frac{\delta_k^2}{2\gamma_k}\left\|\tilde{g}_k\right\|^2 + \frac{\delta_k(1-\delta_k)\gamma_{k-1}}{\gamma_k}\tilde{g}_k^\top(v_{k-1} - y_{k-1}).
\end{aligned}$$

Assuming by induction that $\mathbb{E}\left[d_{k-1}^*\right] \geq \mathbb{E}\left[F(x_{k-1})\right] - \xi_{k-1}$ for some $\xi_{k-1} \geq 0$, we have after taking expectation

$$\begin{aligned}
\mathbb{E}\left[d_k^*\right] \geq (1-\delta_k)\left(\mathbb{E}\left[F(x_{k-1})\right] - \xi_{k-1}\right) + \delta_k \mathbb{E}\left[l_k(y_{k-1})\right] - \frac{\delta_k^2}{2\gamma_k}\mathbb{E}\left\|\tilde{g}_k\right\|^2 \\
+ \frac{\delta_k(1-\delta_k)\gamma_{k-1}}{\gamma_k}\mathbb{E}\left[\tilde{g}_k^\top(v_{k-1} - y_{k-1})\right].
\end{aligned}$$

Then, note that $\mathbb{E}\left[F(x_{k-1})\right] \geq \mathbb{E}\left[l_k(x_{k-1})\right] \geq \mathbb{E}\left[l_k(y_{k-1})\right] + \mathbb{E}\left[\tilde{g}_k^\top(x_{k-1} - y_{k-1})\right]$, and

$$\begin{aligned}
\mathbb{E}\left[d_k^*\right] \geq \mathbb{E}\left[l_k(y_{k-1})\right] - (1-\delta_k)\,\xi_{k-1} - \frac{\delta_k^2}{2\gamma_k}\mathbb{E}\left\|\tilde{g}_k\right\|^2 + \\
(1-\delta_k)\,\mathbb{E}\left[\tilde{g}_k^\top\left(\frac{\delta_k\gamma_{k-1}}{\gamma_k}(v_{k-1} - y_{k-1}) + (x_{k-1} - y_{k-1})\right)\right].
\end{aligned}$$

By Lemma 2.1, we can show that the last term is equal to zero, and we are left with

$$\mathbb{E}\left[d_k^*\right] \geq \mathbb{E}\left[l_k(y_{k-1})\right] - (1-\delta_k)\,\xi_{k-1} - \frac{\delta_k^2}{2\gamma_k}\mathbb{E}\left\|\tilde{g}_k\right\|^2.$$

We may then use Lemma 2.2, which gives us

$$\begin{aligned}\mathbb{E}\left[d_k^*\right] &\geq \mathbb{E}\left[F(x_k)\right] - (1-\delta_k)\,\xi_{k-1} - \eta_k\omega_k^2 + \left(\eta_k - \frac{L\eta_k^2}{2} - \frac{\delta_k^2}{2\gamma_k}\right)\mathbb{E}\left\|\tilde{g}_k\right\|^2 \\ &\geq \mathbb{E}\left[F(x_k)\right] - \xi_k \quad \text{with} \quad \xi_k = (1-\delta_k)\,\xi_{k-1} + \eta_k\omega_k^2,\end{aligned}$$

where we used the fact that $\eta_k \leq 1/L$ and $\delta_k = \sqrt{\gamma_k\eta_k}$.

It remains to choose $d_0^* = F(x_0)$ and $\xi_0 = 0$ to initialize the induction at $k = 0$ and we conclude that

$$\mathbb{E}\left[F(x_k) - F^* + \frac{\gamma_k}{2}\left\|v_k - x^*\right\|^2\right] \leq \mathbb{E}\left[d_k(x^*) - F^*\right] + \xi_k \leq \Gamma_k(d_0(x^*) - F^*) + \xi_k,$$

which gives us the statement of the theorem when noticing that $\xi_k = \Gamma_k \sum_{t=1}^k \frac{\eta_t\omega_t^2}{\Gamma_t}$. □

Next, we specialize the theorem to various practical cases. For the corollaries below, we assume the variances $(\omega_k^2)_{k\geq 1}$ to be upper bounded by σ^2.

Corollary 2.8 (**Proximal accelerated SGD with constant step size, $\mu > 0$**). *Assume that f is μ-strongly convex, and choose $\gamma_0 = \mu$ and $\eta_k = 1/L$ with Algorithm (C). Then,*

$$\mathbb{E}\left[F(x_k) - F^*\right] \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)^k \left(F(x_0) - F^* + \frac{\mu}{2}\left\|x_0 - x^*\right\|^2\right) + \frac{\sigma^2}{\sqrt{\mu L}}. \tag{2.29}$$

We now show that with decreasing step sizes, we obtain an algorithm with optimal complexity similar to [Ghadimi and Lan, 2013].

Corollary 2.9 (**Proximal accelerated SGD with decreasing step sizes and $\mu > 0$**). *Assume that f is μ-strongly convex and that we target an accuracy ε smaller than $2\sigma^2/\sqrt{\mu L}$. First, use a constant step size $\eta_k = 1/L$ with $\gamma_0 = \mu$ within Algorithm (C), leading to the convergence rate (2.29), until $\mathbb{E}\left[F(x_k) - F^*\right] \leq 2\sigma^2/\sqrt{\mu L}$. Then, we restart the optimization procedure with decreasing step sizes $\eta_k = \min\left(\frac{1}{L}, \frac{4}{\mu(k+2)^2}\right)$ and generate a new sequence $(\hat{x}_k)_{k\geq 0}$. The resulting number of gradient evaluations to achieve $\mathbb{E}\left[F(x_k) - F^*\right] \leq \varepsilon$ is upper bounded by*

$$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{F(x_0) - F^*}{\varepsilon}\right)\right) + \mathcal{O}\left(\frac{\sigma^2}{\mu\varepsilon}\right).$$

The proof is provided in Appendix 2.D.9. We note that despite the "optimal" theoretical complexity, we have observed that Algorithm (C) with the parameters of Corollaries 2.8 and 2.9 could be relatively unstable, as shown in Section 2.5, due to the large radius $\sigma^2/\sqrt{\mu L}$ of the noise region. When μ is small, such a quantity may be indeed arbitrarily larger than $F(x_0) - F^*$. Instead, we have found a mini-batch strategy to be more effective in practice. When using a mini-batch of size $b = \lceil L/\mu \rceil$, the theoretical complexity becomes the same as SGD, given in Corollary 2.16, but the algorithm enjoys the benefits of easy parallelization.

Corollary 2.10 (**Proximal accelerated SGD with with $\mu = 0$**). *Assume that f is convex. Consider a step size $\eta \leq 1/L$ and run one iteration of Algorithm (A) with a stochastic gradient estimate. Use the resulting point to initialize Algorithm (C) still with constant step size η, and choose $\gamma_0 = 1/\eta$. Then,*

$$\mathbb{E}\left[F(x_k) - F^*\right] \leq \frac{2\left\|x_0 - x^*\right\|^2}{(1+K)^2\eta} + \sigma^2\eta(K+1).$$

If in addition we choose $\eta = \min\left(\frac{1}{L}, \sqrt{\frac{2\|x_0-x^*\|^2}{\sigma^2}}\frac{1}{(K+1)^{3/2}}\right)$, then

$$\mathbb{E}\left[F(x_k) - F^*\right] \leq \frac{2L\left\|x_0 - x^*\right\|^2}{(1+K)^2} + 2\|x_0 - x^*\|\sigma\sqrt{\frac{2}{1+K}}. \tag{2.30}$$

The proof is given in Appendix 2.D.10. These convergence results are relatively similar to those obtained in [Ghadimi and Lan, 2013] for a different algorithm and is optimal for convex functions.

2.4.2 An Accelerated Algorithm with Variance Reduction

In this section, we show how to combine the previous methodology with variance reduction, and introduce Algorithm 2.4 based on random-SVRG. Then, we present the convergence analysis, which requires controlling the variance of the estimator in a similar manner to [Allen-Zhu, 2017], as stated in the next proposition. Note that the estimator does not require storing the seed of the random perturbations, unlike in the previous section.

Proposition 2.4 (**Variance reduction for random-SVRG estimator**). *Consider problem (2.1) when f is a finite sum of functions $f = \frac{1}{n}\sum_{i=1}^n f_i$ where each f_i is L_i-smooth with $L_i \geq \mu$ and f is μ-strongly convex. Then, the variance of g_k defined in Algorithm 2.4 satisfies*

$$\omega_k^2 \leq 2L_Q\left[f(\tilde{x}_{k-1}) - f(y_{k-1}) - g_k^\top(\tilde{x}_{k-1} - y_{k-1})\right] + 3\rho_Q\tilde{\sigma}^2.$$

The proof is given in Appendix 2.D.11. Then, we extend Lemma 2.2 that was used in the previous analysis to the variance-reduction setting.

Lemma 2.3 (**Lemma for accelerated variance-reduced stochastic optimization**). *Consider the iterates provided by Algorithm 2.4 and call $a_k = 2L_Q\eta_k$. Then,*

$$\mathbb{E}\left[F(x_k)\right] \leq \mathbb{E}\left[a_kF(\tilde{x}_{k-1}) + (1-a_k)l_k(y_{k-1})\right] + \mathbb{E}\left[a_k\tilde{g}_k^\top(y_{k-1} - \tilde{x}_{k-1}) + \left(\frac{L\eta_k^2}{2} - \eta_k\right)\|\tilde{g}_k\|^2\right] + 3\rho_Q\eta_k\tilde{\sigma}^2.$$

The proof of this lemma is given in Appendix 2.D.12. With this lemma in hand, we may now state our main convergence result.

Algorithm 2.4 Accelerated algorithm with random-SVRG estimator

Input: x_0 in $\mathbb{R}^p$ (initial point); K (number of iterations); $(\eta_k)_{k\geq 0}$ (step sizes); $\gamma_0 \geq \mu$;
Initialization: $\tilde{x}_0 = v_0 = x_0$; $\bar{z}_0 = \tilde{\nabla} f(x_0)$;
for $k = 1, \ldots, K$ **do**
    Find (δ_k, γ_k) such that

$$\gamma_k = (1 - \delta_k)\,\gamma_{k-1} + \delta_k \mu \quad \text{and} \quad \delta_k = \sqrt{\frac{5\eta_k \gamma_k}{3n}};$$

    Choose

$$y_{k-1} = \theta_k v_{k-1} + (1 - \theta_k)\tilde{x}_{k-1} \quad \text{with} \quad \theta_k = \frac{3n\delta_k - 5\mu\eta_k}{3 - 5\mu\eta_k};$$

    Sample i_k according to the distribution $Q = \{q_1, \ldots, q_n\}$;
    Compute the gradient estimator, possibly corrupted by stochastic perturbations:

$$g_k = \frac{1}{q_{i_k} n}\left(\tilde{\nabla} f_{i_k}(y_{k-1}) - \tilde{\nabla} f_{i_k}(\tilde{x}_{k-1})\right) + \bar{z}_{k-1};$$

    Obtain the new iterate

$$x_k \leftarrow \mathrm{Prox}_{\eta_k \psi}\left[y_{k-1} - \eta_k g_k\right];$$

    Find the minimizer v_k of the estimate sequence d_k:

$$v_k = \left(1 - \frac{\mu\delta_k}{\gamma_k}\right) v_{k-1} + \frac{\mu\delta_k}{\gamma_k} y_{k-1} + \frac{\delta_k}{\gamma_k \eta_k}(x_k - y_{k-1});$$

    With probability $1/n$, update the anchor point

$$\tilde{x}_k = x_k \quad \text{and} \quad \bar{z}_k = \tilde{\nabla} f(\tilde{x}_k);$$

    Otherwise, with probability $1 - 1/n$, keep the anchor point unchanged $\tilde{x}_k = \tilde{x}_{k-1}$ and $\bar{z}_k = \bar{z}_{k-1}$;
end for
Output: x_K.

Theorem 2.11 (Convergence of the accelerated SVRG algorithm). *Consider the iterates provided by Algorithm 2.4 and assume that the step sizes satisfy* $\eta_k \leq \min\left(\frac{1}{3L_Q}, \frac{1}{15\gamma_k n}\right)$ *for all* $k \geq 1$. *Then,*

$$\mathbb{E}\left[F(x_k) - F^* + \frac{\gamma_k}{2}\left\|v_k - x^*\right\|^2\right] \leq \Gamma_k\left(F(x_0) - F^* + \frac{\gamma_0}{2}\left\|x_0 - x^*\right\|^2 + \frac{3\rho_Q\tilde{\sigma}^2}{n}\sum_{t=1}^{k}\frac{\eta_t}{\Gamma_t}\right).$$

Proof. Following similar steps as in the proof of Theorem 2.7, we have

$$d_k^* \geq (1-\delta_k)\, d_{k-1}^* + \delta_k l_k(y_{k-1}) - \frac{\delta_k^2}{2\gamma_k}\left\|\tilde{g}_k\right\|^2 + \frac{\delta_k\left(1-\delta_k\right)\gamma_{k-1}}{\gamma_k}\tilde{g}_k^\top(v_{k-1} - y_{k-1}).$$

Assume now by induction that $\mathbb{E}\left[d_{k-1}^*\right] \geq \mathbb{E}\left[F(\tilde{x}_{k-1})\right] - \xi_{k-1}$ for some $\xi_{k-1} \geq 0$ and note that $\delta_k \leq \frac{1-a_k}{n}$ since $a_k = 2L_Q\eta_k \leq \frac{2}{3}$ and $\delta_k = \sqrt{\frac{5\eta_k\gamma_k}{3n}} \leq \frac{1}{3n} \leq \frac{1-a_k}{n}$. Then,

$$\begin{aligned}
\mathbb{E}\left[d_k^*\right] &\geq (1-\delta_k)\left(\mathbb{E}\left[F(\tilde{x}_{k-1})\right] - \xi_{k-1}\right) + \delta_k\mathbb{E}\left[l_k(y_{k-1})\right] - \frac{\delta_k^2}{2\gamma_k}\mathbb{E}\left[\left\|\tilde{g}_k\right\|^2\right] \\
&\qquad + \mathbb{E}\left[\tilde{g}_k^\top\left(\frac{\delta_k\left(1-\delta_k\right)\gamma_{k-1}}{\gamma_k}(v_{k-1} - y_{k-1})\right)\right] \\
&\geq \left(1 - \frac{1-a_k}{n}\right)\mathbb{E}\left[F(\tilde{x}_{k-1})\right] + \left(\frac{1-a_k}{n} - \delta_k\right)\mathbb{E}\left[F(\tilde{x}_{k-1})\right] + \delta_k\mathbb{E}\left[l_k(y_{k-1})\right] - \\
&\qquad \frac{\delta_k^2}{2\gamma_k}\mathbb{E}\left[\left\|\tilde{g}_k\right\|^2\right] + \mathbb{E}\left[\tilde{g}_k^\top\left(\frac{\delta_k\left(1-\delta_k\right)\gamma_{k-1}}{\gamma_k}(v_{k-1} - y_{k-1})\right)\right] - (1-\delta_k)\,\xi_{k-1}.
\end{aligned}$$

Note that

$$\mathbb{E}\left[F(\tilde{x}_{k-1})\right] \geq \mathbb{E}\left[l_k(\tilde{x}_{k-1})\right] \geq \mathbb{E}\left[l_k(y_{k-1})\right] + \mathbb{E}\left[\tilde{g}_k^\top(\tilde{x}_{k-1} - y_{k-1})\right].$$

Then,

$$\begin{aligned}
&\mathbb{E}\left[d_k^*\right] \geq \left(1 - \frac{1-a_k}{n}\right)\mathbb{E}\left[F(\tilde{x}_{k-1})\right] + \frac{1-a_k}{n}\mathbb{E}\left[l_k(y_{k-1})\right] - \frac{\delta_k^2}{2\gamma_k}\mathbb{E}\left[\left\|\tilde{g}_k\right\|^2\right] \\
&+\mathbb{E}\left[\tilde{g}_k^\top\left(\frac{\delta_k\left(1-\delta_k\right)\gamma_{k-1}}{\gamma_k}(v_{k-1} - y_{k-1}) + \left(\frac{1-a_k}{n} - \delta_k\right)(\tilde{x}_{k-1} - y_{k-1})\right)\right] - (1-\delta_k)\,\xi_{k-1}.
\end{aligned}$$

We may now use Lemma 2.3, which gives us

$$\begin{aligned}
\mathbb{E}\left[d_k^*\right] &\geq \left(1 - \frac{1}{n}\right)\mathbb{E}\left[F(\tilde{x}_{k-1})\right] + \frac{1}{n}\mathbb{E}\left[F(x_k)\right] + \left(\frac{1}{n}\left(\eta_k - \frac{L\eta_k^2}{2}\right) - \frac{\delta_k^2}{2\gamma_k}\right)\mathbb{E}\left[\left\|\tilde{g}_k\right\|^2\right] \\
&\quad + \mathbb{E}\left[\tilde{g}_k^\top\left(\frac{\delta_k\left(1-\delta_k\right)\gamma_{k-1}}{\gamma_k}(v_{k-1} - y_{k-1}) + \left(\frac{1}{n} - \delta_k\right)(\tilde{x}_{k-1} - y_{k-1})\right)\right] - \xi_k, \quad (2.31)
\end{aligned}$$

with $\xi_k = (1-\delta_k)\,\xi_{k-1} + \frac{3\rho_Q\eta_k\tilde{\sigma}^2}{n}$. Then, since $\delta_k = \sqrt{\frac{5\eta_k\gamma_k}{3n}}$ and $\eta_k \leq \frac{1}{3L_Q} \leq \frac{1}{3L}$,

$$\frac{1}{n}\left(\eta_k - \frac{L\eta_k^2}{2}\right) - \frac{\delta_k^2}{2\gamma_k} \geq \frac{5\eta_k}{6n} - \frac{\delta_k^2}{2\gamma_k} = 0,$$

and the term in (2.31) involving $\|\tilde{g}_k\|^2$ may disappear. Similarly, we have

$$\frac{\delta_k(1-\delta_k)\gamma_{k-1}}{\delta_k(1-\delta_k)\gamma_{k-1}+\gamma_k/n-\delta_k\gamma_k} = \frac{\delta_k\gamma_k-\delta_k^2\mu}{\gamma_k/n-\delta_k^2\mu}$$
$$= \frac{3n\delta_k^3/5\eta_k-\delta_k^2\mu}{3\delta_k^2/5\eta_k-\delta_k^2\mu} = \frac{3n-5\mu\eta_k}{3-5\mu\eta_k} = \theta_k,$$

and the term in (2.31) that is linear in $\tilde{g}_k$ may disappear as well. Then, we are left with $\mathbb{E}[d_k^*] \geq \mathbb{E}[F(\tilde{x}_k)] - \xi_k$. Initializing the induction requires choosing $\xi_0 = 0$ and $d_0^* = F(x_0)$. Ultimately, we note that $\mathbb{E}[d_k(x^*) - F^*] \leq (1-\delta_k)\mathbb{E}[d_{k-1}(x^*) - F^*]$ for all $k \geq 1$, and

$$\mathbb{E}\left[F(\tilde{x}_k) - F^* + \frac{\gamma_k}{2}\|x^* - v_k\|^2\right] \leq \mathbb{E}[d_k(x^*) - F^*] + \xi_k$$
$$\leq \Gamma_k\left(F(x_0) - F^* + \frac{\gamma_0}{2}\|x^* - x_0\|^2\right) + \xi_k,$$

and we obtain the statement. □

We may now derive convergence rates of our accelerated SVRG algorithm under various settings. The proofs of the following corollaries, when not straightforward, are given in the appendix. The first corollary simply uses Lemma 2.9.

Corollary 2.12 (**Accelerated proximal SVRG - constant step size -** $\mu > 0$)**.** *With* $\eta_k = \min\left(\frac{1}{3L_Q}, \frac{1}{15\mu n}\right)$ *and* $\gamma_0 = \mu$*, the iterates produced by Algorithm 2.4 satisfy*

— *if* $\frac{1}{3L_Q} \leq \frac{1}{15\mu n}$*,*

$$\mathbb{E}[F(x_k) - F^*] \leq \left(1 - \sqrt{\frac{5\mu}{9L_Q n}}\right)^k \left(F(x_0) - F^* + \frac{\mu}{2}\|x_0 - x^*\|^2\right) + \frac{3\rho_Q\tilde{\sigma}^2}{\sqrt{5\mu L_Q n}};$$

— *otherwise,*

$$\mathbb{E}[F(x_k) - F^*] \leq \left(1 - \frac{1}{3n}\right)^k \left(F(x_0) - F^* + \frac{\mu}{2}\|x_0 - x^*\|^2\right) + \frac{3\rho_Q\tilde{\sigma}^2}{5\mu n}.$$

The corollary uses the fact that $\Gamma_k \sum_{t=1}^k \eta/\Gamma_t \leq \eta/\delta = \sqrt{3n\eta/5\mu}$ and thus the algorithm converges linearly to an area of radius $3\rho_Q\tilde{\sigma}^2\sqrt{3\eta/5\mu n} = \mathcal{O}\left(\rho_Q\tilde{\sigma}^2\min\left(\frac{1}{\sqrt{n\mu L_Q}}, \frac{1}{\mu n}\right)\right)$, where as before, $\rho_Q = 1$ if the distribution Q is uniform. When $\tilde{\sigma}^2 = 0$, the corresponding algorithm achieves the optimal complexity for finite sums [Arjevani and Shamir, 2016]. Interestingly, we see that here non-uniform sampling may hurt the convergence guarantees in some situations. Whenever $\frac{1}{\max_i L_i} > \frac{1}{5\mu n}$, the optimal sampling strategy is indeed the uniform one. Next, we show how to obtain a converging algorithm in the next corollary.

Corollary 2.13 (**Accelerated proximal SVRG - diminishing step sizes -** $\mu > 0$)**.** *Assume that* f *is* μ*-strongly convex and that we target an accuracy* ε *smaller than* $B = 3\rho_Q\tilde{\sigma}^2\sqrt{\eta/\mu}$ *with the same step size* η *as in the previous corollary. First, use such a constant step size strategy* $\eta_k = \eta$ *with* $\gamma_0 = \mu$ *within Algorithm 2.4, leading to the*

convergence rate of the previous corollary, until $\mathbb{E}\left[F(x_k) - F^*\right] \leq B$. *Then, we restart the optimization procedure with decreasing step sizes* $\eta_k = \min\left(\eta, \frac{12n}{5\mu(k+2)^2}\right)$ *and generate a new sequence* $(\hat{x}_k)_{k\geq 0}$. *The resulting number of gradient evaluations to achieve* $\mathbb{E}\left[F(x_k) - F^*\right] \leq \varepsilon$ *is upper bounded by*

$$\mathcal{O}\left(\left(n + \sqrt{\frac{nL_Q}{\mu}}\right)\log\left(\frac{F(x_0) - F^*}{\varepsilon}\right)\right) + \mathcal{O}\left(\frac{\rho_Q \sigma^2}{\mu\varepsilon}\right).$$

The proof is given in Appendix 2.D.13. Next, we study the case when $\mu = 0$.

Corollary 2.14 (**Accelerated proximal SVRG -** $\mu = 0$)**.** *Consider the same setting as Theorem 2.11, where* f *is convex and proceed in two steps. First, run one iteration of (A) with step size* $\eta \leq \frac{1}{3L_Q}$ *with the gradient estimator* $(1/n)\sum_{i=1}^n \tilde{\nabla} f_i(x_0)$. *Second, use the resulting point to initialize Algorithm 2.4 and use step size* $\eta_t = \min\left(\frac{1}{3L_Q}, \frac{1}{15\gamma_t n}\right)$, *with* $\gamma_0 = 1/\eta$, *for a total of* $K \geq 6n\log(15n) + 1$ *iterations. Then*

$$\mathbb{E}\left[F(x_K) - F^*\right] \leq \frac{6n\left\|x_0 - x^*\right\|^2}{\eta(K+1)^2} + \frac{3\eta\rho_Q\tilde{\sigma}^2(K+1)}{n}.$$

If in addition we choose $\eta = \min\left(\frac{1}{3L_Q}, \frac{\sqrt{2}n\|x_0 - x^*\|}{\tilde{\sigma}\sqrt{\rho_Q}(K+1)^{3/2}}\right)$,

$$\mathbb{E}\left[F(x_K) - F^*\right] \leq \frac{18L_Q n\left\|x_0 - x^*\right\|^2}{(K+1)^2} + \frac{6\tilde{\sigma}\left\|x_0 - x^*\right\|\sqrt{2\rho_Q}}{\sqrt{K+1}}. \tag{2.32}$$

The proof is provided in Appendix 2.D.14. When $\tilde{\sigma}^2 = 0$ (deterministic setting), the first part of the corollary with $\eta = 1/3L_Q$ gives us the same complexity as Katyusha [Allen-Zhu, 2017] and RPDG [Lan and Zhou, 2018a], and in the stochastic case, we obtain a significantly better complexity than the same algorithm without acceleration, which was analyzed in Corollary 2.6.

2.5 Experiments

In this section, we evaluate numerically the approaches introduced in the previous sections.

2.5.1 Datasets, Formulations, and Methods

Following classical benchmarks in optimization methods for machine learning [see, *e.g.* Schmidt et al., 2017], we consider empirical risk minimization formulations. Given training data $(a_i, b_i)_{i=1,\ldots,n}$, with a_i in $\mathbb{R}^p$ and b_i in $\{-1, +1\}$, we consider the optimization problem

$$\min_{x\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \phi(b_i a_i^\top x) + \frac{\lambda}{2}\left\|x\right\|^2,$$

where ϕ is either the logistic loss $\phi(u) = \log(1 + e^{-u})$, or the squared hinge loss $\phi(u) = \max(0, 1-u)^2$. Both functions are L-smooth; when the vectors a_i have unit norm, we may

indeed choose $L = 0.25$ for the logistic loss and $L = 1$ for the squared hinge loss. Studying the squared hinge loss is interesting: whereas the logistic loss has bounded gradients on $\mathbb{R}^p$, this is not the case for the squared hinge loss. With unbounded optimization domain, the gradient norms may be indeed large in some regions of the solution space, which may lead in turn to large variance σ^2 of the gradient estimates obtained by SGD, causing instabilities.

The scalar λ is a regularization parameter that acts as a lower bound on the strong convexity constant of the problem. We consider the parameters $\mu = \lambda = 1/10n$ in our problems, which is of the order of the smallest values that one would try when doing a parameter search, *e.g.*, by cross-validation. For instance, this is empirically observed for the dataset cifar-ckn described below, where a test set is available, allowing us to check that the "optimal" regularization parameter leading to the lowest generalization error is indeed of this order. We also report an experiment with $\lambda = 1/100n$ in order to study the effect of the problem conditioning on the method's performance.

Following Bietti and Mairal [2017], Zheng and Kwok [2018], we consider DropOut perturbations [Srivastava et al., 2014] to illustrate the robustness to noise of the algorithms. DropOut consists of randomly setting to zero each entry of a data point with probability δ, leading to the optimization problem

$$\min_{x \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\rho \left[\phi(b_i(\rho \circ a_i)^\top x) \right] + \frac{\lambda}{2} \|x\|^2 , \tag{2.33}$$

where ρ is a binary vector in $\{0, 1\}^p$ with i.i.d. Bernoulli entries, and $\circ$ denotes the element-wise multiplication between two vectors. We consider two DropOut regimes, with δ in $\{0.01, 0.1\}$, representing small and medium perturbations, respectively.

Then, we consider three datasets with various number of points n and dimension p, coming from different scientific fields:

— alpha is from the Pascal Large Scale Learning Challenge website [2] and contains $n = 250\,000$ points in dimension $p = 500$.

— gene consists of gene expression data and the binary labels b_i characterize two different types of breast cancer. This is a small dataset with $n = 295$ and $p = 8\,141$.

— ckn-cifar is an image classification task where each image from the CIFAR-10 dataset [3] is represented by using a two-layer unsupervised convolutional neural network [Mairal, 2016]. Since CIFAR-10 originally contains 10 different classes, we consider the binary classification task consisting of predicting the class 1 vs. other classes. The dataset contains $n = 50\,000$ images and the dimension of the representation is $p = 9\,216$.

For simplicity, we normalize the features of all datasets and thus we use a uniform sampling strategy Q in all algorithms. Then, we consider several methods with their theoretical step sizes, described in Table 2.1. Note that we also evaluate the strategy random-SVRG with step size $1/3L$, even though our analysis requires $1/12L$, in order to get a fair comparison with the accelerated SVRG method. In all figures, we consider that n iterations of SVRG count as 2 effective passes over the data since it appears to be a good proxy of the

2. `http://largescale.ml.tu-berlin.de/`

3. `https://www.cs.toronto.edu/~kriz/cifar.html`

Algorithm	step size η_k	Cor.	Complexity $\mathcal{O}(.)$	Bias $\mathcal{O}(.)$
SGD	$\frac{1}{L}$	2.15	$\frac{L}{\mu}\log\left(\frac{C_0}{\varepsilon}\right)$	$\frac{\sigma^2}{L}$
SGD-d	$\min\left(\frac{1}{L}, \frac{2}{\mu(k+2)}\right)$	2.16	$\frac{L}{\mu}\log\left(\frac{C_0}{\varepsilon}\right) + \frac{\sigma^2}{\mu\varepsilon}$	0
acc-SGD	$\frac{1}{L}$	2.8	$\sqrt{\frac{L}{\mu}}\log\left(\frac{C_0}{\varepsilon}\right)$	$\frac{\sigma^2}{\sqrt{\mu L}}$
acc-SGD-d	$\min\left(\frac{1}{L}, \frac{4}{\mu(k+2)^2}\right)$	2.9	$\sqrt{\frac{L}{\mu}}\log\left(\frac{C_0}{\varepsilon}\right) + \frac{\sigma^2}{\mu\varepsilon}$	0
acc-mb-SGD-d	$\min\left(\frac{1}{L}, \frac{4}{\mu(k+2)^2}\right)$	2.9	$\frac{L}{\mu}\log\left(\frac{C_0}{\varepsilon}\right) + \frac{\sigma^2}{\mu\varepsilon}$	0
rand-SVRG	$\frac{1}{3L}$	2.3	$\left(n + \frac{L}{\mu}\right)\log\left(\frac{C_0}{\varepsilon}\right)$	$\frac{\tilde{\sigma}^2}{L}$
rand-SVRG-d	$\min\left(\frac{1}{12L_Q}, \frac{1}{5\mu n}, \frac{2}{\mu(k+2)}\right)$	2.5	$\left(n + \frac{L}{\mu}\right)\log\left(\frac{C_0}{\varepsilon}\right) + \frac{\tilde{\sigma}^2}{\mu\varepsilon}$	0
acc-SVRG	$\min\left(\frac{1}{3L_Q}, \frac{1}{15\mu n}\right)$	2.12	$\left(n + \sqrt{\frac{nL}{\mu}}\right)\log\left(\frac{C_0}{\varepsilon}\right)$	$\frac{\tilde{\sigma}^2}{\sqrt{n\mu L}+n\mu}$
acc-SVRG-d	$\min\left(\frac{1}{3L_Q}, \frac{1}{15\mu n}, \frac{12n}{5\mu(k+2)^2}\right)$	2.13	$\left(n + \sqrt{\frac{nL}{\mu}}\right)\log\left(\frac{C_0}{\varepsilon}\right) + \frac{\tilde{\sigma}^2}{\mu\varepsilon}$	0

Table 2.1 – List of algorithms used in the experiments, along with the step size used and the pointer to the corresponding convergence guarantees, with $C_0 = F(x_0) - F^*$. In the experiments, we also use the method rand-SVRG with step size $\eta = 1/3L$, even though our analysis requires $\eta \leq 1/12L$. The approach acc-mb-SGD-d uses mini-batches of size $\lceil\sqrt{L/\mu}\rceil$ and could thus easily be parallelized. Note that we potentially have $\tilde{\sigma} \ll \sigma$.

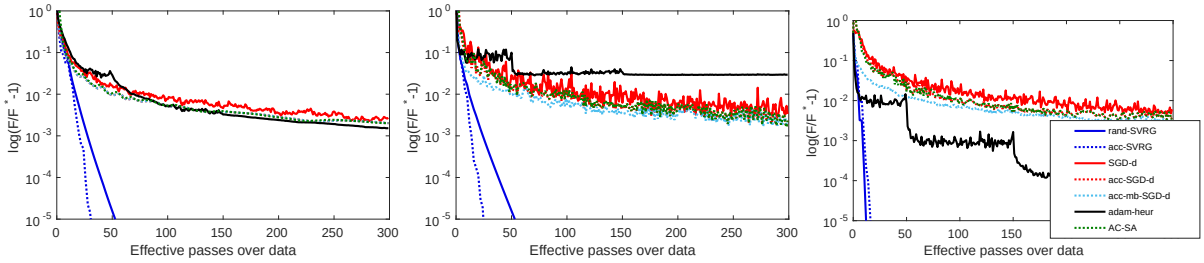
computational time. Indeed, (i) if one is allowed to store the variables z_i^k, then n iterations exactly correspond to two passes over the data; (ii) the gradients $\tilde{\nabla} f_i(x_{k-1}) - \tilde{\nabla} f_i(\tilde{x}_{k-1})$ access the same training point which reduces the data access overhead; (iii) computing the full gradient $\bar{z}_k$ can be done in practice in a much more efficient manner than computing individually the n gradients $\tilde{\nabla} f_i(x_k)$, either through parallelization or by using more efficient routines (*e.g.*, BLAS2 vs BLAS1 routines for linear algebra). Each experiment is conducted five times and we always report the average of the five experiments in each figure. We also include in the comparison two baselines from the literature: AC-SA is the accelerated stochastic gradient descent method of Ghadimi and Lan [2013], and adam-heur is the Adam method of Kingma and Ba [2014] with its recommended step size. As Adam is not converging, we adopt a standard heuristics from the deep learning literature, consisting of reducing the step size by 10 after 50 and 150 passes over the data, respectively, which performs much better than using a constant step size in practice.

2.5.2 Evaluation of Algorithms without Perturbations

First, we study the behavior of all methods when $\tilde{\sigma}^2 = 0$. We report the corresponding results in Figures 2.1, 2.2, and 2.3. Since the problem is deterministic, we can check that the value F^* we consider is indeed optimal by computing a duality gap using Fenchel duality. For SGD and random-SVRG, we do not use any averaging strategy, which we found to empirically slow down convergence, when used from the start; knowing when to start averaging is indeed not easy and requires heuristics which we do not evaluate here.

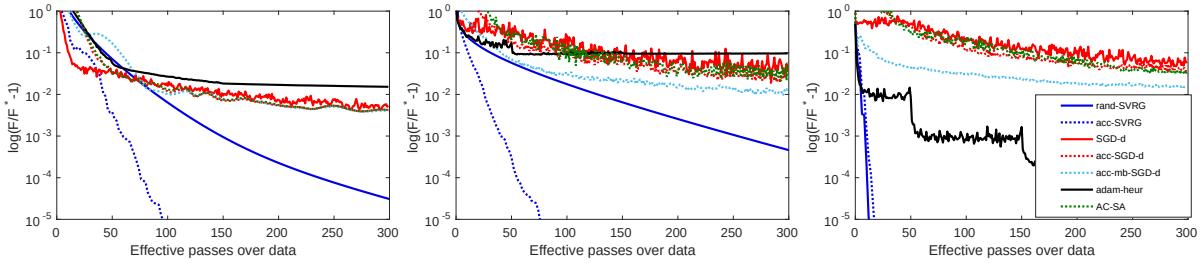
From these experiments, we obtain the following conclusions:

— Acceleration for SVRG is effective on the datasets gene and ckn-cifar except on alpha, where all SVRG-like methods perform already well. This may be due to strong



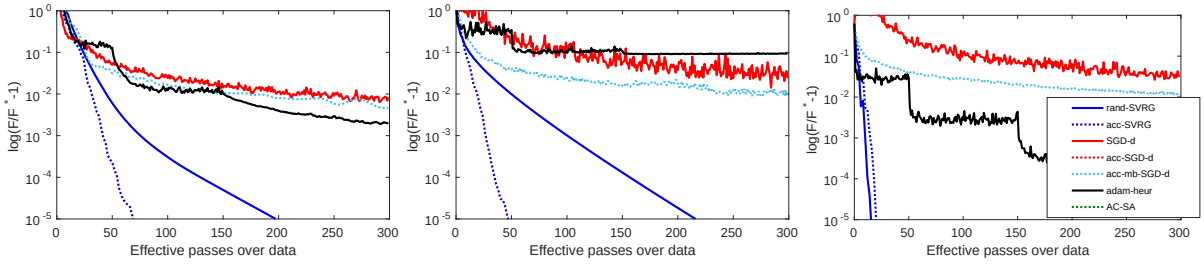
(a) Dataset gene (b) Dataset ckn-cifar (c) Dataset alpha

Figure 2.1 – Optimization curves without perturbations when using the logistic loss and the parameter $\lambda = 1/10n$. We plot the value of the objective function on a logarithmic scale as a function of the effective passes over the data (see main text for details). Best seen in color by zooming on a computer screen. Note that the method Adam is not converging.



(a) Dataset gene (b) Dataset ckn-cifar (c) Dataset alpha

Figure 2.2 – Same experiment as in Figure 2.1 with $\lambda = 1/100n$.



(a) Dataset gene (b) Dataset ckn-cifar (c) Dataset alpha

Figure 2.3 – Same experiment as in Figure 2.1 with squared hinge loss instead of logistic. ACC-SA and acc-SGD-d were unstable for this setting due to the large size of the noise region $\sigma^2/\sqrt{\mu L} = \sqrt{10n}\sigma^2$ and potentially large gradients of the loss function over the optimization domain.

convexity hidden in alpha leading to a regime where acceleration does not occur—that is, when the complexity is $\mathcal{O}\left(n \log(1/\varepsilon)\right)$, which is independent of the condition number.

— Acceleration is more effective when the problem is badly conditioned. When $\lambda = 1/100n$, acceleration brings several orders of magnitude improvement in

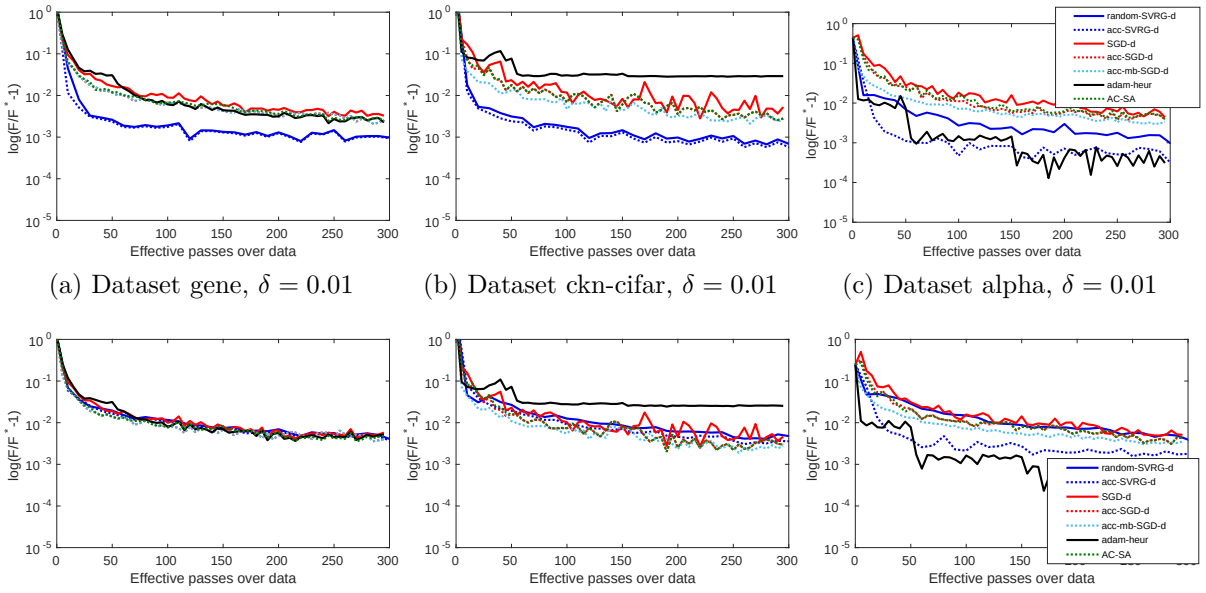
complexity.

— Accelerated SGD is unstable with the squared hinge loss. During the initial phase with constant step size $1/L$, the expected primal gap is in a region of radius $\mathcal{O}\left(\sigma^2/\sqrt{\mu L}\right) \approx \sqrt{n}\sigma^2$, which is potentially huge, causing large gradients and instabilities.

— Accelerated mini-batch SGD performs best among the SGD methods and is competitive with SVRG in the low precision regime. The performance of Adam on these datasets is inconsistent; it performs best among SGD methods on alpha, but is significantly worse on ckn-cifar. Note also that AC-SA performs in general similarly to acc-SGD-d.

2.5.3 Evaluation of Algorithms with Perturbations

We now consider the same setting as in the previous section, but we add DropOut perturbations with rate δ in $\{0.01, 0.1\}$. As predicted by theory, all approaches with constant step size do not converge. Therefore, we only report the results for decreasing step sizes in Figures 2.4, 2.5, and 2.6. We evaluate the loss function every 5 data passes and we estimate the expectation (2.33) by drawing 5 random perturbations per data point, resulting in $5n$ samples. The optimal value F^* is estimated by letting the methods run for 1000 epochs and selecting the best point found as a proxy of F^*.



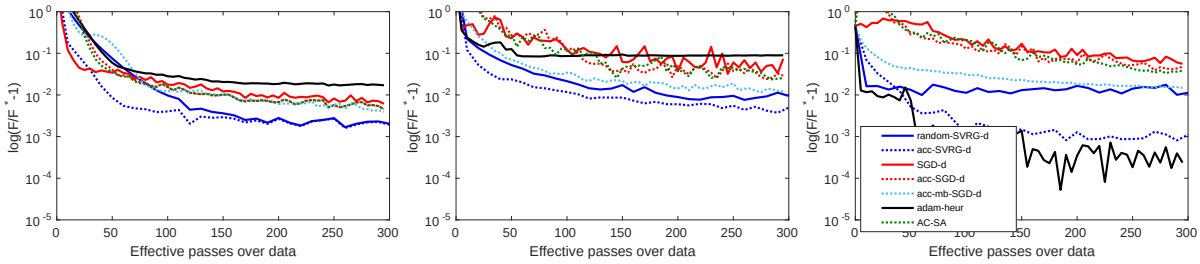
(a) Dataset gene, $\delta = 0.01$ (b) Dataset ckn-cifar, $\delta = 0.01$ (c) Dataset alpha, $\delta = 0.01$

(d) Dataset gene, $\delta = 0.1$ (e) Dataset ckn-cifar, $\delta = 0.1$ (f) Dataset alpha, $\delta = 0.1$

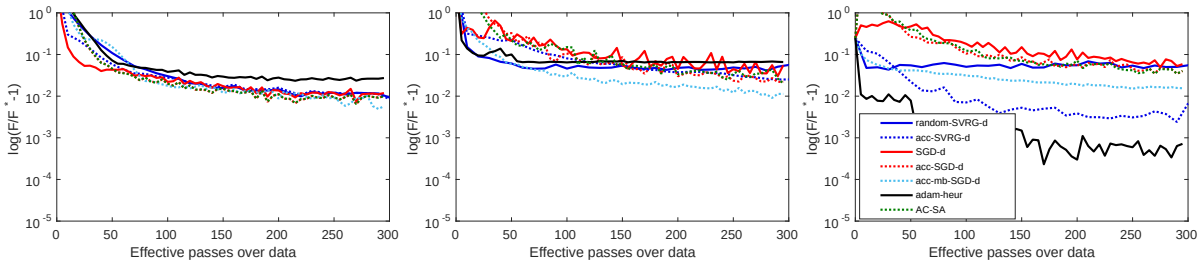
Figure 2.4 – Optimization curves with DropOut rate δ when using the logistic loss and $\lambda = 1/10n$. We plot the value of the objective function on a logarithmic scale as a function of the effective passes over the data. Best seen in color by zooming on a computer screen.

The conclusions of these experiments are the following:

— accelerated mini-batch SGD performs the best among SGD approaches in general except on alpha where Adam performs best.

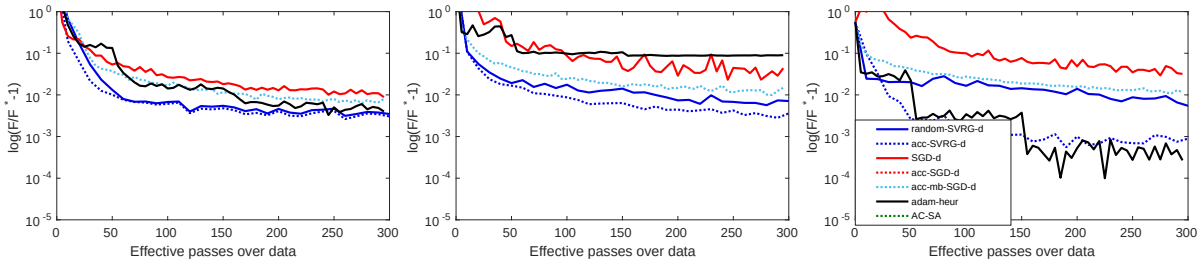


(a) Dataset gene, $\delta = 0.01$ (b) Dataset ckn-cifar, $\delta = 0.01$ (c) Dataset alpha, $\delta = 0.01$

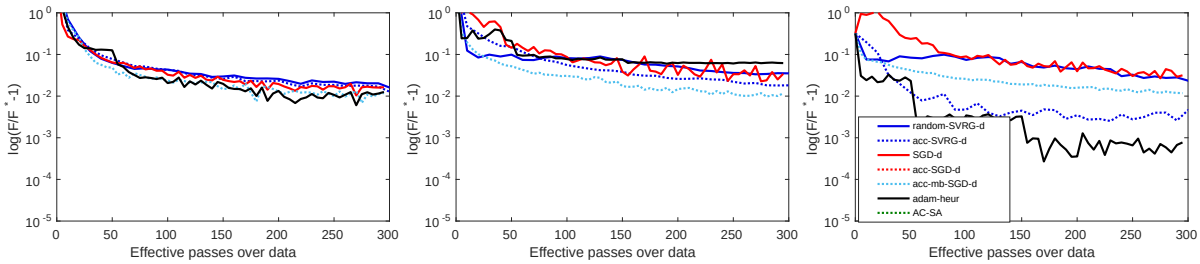


(d) Dataset gene, $\delta = 0.1$ (e) Dataset ckn-cifar, $\delta = 0.1$ (f) Dataset alpha, $\delta = 0.1$

Figure 2.5 – Same setting as in Figure 2.4 but with $\lambda = 1/100n$.



(a) Dataset gene, $\delta = 0.01$ (b) Dataset ckn-cifar, $\delta = 0.01$ (c) Dataset alpha, $\delta = 0.01$



(d) Dataset gene, $\delta = 0.1$ (e) Dataset ckn-cifar, $\delta = 0.1$ (f) Dataset alpha, $\delta = 0.1$

Figure 2.6 – Same setting as in Figure 2.4 but with the squared hinge loss.

— accelerated SVRG performs better than SVRG in general, or they achieve the same performance. As in the deterministic case, the gains are typically more important in ill-conditioned cases.
— accelerated SVRG performs better than SGD approaches in the low perturbation regime $\delta = 0.01$ and only on the alpha dataset when $\delta = 0.1$. Otherwise, the methods perform similarly.
— not reported on these figures, high perturbation regimes, *e.g.*, $\delta = 0.3$ make variance reduction less useful since the noise due to data sampling becomes potentially of the same order as $\tilde{\sigma}^2$; Yet, benefits are still seen on the alpha dataset, whereas SGD approaches perform slightly better than SVRG approaches on ckn-cifar and gene.

2.6 Discussion

In this chapter, we have studied simple stochastic gradient-based rules with or without variance reduction, and presented an accelerated algorithm dedicated to finite-sums minimization under the presence of stochastic perturbations. The approach we propose achieves the classical optimal worst-case complexities for finite-sum optimization when there is no perturbation [Arjevani and Shamir, 2016], and exhibits an optimal dependency in the noise variance $\tilde{\sigma}^2$ for convex and strongly convex problems.

Our work is based on stochastic variants of estimate sequences introduced by Nesterov [1983, 2014]. The framework leads naturally to many algorithms with relatively generic proofs of convergence, where convergence is proven at the same time as the algorithm's design. With iterate averaging techniques inspired by Ghadimi and Lan [2013], we show that a large class of variance-reduction stochastic optimization methods can be made robust to stochastic perturbations. Estimate sequences also naturally lead to several accelerated algorithms, some of them we did not present in this chapter. For instance, it is possible to show that replacing in (2.28) the lower bound $\psi(x_k) + \psi'(x_k)^\top(x - x_k)$ by $\psi(x)$ itself—in a similar way as we proceeded to obtain iteration (B) from iteration (A)—also leads to an accelerated algorithm with similar guarantees as (C).

Possibilities offered by estimate sequences are large, but our framework also admits a few limitations, paving the way for future work. In particular, our results are currently limited to Euclidean metrics—meaning that our convergence rates typically depend on quantities involving the Euclidean norm (e.g., strong convexity or L-smooth inequalities), and one may expect extensions of our work to other metrics such as Bregman distances. Estimate sequences admit indeed known extensions to such metrics, and can also deal with higher-order smoothness assumptions than Lipschitz continuity of the gradient [Baes, 2009]—*e.g.*, cubic regularization [Nesterov and Polyak, 2006]. We leave such directions for the future.

Another limitation we encountered was the inability to propose robust accelerated variants of SAGA, MISO, or SDCA based on our stochastic estimate sequences framework. To address this problem, we investigate in Chapter 3 a significantly different approach based on the Catalyst method [Lin et al., 2018], allowing us to accelerate stochastic first-order methods in a generic fashion, at the price of a logarithmic factor in the optimal complexity—in other words, we were able to obtain for SAGA, MISO, and SDCA a

complexity close to (2.4) up to a logarithmic factor in the condition number L_Q/μ. We believe that estimate sequences may be useful to obtain the optimal complexity without this logarithmic term, but the construction would be non-trivial and would rely on a different lower bound than the one we used in Section 2.4.

Finally, we note that the optimal complexities we have obtained with diminishing step sizes for strongly convex objectives can also be achieved by using instead a constant step size combined with mini-batch and restart strategies. As a constant step size yields a linear rate of convergence to a noise-dominated region of radius $\mathcal{O}(\tilde{\sigma}^2)$, we can indeed use the restart procedure described in Section 3.2 of the next chapter, which would yield an optimal or near-optimal complexity.

Appendix

2.A Useful Mathematical Results

2.A.1 Simple Results about Convexity and Smoothness

The next three lemmas are classical upper and lower bounds for smooth or strongly convex functions [Nesterov, 2014].

Lemma 2.4 (**Quadratic upper bound for L-smooth functions**). *Let $f : \mathbb{R}^p \to \mathbb{R}$ be L-smooth. Then, for all x, x' in $\mathbb{R}^p$,*

$$\left|f(x') - f(x) - \nabla f(x)^\top (x' - x)\right| \leq \frac{L}{2}\|x - x'\|_2^2.$$

Lemma 2.5 (**Lower bound for strongly convex functions**). *Let $f : \mathbb{R}^p \to \mathbb{R}$ be a μ-strongly convex function. Let z be in $\partial f(x)$ for some x in $\mathbb{R}^p$. Then, the following inequality holds for all x' in $\mathbb{R}^p$:*

$$f(x') \geq f(x) + z^\top (x' - x) + \frac{\mu}{2}\|x - x'\|_2^2.$$

Lemma 2.6 (**Second-order growth property**). *Let $f : \mathbb{R}^p \to \mathbb{R}$ be a μ-strongly convex function and $\mathcal{X} \subseteq \mathbb{R}^p$ be a convex set. Let x^* be the minimizer of f on $\mathcal{X}$. Then, the following condition holds for all x in $\mathcal{X}$:*

$$f(x) \geq f(x^*) + \frac{\mu}{2}\|x - x^*\|_2^2.$$

Lemma 2.7 (**Useful inequality for smooth and convex functions**). *Consider an L-smooth μ-strongly convex function f defined on $\mathbb{R}^p$ and a parameter β in $[0, \mu]$. Then, for all x, y in $\mathbb{R}^p$,*

$$\|\nabla f(x) - \nabla f(y) - \beta(x - y)\|^2 \leq 2L(f(x) - f(y) - \nabla f(y)^\top (x - y)).$$

Proof. Let us define the function $\phi(x) = f(x) - \frac{\beta}{2}\|x\|^2$, which is $(\mu - \beta)$-strongly convex. It is then easy to show that ϕ is $(L - \beta)$-smooth, according to Theorem 2.1.5 in [Nesterov,

2014]: indeed, for all x, y in $\mathbb{R}^p$,

$$\begin{aligned}\phi(x) = f(x) - \frac{\beta}{2}\|x\|^2 &\leq f(y) + \nabla f(y)^\top (x-y) + \frac{L}{2}\|x-y\|^2 - \frac{\beta}{2}\|x\|^2 \\ &= \phi(y) + \nabla\phi(y)^\top (x-y) + \frac{L-\beta}{2}\|x-y\|^2,\end{aligned}$$

and again according to Theorem 2.1.5 of [Nesterov, 2014],

$$\begin{aligned}\|\nabla\phi(x) - \nabla\phi(y)\|^2 &\leq 2L(\phi(x) - \phi(y) - \nabla\phi(y)^\top(x-y)) \\ &= 2L\left(f(x) - f(y) - \nabla f(y)^\top(x-y) - \frac{\beta}{2}\|x-y\|^2\right) \\ &\leq 2L\left(f(x) - f(y) - \nabla f(y)^\top(x-y)\right).\end{aligned}$$

□

2.A.2 Useful Results to Select Step Sizes

In this section, we present basic mathematical results regarding the choice of step sizes. The proofs of the first two lemmas are trivial by induction.

Lemma 2.8 (**Relation between** $(\delta_k)_{k\geq 0}$ **and** $(\Gamma_k = \prod_{t=1}^k (1-\delta_t))_{k\geq 0}$)**.** *Consider the following cases:*

- $\delta_k = \delta$ *(constant). Then* $\Gamma_k = (1-\delta)^k$*;*
- $\delta_k = 1/(k+1)$*. Then,* $\Gamma_k = \delta_k = \frac{1}{(k+1)}$*;*
- $\delta_k = 2/(k+2)$*. Then,* $\Gamma_k = \frac{2}{(k+1)(k+2)}$*;*
- $\delta_k = \min(1/(k+1), \delta)$*. then,*

$$\Gamma_k = \begin{cases} (1-\delta)^k & \text{if } k < k_0 \quad \text{with} \quad k_0 = \left\lceil \frac{1}{\delta} - 1 \right\rceil \\ \Gamma_{k_0 - 1}\frac{k_0}{k+1} & \text{otherwise.} \end{cases}$$

- $\delta_k = \min(2/(k+2), \delta)$*. then,*

$$\Gamma_k = \begin{cases} (1-\delta)^k & \text{if } k < k_0 \quad \text{with} \quad k_0 = \left\lceil \frac{2}{\delta} - 2 \right\rceil \\ \Gamma_{k_0 - 1}\frac{k_0(k_0+1)}{(k+1)(k+2)} & \text{otherwise.} \end{cases}$$

Lemma 2.9 (**Simple relation**)**.** *Consider a sequence of weights* $(\delta_k)_{k\geq 0}$ *in* $(0,1)$*. Then,*

$$\sum_{t=1}^k \frac{\delta_t}{\Gamma_t} + 1 = \frac{1}{\Gamma_k} \qquad \text{where} \qquad \Gamma_t \triangleq \prod_{i=1}^t (1-\delta_i). \tag{2.34}$$

Lemma 2.10 (**Convergence rate of** Γ_k)**.** *Consider the same quantities defined in the previous lemma and consider the sequence* $\gamma_k = (1-\delta_k)\gamma_{k-1} + \delta_k\mu = \Gamma_k\gamma_0 + (1-\Gamma_k)\mu$ *with* $\gamma_0 \geq \mu$*, and assume the relation* $\delta_k = \gamma_k\eta$*. Then, for all* $k \geq 0$*,*

$$\Gamma_k \leq \min\left((1-\mu\eta)^k, \frac{1}{1+\gamma_0\eta k}\right). \tag{2.35}$$

Besides,

— *when* $\gamma_0 = \mu$, *then* $\Gamma_k = (1 - \mu\eta)^k$.
— *when* $\mu = 0$, $\Gamma_k = \frac{1}{1+\gamma_0\eta k}$.

Proof. First, we have for all k, $\gamma_k \geq \mu$ such that $\delta_k \geq \eta\mu$, which leads then to $\Gamma_k \leq (1 - \eta\mu)^k$. Besides, $\gamma_k \geq \Gamma_k\gamma_0$ and thus $\Gamma_k = (1 - \delta_k)\,\Gamma_{k-1} \leq (1 - \Gamma_k\gamma_0\eta)\,\Gamma_{k-1}$. Then, $\frac{1}{\Gamma_k}(1 - \Gamma_k\gamma_0\eta) \geq \frac{1}{\Gamma_{k-1}}$, and

$$\frac{1}{\Gamma_k} \geq \frac{1}{\Gamma_{k-1}} + \gamma_0\eta \geq 1 + \gamma_0\eta k,$$

which is sufficient to obtain (2.35). Then, the fact that $\gamma_0 = \mu$ leads to $\Gamma_k = (1 - \mu\eta)^k$ is trivial, and the fact that $\mu = 0$ yields $\Gamma_k = \frac{1}{1+\gamma_0\eta k}$ can be shown by induction. Indeed, the relation is true for Γ_0 and then, assuming the relation is true for $k - 1$, we have for $k \geq 1$,

$$\Gamma_k = (1 - \delta_k)\,\Gamma_{k-1} = (1 - \eta\gamma_k)\Gamma_{k-1} = (1 - \eta\gamma_0\Gamma_k)\Gamma_{k-1} \geq (1 - \eta\gamma_0\Gamma_k)\,\frac{1}{1 + \gamma_0\eta(k–1)},$$

which leads to $\Gamma_k = \frac{1}{1+\gamma_0\eta k}$. □

Lemma 2.11 (**Accelerated convergence rate of** Γ_k)**.** *Consider the same quantities defined in Lemma 2.9 and consider the sequence* $\gamma_k = (1 - \delta_k)\,\gamma_{k-1} + \delta_k\mu = \Gamma_k\gamma_0 + (1 - \Gamma_k)\,\mu$ *with* $\gamma_0 \geq \mu$, *and assume the relation* $\delta_k = \sqrt{\gamma_k\eta}$. *Then, for all* $k \geq 0$,

$$\Gamma_k \leq \min\left((1 - \sqrt{\mu\eta})^k, \frac{4}{(2 + \sqrt{\gamma_0\eta}k)^2}\right).$$

Besides, when $\gamma_0 = \mu$, *then* $\Gamma_k = (1 - \sqrt{\mu\eta})^k$.

Proof. see Lemma 2.2.4 of [Nesterov, 2014]. □

2.A.3 Averaging Strategies

Next, we show a generic convergence result and an appropriate averaging strategy given a recursive relation between quantities acting as Lyapunov function.

Lemma 2.12 (**Averaging strategy**)**.** *Assume that an algorithm generates a sequence* $(x_k)_{k\geq 1}$ *for minimizing a convex function* F, *and that there exist non-negative sequences* $(T_k)_{k\geq 0}$, $(\delta_k)_{k\geq 1}$ *in* $(0, 1)$, $(\beta_k)_{k\geq 1}$ *and a scalar* $\alpha > 0$ *such that for all* $k \geq 1$,

$$\frac{\delta_k}{\alpha}\mathbb{E}\left[F(x_k) - F^*\right] + T_k \leq (1 - \delta_k)\,T_{k-1} + \beta_k, \tag{2.36}$$

where the expectation is taken with respect to any random parameter used by the algorithm. Then,

$$\mathbb{E}\left[F(x_k) - F^*\right] + \frac{\alpha}{\delta_k}T_k \leq \frac{\alpha\Gamma_k}{\delta_k}\left(T_0 + \sum_{t=1}^{k}\frac{\beta_t}{\Gamma_t}\right) \quad \textit{where} \quad \Gamma_k \triangleq \prod_{t=1}^{k}(1 - \delta_t). \tag{2.37}$$

Generic averaging strategy. *For any point $\hat{x}_0$, consider the averaging sequence $(\hat{x}_k)_{k\geq 0}$,*

$$\hat{x}_k = \Gamma_k \left(\hat{x}_0 + \sum_{t=1}^{k} \frac{\delta_t}{\Gamma_t} x_t \right) = (1-\delta_k)\, \hat{x}_{k-1} + \delta_k x_k \quad \textit{(for } k \geq 1\textit{)},$$

then,

$$\mathbb{E}\left[F(\hat{x}_k) - F^*\right] + \alpha T_k \leq \Gamma_k \left(F(\hat{x}_0) - F^* + \alpha T_0 + \alpha \sum_{t=1}^{k} \frac{\beta_t}{\Gamma_t} \right). \tag{2.38}$$

Uniform averaging strategy. *Assume that $\delta_k = \frac{1}{k+1}$ and consider the average sequence $\hat{x}_k = \frac{1}{k}\sum_{i=1}^{k} x_i$. Then,*

$$\mathbb{E}\left[F(\hat{x}_k) - F^*\right] + \alpha T_k \leq \frac{\alpha}{k} \left(T_0 + \sum_{t=1}^{k} \frac{\beta_t}{\Gamma_t} \right). \tag{2.39}$$

Proof. Given that $T_k \leq (1-\delta_k)\, T_{k-1} + \beta_k$, we obtain (2.36) by simply unrolling the recursion. To analyze the effect of the averaging strategies, divide now (2.36) by Γ_k:

$$\frac{\delta_k}{\alpha \Gamma_k} \mathbb{E}\left[F(x_k) - F^*\right] + \frac{T_k}{\Gamma_k} \leq \frac{T_{k-1}}{\Gamma_{k-1}} + \frac{\beta_k}{\Gamma_k}.$$

Sum from $t = 1$ to k and notice that we have a telescopic sum:

$$\frac{1}{\alpha} \sum_{t=1}^{k} \frac{\delta_t}{\Gamma_t} \mathbb{E}\left[F(x_t) - F^*\right] + \frac{T_k}{\Gamma_k} \leq T_0 + \sum_{t=1}^{k} \frac{\beta_t}{\Gamma_t}. \tag{2.40}$$

Then, add $(1/\alpha)\mathbb{E}\left[F(\hat{x}_0) - F^*\right]$ on both sides and multiply by $\alpha\Gamma_k$:

$$\sum_{t=1}^{k} \frac{\delta_t \Gamma_k}{\Gamma_t} \mathbb{E}\left[F(x_t) - F^*\right] + \Gamma_k \mathbb{E}\left[F(\hat{x}_0) - F^*\right] + \alpha T_k \leq \Gamma_k \left(\alpha T_0 + \mathbb{E}\left[F(\hat{x}_0) - F^*\right] + \alpha \sum_{t=1}^{k} \frac{\beta_t}{\Gamma_t} \right).$$

By exploiting the relation (2.34), we may then use Jensen's inequality and we obtain (2.38).

Consider now the specific case $\delta_k = \frac{1}{k+1}$, which yields $\Gamma_k = \frac{1}{k+1}$. Multiply then Eq. (2.40) by α/k and use Jensen's inequality; we obtain Eq. (2.39). □

2.B Relation Between Iteration (B) and MISO/SDCA

In this section, we derive explicit links between the proximal MISO algorithm [Lin et al., 2015], a primal version of SDCA [Shalev-Shwartz, 2016], and iteration (B) when used with the gradient estimator (2.14) without stochastic perturbations. Under the big data condition $L/\mu \leq n$, consider indeed $\beta = \mu$, constant step sizes $\eta_k = \eta = \frac{1}{n\mu}$, $\gamma_k = \mu$, and a uniform sampling distribution Q; then, we obtain the following algorithm

$$\bar{x}_k \leftarrow (1-\mu\eta)\bar{x}_{k-1} + \mu\eta x_{k-1} - \eta\left(\nabla f_{i_k}(x_{k-1}) - z_{k-1}^{i_k} + \bar{z}_{k-1}\right) \quad \text{and} \quad x_k = \mathrm{Prox}_{\psi/\mu}\left[\bar{x}_k\right]$$

$$\bar{z}_k = \bar{z}_{k-1} + \frac{1}{n}(z_k^{i_k} - z_{k-1}^{i_k}) \quad \text{and} \quad z_k^{i_k} = \nabla f_{i_k}(x_{k-1}) - \mu x_{k-1},$$

with $\bar{z}_0 = \bar{x}_0 = 0$. Then, since $\mu\eta = \frac{1}{n}$, it is easy to show that in fact $\bar{z}_k = \mu\bar{x}_k$ for all $k \geq 0$. This is then exactly the proximal MISO algorithm [see Bietti and Mairal, 2017]. For the relation between primal variants of SDCA and MISO, see page 4 and Equation (3) of Bietti and Mairal [2017].

2.C Recovering Classical Results for Proximal SGD

In this section, we present several corollaries of Theorem 2.1 to recover classical results for proximal variants of the stochastic gradient descent method. Throughout the section, we assume that the gradient estimates have variance bounded by σ^2:

$$\omega_k^2 = \mathbb{E}\left[\|g_k - \nabla f(x_{k-1})\|^2\right] \leq \sigma^2.$$

Convergence results for the deterministic case $\sigma^2 = 0$ can be also recovered naturally from the corollaries. We start by applying Theorem 2.1 with a constant step size strategy $\eta_k = 1/L$, which shows convergence to a noise-dominated region of radius σ^2/L. In all the corollaries below, we use the notation from Theorem 2.1.

Corollary 2.15 (**Proximal variants of SGD with constant step size,** $\mu > 0$)**.** *Assume that f is μ-strongly convex, choose $\gamma_0 = \mu$ and $\eta_k = 1/L$ with Algorithm (A) or (B). Then, for any point $\hat{x}_0$,*

$$\mathbb{E}\left[F(\hat{x}_k) - F^* + d_k(x^*) - d_k^*\right] \leq \left(1 - \frac{\mu}{L}\right)^k \left(F(\hat{x}_0) - F^* + d_0(x^*) - d_0^*\right) + \frac{\sigma^2}{L}, \qquad (2.41)$$

when using the averaging strategy from Theorem 2.1. Note that $d_k(x^*) - d_k^* \geq \frac{\mu}{2}\|x_k - x^*\|^2$ for all $k \geq 0$ with equality for Algorithm (A).

Next, we show how to obtain converging algorithms by using decreasing step sizes.

Corollary 2.16 (**Proximal variants of SGD with decreasing step sizes,** $\mu > 0$)**.** *Assume that f is μ-strongly convex and that we target an accuracy ε smaller than $2\sigma^2/L$. First, use a constant step size $\eta_k = 1/L$ with $\gamma_0 = \mu$ within Algorithm (A) or (B), using $\hat{x}_0 = x_0$, leading to the convergence rate (2.41), until $\mathbb{E}\left[F(\hat{x}_k) - F^* + d_k(x^*) - d_k^*\right] \leq 2\sigma^2/L$. Then, we restart the optimization procedure, using the previously obtained $\hat{x}_k, x_k$ as new initial points, with decreasing step sizes $\eta_k = \min\left(\frac{1}{L}, \frac{2}{\mu(k+2)}\right)$, and generate new sequences $(\hat{x}'_k, x'_k)_{k\geq 0}$. The total number of iterations to achieve $\mathbb{E}\left[F(\hat{x}'_k) - F^*\right] \leq \varepsilon$ is upper bounded by*

$$\mathcal{O}\left(\frac{L}{\mu}\log\left(\frac{F(x_0) - F^* + d_0(x^*) - d_0^*}{\varepsilon}\right)\right) + \mathcal{O}\left(\frac{\sigma^2}{\mu\varepsilon}\right). \qquad (2.42)$$

Note that $d_0(x^*) - d_0^* = \frac{\mu}{2}\|x_0 - x^*\|^2 \leq F(x_0) - F^*$ for Algorithm (A).

Proof. Given the linear convergence rate (2.41), the number of iterations of the first the constant step size strategy is upper bounded by the left term of (2.42). Then, after restarting the algorithm, we may apply Theorem 2.1 with $\mathbb{E}\left[F(\hat{x}_0) - F^* + d_0(x^*) - d_0^*\right] \leq 2\sigma^2/L$. With $\gamma_0 = \mu$, we have $\gamma_k = \mu$ for all $k \geq 0$, and the rate of Γ_k is given by Lemma 2.8,

which yields for $k \geq k_0 = \left\lceil \frac{2L}{\mu} - 2 \right\rceil$,

$$\begin{aligned}
\mathbb{E}\left[F(\hat{x}_k') - F^*\right] &\leq \Gamma_k \left(\frac{2\sigma^2}{L} + \sigma^2 \sum_{t=1}^{k} \frac{\delta_t \eta_t}{\Gamma_t} \right) \\
&= \Gamma_k \left(\frac{2\sigma^2}{L} + \frac{\sigma^2}{L} \sum_{t=1}^{k_0-1} \frac{\delta_t}{\Gamma_t} + \sigma^2 \sum_{t=k_0}^{k} \frac{2\delta_t}{\Gamma_t \mu (t+2)} \right) \\
&= \frac{k_0(k_0+1)}{(k+1)(k+2)} \left(\Gamma_{k_0-1} \frac{2\sigma^2}{L} + \frac{\sigma^2}{L} \Gamma_{k_0-1} \sum_{t=1}^{k_0-1} \frac{\delta_t}{\Gamma_t} \right) + \sigma^2 \sum_{t=k_0}^{k} \frac{2\delta_t \Gamma_k}{\Gamma_t \mu (t+2)} \\
&= \frac{k_0(k_0+1)}{(k+1)(k+2)} \left(\Gamma_{k_0-1} \frac{2\sigma^2}{L} + (1 - \Gamma_{k_0-1}) \frac{\sigma^2}{L} \right) + \sigma^2 \sum_{t=k_0}^{k} \frac{2\delta_t \Gamma_k}{\Gamma_t \mu (t+2)} \\
&\leq \frac{k_0(k_0+1)}{(k+1)(k+2)} \frac{2\sigma^2}{L} + \sigma^2 \frac{1}{(k+1)(k+2)} \left(\sum_{t=k_0+1}^{k} \frac{4(t+1)(t+2)}{\mu (t+2)^2} \right) \\
&\leq \frac{k_0}{(k+1)(k+2)} \frac{4\sigma^2}{\mu} + \frac{4\sigma^2}{\mu(k+2)},
\end{aligned}$$

where the second inequality uses the fact that $(\mu/2)\left\|x_0 - x^*\right\|^2 \leq F(x_0) - F^* \leq \frac{2\sigma^2}{L}$, and then we use Lemmas 2.8 and 2.9. The term on the right is of order $\mathcal{O}\left(\sigma^2/\mu k\right)$ whereas the term on the left becomes of the same order or smaller whenever $k \geq k_0 = \mathcal{O}\left(L/\mu\right)$. This leads to the desired iteration complexity. □

We may now study the case $\mu = 0$, first with a constant step size. The next corollary consists of simply applying the uniform averaging strategy of Lemma 2.12 to Proposition 2.1, noting that $\delta_k = \frac{1}{k+1}$ for all $k \geq 0$ if $\mu = 0$ and $\gamma_0 = 1/\eta$.

Corollary 2.17 (**Proximal variants of SGD with constant step size, $\mu = 0$**). *Assume that f is convex, choose a constant step size $\eta_k = \eta \leq \frac{1}{L}$ with Algorithm (A) or (B) with $\gamma_0 = 1/\eta$.*

Then,

$$\mathbb{E}\left[F(\hat{x}_k) - F^*\right] \leq \frac{d_0(x^*) - d_0^*}{k} + \eta\sigma^2, \tag{2.43}$$

where $\hat{x}_k = \frac{1}{k}\sum_{i=1}^{k} x_i$. Note that $d_0(x^*) - d_0^* = \frac{1}{2\eta}\left\|x_0 - x^*\right\|^2$ for Algorithm (A).

The noise dependency is now illustrated for Algorithm (A) in the next corollary, obtained in a finite horizon setting.

Corollary 2.18 (**Proximal variants of SGD with $\mu = 0$, finite horizon**). *Consider the same setting as in the previous corollary. Assume that we have a budget of K iterations for Algorithm (A). Choose a constant step size*

$$\eta_k = \min\left(\frac{1}{L}, \sqrt{\frac{T_0}{K\sigma^2}} \right) \quad \textit{with} \quad T_0 = \frac{1}{2}\left\|x_0 - x^*\right\|^2.$$

Then, with $\gamma_0 = 1/\eta$ and when using the averaging strategy from Corollary 2.17,

$$\mathbb{E}\left[F(\hat{x}_K) - F^*\right] \leq \frac{LT_0}{K} + 2\sigma\sqrt{\frac{T_0}{K}}. \tag{2.44}$$

This corollary is obtained by optimizing the right side of (2.43) with respect to η under the constraint $\eta \leq 1/L$. Considering both cases $\eta = 1/L$ and $\eta = \sqrt{T_0/K\sigma^2}$, it is easy to check that we have (2.44) in all cases. Whereas this last result is not a practical one since the step size depends on unknown quantities, it shows that our analysis is nevertheless able to recover the optimal noise-dependency in $\mathcal{O}\left(\sigma\sqrt{T_0/K}\right)$, [see Nemirovski et al., 2009].

2.D Proofs of the Main Results

2.D.1 Proof of Proposition 2.2

Proof. The proof borrows a large part of the analysis of Xiao and Zhang [2014] for controlling the variance of the gradient estimate in the SVRG algorithm. First, we note that all the gradient estimators we consider may be written in the generic form (2.14), with $\beta = 0$ for SAGA or SVRG. Then, we will write $\tilde{\nabla} f_{i_k}(x_{k-1}) = \nabla f_{i_k}(x_{k-1}) + \zeta_k$, where ζ_k is a zero-mean variable with variance $\tilde{\sigma}^2$ drawn at iteration k, and $z_k^i = u_k^i + \zeta_k^i$ for all k, i, where ζ_k^i has zero-mean with variance $\tilde{\sigma}^2$ and was drawn during the previous iterations. Let us denote by $\omega_k^2 = \mathbb{E}\left[\|g_k - f(x_{k-1})\|^2\right]$ and let us introduce the quantity $A_k = \mathbb{E}\left[\frac{1}{(q_{i_k}n)^2}\|\zeta_k\|^2\right]$. Then,

$$\begin{aligned}
\omega_k^2 &= \mathbb{E}\left\|\frac{1}{q_{i_k}n}\left(\tilde{\nabla} f_{i_k}(x_{k-1}) - \beta x_{k-1} - z_{k-1}^{i_k}\right) + \bar{z}_{k-1} + \beta x_{k-1} - \nabla f(x_{k-1})\right\|^2 \\
&= \mathbb{E}\left\|\frac{1}{q_{i_k}n}\left(\nabla f_{i_k}(x_{k-1}) - \beta x_{k-1} - z_{k-1}^{i_k}\right) + \bar{z}^{k-1} + \beta x_{k-1} - \nabla f(x_{k-1})\right\|^2 \\
&\qquad\qquad + \mathbb{E}\left[\frac{1}{(q_{i_k}n)^2}\|\zeta_k\|^2\right] \\
&\leq \mathbb{E}\left\|\frac{1}{q_{i_k}n}\left(\nabla f_{i_k}(x_{k-1}) - \beta x_{k-1} - z_{k-1}^{i_k}\right)\right\|^2 + A_k \\
&= \frac{1}{n}\sum_{i=1}^n \frac{1}{q_i n}\mathbb{E}\left[\left\|\nabla f_i(x_{k-1}) - \beta x_{k-1} - z_{k-1}^i\right\|^2\right] + A_k \\
&= \frac{1}{n}\sum_{i=1}^n \frac{1}{q_i n}\mathbb{E}\left[\left\|\nabla f_i(x_{k-1}) - \beta x_{k-1} - u_*^i + u_*^i - z_{k-1}^i\right\|^2\right] + A_k \\
&\leq \frac{2}{n}\sum_{i=1}^n \frac{1}{q_i n}\mathbb{E}\left[\left\|\nabla f_i(x_{k-1}) - \beta x_{k-1} - u_*^i\right\|^2\right] + \frac{2}{n}\sum_{i=1}^n \frac{1}{q_i n}\mathbb{E}\left[\left\|z_{k-1}^i - u_*^i\right\|^2\right] + A_k
\end{aligned}$$

$$\leq \frac{2}{n}\sum_{i=1}^{n}\frac{1}{q_i n}\mathbb{E}\left[\|\nabla f_i(x_{k-1}) - \nabla f_i(x^*) - \beta(x_{k-1}-x^*)\|^2\right] + \frac{2}{n}\sum_{i=1}^{n}\frac{1}{q_i n}\mathbb{E}\left[\left\|u_{k-1}^i - u_*^i\right\|^2\right] + 3A_k$$

$$\leq \frac{4}{n}\sum_{i=1}^{n}\frac{L_i}{q_i n}\mathbb{E}\left[f_i(x_{k-1}) - f_i(x^*) - \nabla f_i(x^*)^\top(x_{k-1}-x^*)\right] + \frac{2}{n}\sum_{i=1}^{n}\frac{1}{q_i n}\mathbb{E}\left[\left\|u_{k-1}^i - u_*^i\right\|^2\right] + 3A_k$$

$$\leq 4L_Q\mathbb{E}\left[f(x_{k-1}) - f(x^*) - \nabla f(x^*)^\top(x_{k-1}-x^*)\right] + \frac{2}{n}\sum_{i=1}^{n}\frac{1}{q_i n}\mathbb{E}\left[\|u_{k-1}^i - u_*^i\|^2\right] + 3A_k, \tag{2.45}$$

where the first inequality uses the relation $\mathbb{E}\left[\|X - \mathbb{E}[X]\|^2\right] \leq \mathbb{E}\left[\|X\|^2\right]$ for all random variable X, taking here expectation with respect to the index $i_k \sim Q$ and conditioning on $\mathcal{F}_{k-1}$; the second inequality uses the relation $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$; the last inequality uses Lemma 2.7.

We have now two possibilities to control the quantity A_k related to ζ_k. First, we may simply upper bound it as follows

$$A_k = \mathbb{E}\left[\frac{1}{(q_{i_k}n)^2}\|\zeta_k\|^2\right] \leq \rho_Q\tilde{\sigma}^2.$$

Then, since x^* minimizes F, we have $0 \in \nabla f(x^*) + \partial\psi(x^*)$ and thus $-\nabla f(x^*)$ is a subgradient in $\partial\psi(x^*)$. By using as well the convexity inequality $\psi(x) \geq \psi(x^*) - \nabla f(x^*)^\top(x-x^*)$, we have

$$f(x_{k-1}) - f(x^*) - \nabla f(x^*)^\top(x_{k-1}-x^*) \leq F(x_{k-1}) - F^*, \tag{2.46}$$

leading finally to (2.20).

The second possibility is to relate A_k to $\tilde{\sigma}_*^2$, under the assumption that each f_i may be written as $f_i(x) = \mathbb{E}_\xi\left[\tilde{f}_i(x,\xi)\right], i \in [1,\dots,n]$ with $\tilde{f}_i(.,\xi)$ L_i-smooth with $L_i \geq \mu$ for all ξ. Then,

$$\mathbb{E}\left[\frac{1}{(q_{i_k}n)^2}\|\zeta_k\|^2\right] = \mathbb{E}\left[\frac{1}{(q_{i_k}n)^2}\left\|\tilde{\nabla} f_{i_k}(x_{k-1}) - \nabla f_{i_k}(x_{k-1})\right\|^2\right]$$

$$= \mathbb{E}\Bigg[\frac{1}{(q_{i_k}n)^2}\Big\|\tilde{\nabla} f_{i_k}(x_{k-1}) - \tilde{\nabla} f_{i_k}(x^*) + \tilde{\nabla} f_{i_k}(x^*) - \nabla f_{i_k}(x^*) + \nabla f_{i_k}(x^*) - \nabla f_{i_k}(x_{k-1})\Big\|\Bigg]$$

$$\leq \mathbb{E}\left[\frac{1}{(q_{i_k}n)^2}\left[\left\|\tilde{\nabla} f_{i_k}(x_{k-1}) - \tilde{\nabla} f_{i_k}(x^*) + \tilde{\nabla} f_{i_k}(x^*) - \nabla f_{i_k}(x^*)\right\|^2\right]\right]$$

$$\leq 2\mathbb{E}\left[\frac{1}{(q_{i_k}n)^2}\left[\left\|\tilde{\nabla} f_{i_k}(x_{k-1}) - \tilde{\nabla} f_{i_k}(x^*)\right\|^2 + \left\|\tilde{\nabla} f_{i_k}(x^*) - \nabla f_{i_k}(x^*)\right\|^2\right]\right]$$

$$
\begin{aligned}
&\leq 4\mathbb{E}\left[\frac{L_{i_k}}{(q_{i_k}n)^2}\left(f_{i_k}(x_{k-1}) - f_{i_k}^* - \langle \nabla f_{i_k}(x^*), x_{k-1} - x^*\rangle\right)\right] + 2\mathbb{E}\left[\frac{1}{(q_{i_k}n)^2}\tilde{\sigma}_{i_k,*}^2\right] \\
&\leq 4L_Q\mathbb{E}\left[\frac{1}{q_{i_k}n}\left(f_{i_k}(x_{k-1}) - f_{i_k}^* - \langle \nabla f_{i_k}(x^*), x_{k-1} - x^*\rangle\right)\right] + 2\rho_Q\tilde{\sigma}_*^2 \\
&= 4L_Q\left(f(x_{k-1}) - f^* - \langle \nabla f(x^*), x_{k-1} - x^*\rangle\right) + 2\rho_Q\tilde{\sigma}_*^2, \qquad (2.47)
\end{aligned}
$$

where we use the relation $\mathbb{E}\left[\|X - \mathbb{E}[X]\|^2\right] \leq \mathbb{E}\left[\|X\|^2\right]$ for the first inequality, the well-known inequality for a convex norm $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ for the second inequality and the definition $\tilde{\sigma}_* = \frac{1}{n}\sum_{i=1}^n \tilde{\sigma}_{i,*}^2$.

Then, we may combine (2.47) with (2.45) and use (2.46) to obtain (2.21).

□

2.D.2 Proof of Proposition 2.3

Proof. To make the notation more compact, we call

$$
F_k = \mathbb{E}\left[F(x_k) - F^*\right], \qquad D_k = \mathbb{E}\left[d_k(x^*) - d_k^*\right] \qquad \text{and} \qquad C_k = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \frac{1}{q_i n}\|u_k^i - u_*^i\|^2\right].
$$

Then, according to Proposition 2.2, we have

$$
\omega_k^2 \leq 4L_Q F_{k-1} + 2C_{k-1} + 3\rho_Q\tilde{\sigma}^2,
$$

and according to Proposition 2.1,

$$
\delta_k F_k + D_k \leq (1-\delta_k)\, D_{k-1} + 4L_Q\eta_k\delta_k F_{k-1} + 2\eta_k\delta_k C_{k-1} + 3\rho_Q\eta_k\delta_k\tilde{\sigma}^2. \qquad (2.48)
$$

Then, we note that both for the SVRG and SAGA/MISO/SDCA strategies, we have (with $\beta = 0$ for SVRG),

$$
\mathbb{E}\left[\|u_k^i - u_*^i\|^2\right] = \left(1 - \frac{1}{n}\right)\mathbb{E}\left[\|u_{k-1}^i - u_*^i\|^2\right] + \frac{1}{n}\mathbb{E}\|\nabla f_i(x_k) - \nabla f_i(x^*) + \beta(x_k - x^*)\|^2.
$$

By taking a weighted average, this yields

$$
\begin{aligned}
C_k &\leq \left(1 - \frac{1}{n}\right)C_{k-1} + \frac{1}{n^2}\sum_{i=1}^n \frac{1}{q_i n}\mathbb{E}\left[\|\nabla f_i(x_k) - \nabla f_i(x^*) - \beta(x_k - x^*)\|^2\right] \\
&\leq \left(1 - \frac{1}{n}\right)C_{k-1} + \frac{1}{n^2}\sum_{i=1}^n \frac{2L_i}{q_i n}\mathbb{E}\left[f_i(x_k) - f_i(x^*) - \nabla f_i(x^*)^\top(x_k - x^*)\right] \\
&\leq \left(1 - \frac{1}{n}\right)C_{k-1} + \frac{2L_Q F_k}{n},
\end{aligned}
$$

where the second inequality comes from Lemma 2.7 and the last one uses similar arguments as in the proof of Proposition 2.2. Then, we add a quantity $\beta_k C_k$ on both sides of the relation (2.48) with some $\beta_k > 0$ that we will specify later:

$$
\begin{aligned}
&\left(\delta_k - \beta_k\frac{2L_Q}{n}\right)F_k + D_k + \beta_k C_k \\
&\qquad \leq (1-\delta_k)\, D_{k-1} + \left(\beta_k\left(1 - \frac{1}{n}\right) + 2\eta_k\delta_k\right)C_{k-1} + 4L_Q\eta_k\delta_k F_{k-1} + 3\rho_Q\eta_k\delta_k\tilde{\sigma}^2,
\end{aligned}
$$

and then choose $\beta_k/n = (5/2)\eta_k\delta_k$, which yields

$$\begin{aligned}\delta_k\left(1-5L_Q\eta_k\right)F_k + D_k + \beta_k C_k \\ \leq \left(1-\delta_k\right)D_{k-1} + \beta_k\left(1-\frac{1}{5n}\right)C_{k-1} + 4L_Q\eta_k\delta_k F_{k-1} + 3\rho_Q\eta_k\delta_k\tilde{\sigma}^2.\end{aligned}$$

Remember that $\tau_k = \min\left(\delta_k, \frac{1}{5n}\right)$, notice that the sequences $(\beta_k)_{k\geq 0}, (\eta_k)_{k\geq 0}$ and $(\delta_k)_{k\geq 0}$ are non-increasing and note that $4 \leq 5\left(1-\frac{1}{5n}\right)$ for all $n \geq 1$. Then,

$$\begin{aligned}\delta_k\left(1-10L_Q\eta_k\right)F_k + \underbrace{5L_Q\eta_k\delta_k + D_k + \beta_k C_k}_{T_k} \\ \leq \left(1-\tau_k\right)\left(D_{k-1} + \beta_{k-1}C_{k-1} + 5L_Q\eta_{k-1}\delta_{k-1}F_{k-1}\right) + 3\rho_Q\eta_k\delta_k\tilde{\sigma}^2,\end{aligned}$$

which immediately yields (2.22) with the appropriate definition of T_k, and by noting that $(1-10L_Q\eta_k) \geq \frac{1}{6}$. □

2.D.3 Proof of Theorem 2.2

Proof. The first part of the theorem is a direct application of Lemma 2.12 to Proposition 2.3, by noting that (2.36) holds—when replacing the notation δ_t by τ_t in (2.36)—since for a fixed number of iterations K, we have the relation $\frac{\tau_k\delta_K}{6\tau_K}\mathbb{E}\left[F(x_k) - F^*\right] + T_k \leq \left(1-\tau_k\right)T_{k-1} + 3\rho_Q\eta_k\delta_k\tilde{\sigma}^2$ for all $k \leq K$. Indeed, $\delta_k = \frac{\tau_k\delta_k}{\tau_k} \geq \frac{\tau_k\delta_K}{\tau_K}$ since the ratio δ_t/τ_t is non-increasing. Then, we may now prove (2.24):

$$\begin{aligned}T_0 &= 5L_Q\eta_0\delta_0\left(F(x_0) - F^*\right) + d_0(x^*) - d_0^* + \frac{5\eta_0\delta_0}{2}\frac{1}{n}\sum_{i=1}^n \frac{1}{q_i n}\left\|u_0^i - u_*^i\right\|^2 \\ &\leq 5L_Q\eta_0\delta_0\left(F(x_0) - F^*\right) + d_0(x^*) - d_0^* \\ &\qquad + \frac{5\eta_0\delta_0}{2}\frac{1}{n}\sum_{i=1}^n \frac{2L_i}{q_i n}\left(f_i(x_0) - f_i(x^*) - \nabla f_i(x^*)^\top(x_0 - x^*)\right) \\ &\leq 5L_Q\eta_0\delta_0\left(F(x_0) - F^*\right) + d_0(x^*) - d_0^* + 5\eta_0\delta_0 L_Q(f(x_0) - f(x^*) - \nabla f(x^*)^\top(x_0 - x^*)) \\ &\leq 10L_Q\eta_0\delta_0\left(F(x_0) - F^*\right) + d_0(x^*) - d_0^*,\end{aligned}$$

where the first inequality uses Lemma 2.7, and the second one uses the definition of L_Q, whereas the last one uses (2.46). □

2.D.4 Proof of Corollary 2.3

Proof. First, notice that $\delta_k = \eta_k\gamma_k = \frac{\mu}{12L_Q}$ and that $\alpha = \frac{6\tau_k}{\delta_k}$. Then, we apply Theorem 2.2 and obtain

$$\begin{aligned}\mathbb{E}\left[F(\hat{x}_k) - F^* + \alpha T_k\right] &\leq \Theta_k\left(F(\hat{x}_0) - F^* + \alpha T_0 + \frac{18\rho_Q\tau_k\tilde{\sigma}^2}{\delta_k}\sum_{t=1}^k \frac{\eta_t\delta_t}{\Theta_t}\right) \\ &= \Theta_k\left(F(\hat{x}_0) - F^* + \alpha T_0 + \frac{3\rho_Q\tilde{\sigma}^2}{2L_Q}\sum_{t=1}^k \frac{\tau_t}{\Theta_t}\right) \\ &\leq \Theta_k\left(F(\hat{x}_0) - F^* + \alpha T_0\right) + \frac{3\rho_Q\tilde{\sigma}^2}{2L_Q}.\end{aligned}$$

□

2.D.5 Proof of Corollary 2.5

Proof. Since the convergence rate (2.26) applies for the first stage with a constant step size, the number of iterations to ensure the condition $\mathbb{E}\left[F(\hat{x}_k) - F^* + 6T_k\right] \leq 24\eta\rho_Q\tilde{\sigma}^2$ is upper bounded by K with

$$K = \mathcal{O}\left(\left(n + \frac{L_Q}{\mu}\right)\log\left(\frac{F(x_0) - F^* + d_0(x^*) - d_0^*}{\varepsilon}\right)\right),$$

when using the upper-bound (2.24) on T_0. Then, we restart the optimization procedure, using $x_0' = x_K$ and $\hat{x}_0' = \hat{x}_K'$, assuming from now on that $\mathbb{E}\left[F(\hat{x}_0') - F^* + 6T_0'\right] \leq 24\eta\rho_Q\tilde{\sigma}^2$, with decreasing step sizes $\eta_k = \min\left(\frac{2}{\mu(k+2)}, \eta\right)$. Then, since $\delta_k = \mu\eta_k \leq \frac{1}{5n}$, we have that $\tau_k = \delta_k$ for all k, and Theorem 2.2 gives us—note that here $\Gamma_k = \Theta_k$—

$$\mathbb{E}\left[F(\hat{x}_k') - F^*\right] \leq \Gamma_k\left(F(\hat{x}_0') - F^* + 6T_0' + 18\rho_Q\tilde{\sigma}^2\sum_{t=1}^{k}\frac{\eta_t\delta_t}{\Gamma_t}\right) \quad \text{with} \quad \Gamma_k = \prod_{t=1}^{k}(1-\delta_t).$$

Then, after taking the expectation with respect to the output of the first stage,

$$\mathbb{E}\left[F(\hat{x}_k') - F^*\right] \leq \Gamma_k\left(24\rho_Q\eta\tilde{\sigma}^2 + 18\rho_Q\tilde{\sigma}^2\sum_{t=1}^{k}\frac{\eta_t\delta_t}{\Gamma_t}\right).$$

Denote now by k_0 the largest index such that $\frac{2}{\mu(k_0+2)} \geq \eta$ and thus $k_0 = \lceil 2/(\mu\eta) - 2\rceil$. Then, according to Lemma 2.8, for $k \geq k_0$,

$$\begin{aligned}
\mathbb{E}\left[F(\hat{x}_k) - F^*\right] &\leq \Gamma_k\left(24\rho_Q\eta\tilde{\sigma}^2 + 18\rho_Q\eta\tilde{\sigma}^2\sum_{t=1}^{k_0-1}\frac{\delta_t}{\Gamma_t} + 18\rho_Q\tilde{\sigma}^2\sum_{t=k_0}^{k}\frac{2\delta_t}{\mu\Gamma_t(t+2)}\right) \\
&\leq \frac{k_0(k_0+1)}{(k+1)(k+2)}\left(\Gamma_{k_0-1}24\rho_Q\eta\tilde{\sigma}^2 + 18\eta\rho_Q\tilde{\sigma}^2\Gamma_{k_0-1}\sum_{t=1}^{k_0-1}\frac{\delta_t}{\Gamma_t}\right) \\
&\qquad + 36\rho_Q\tilde{\sigma}^2\sum_{t=k_0}^{k}\frac{\delta_t\Gamma_k}{\mu\Gamma_t(t+2)} \\
&\leq \frac{k_0(k_0+1)}{(k+1)(k+2)}24\eta\rho_Q\tilde{\sigma}^2 + 36\rho_Q\tilde{\sigma}^2\sum_{t=k_0}^{k}\frac{(t+1)(t+2)}{\mu(k+1)(k+2)(t+2)^2} \\
&\leq \frac{k_0\eta}{k+2}24\rho_Q\tilde{\sigma}^2 + \frac{36\rho_Q\tilde{\sigma}^2}{\mu(k+2)} = \mathcal{O}\left(\frac{\rho_Q\tilde{\sigma}^2}{\mu k}\right),
\end{aligned}$$

which gives the desired complexity. □

2.D.6 Proof of Corollary 2.6

Proof. Let us call x_0' the point obtained by running one iteration of (A) with step size $\eta \leq \frac{1}{12L_Q}$ and gradient estimator $(1/n)\sum_{i=1}^{n}\tilde{\nabla}f_i(x_0)$, whose variance is $\tilde{\sigma}^2/n$. Then, since

$\delta_1 = \Gamma_1 = 1/2$, according to Theorem 2.1, we have

$$\mathbb{E}\left[F(x_0') - F^* + \frac{1}{2\eta}\|x_0' - x^*\|^2\right] \leq \frac{1}{2\eta}\|x_0 - x^*\|^2 + \frac{\eta\tilde{\sigma}^2}{n}. \tag{2.49}$$

Then, we consider the main run of the algorithm, and apply Theorem 2.2, replacing x_0 by x_0'. With the chosen setup, we have $\delta_k = \frac{1}{k+1}$ and since $K \geq 5n$, we have $\delta_K = \tau_K$, such that (2.23) becomes

$$\mathbb{E}\left[F(\hat{x}_K) - F^*\right] \leq \Theta_K\left(F(x_0') - F^* + 6T_0 + 18\rho_Q\eta\tilde{\sigma}^2\sum_{t=1}^{k}\frac{\delta_t}{\Theta_t}\right),$$

and from (2.24), we have

$$T_0 \leq 10L_Q\eta(F(x_0') - F^*) + \frac{1}{2\eta}\|x_0' - x^*\|^2 \leq \frac{5}{6}(F(x_0') - F^*) + \frac{1}{2\eta}\|x_0' - x^*\|^2,$$

which yields, combined with (2.49),

$$\mathbb{E}\left[F(x_0') - F^* + 6T_0\right] \leq 6\mathbb{E}\left[F(x_0') - F^* + \frac{1}{2\eta}\|x_0' - x^*\|^2\right] \leq \frac{3}{\eta}\|x_0 - x^*\|^2 + \frac{6\eta\tilde{\sigma}^2}{n}.$$

Note that Lemma 2.8 gives us that $\Theta_k = (1 - 1/5n)^{5n-1}\frac{5n}{k+1} \leq \frac{3n}{k+1}$ for $k \geq 5n$ and since $1 + \sum_{t=1}^{K}\frac{\tau_t}{\Theta_t} = \frac{1}{\Theta_K}$ according to Lemma 2.9,

$$\begin{aligned}
\mathbb{E}\left[F(\hat{x}_K) - F^*\right] &\leq \Theta_K\left(\frac{3}{\eta}\|x_0 - x^*\|^2 + \frac{6\eta\tilde{\sigma}^2}{n} + 18\rho_Q\eta\tilde{\sigma}^2\sum_{t=1}^{K}\frac{\delta_t}{\Theta_t}\right) \\
&\leq \frac{9n}{\eta(K+1)}\|x_0 - x^*\|^2 + 6\eta\tilde{\sigma}^2\rho_Q\Theta_K\left(\frac{1}{n} + 3\sum_{t=1}^{K}\frac{\tau_t}{\Theta_t} + 3\sum_{t=1}^{5n-1}\frac{\delta_t}{\Theta_t}\right) \\
&\leq \frac{9n}{\eta(K+1)}\|x_0 - x^*\|^2 + 6\eta\tilde{\sigma}^2\rho_Q\left(\frac{\Theta_K}{n} + 3(1 - \Theta_K) + \frac{15n}{K+1}\sum_{t=1}^{5n-1}\delta_t\right) \\
&\leq \frac{9n}{\eta(K+1)}\|x_0 - x^*\|^2 + 18\eta\tilde{\sigma}^2\rho_Q\left(1 + \frac{5n}{K+1}\log(5n)\right) \\
&\leq \frac{9n}{\eta(K+1)}\|x_0 - x^*\|^2 + 36\eta\tilde{\sigma}^2\rho_Q.
\end{aligned}$$

It remains to optimize it over η to get the left side of (2.27). □

2.D.7 Proof of Lemma 2.1

Proof. Let us assume that the relation $y_{k-1} = (1 - \theta_{k-1})x_{k-1} + \theta_{k-1}v_{k-1}$ holds and let us show that it also holds for y_k. Since the estimate sequences d_k are quadratic functions, we

have

$$\begin{aligned}
v_k &= (1-\delta_k)\,\frac{\gamma_{k-1}}{\gamma_k}v_{k-1} + \frac{\mu\delta_k}{\gamma_k}y_{k-1} - \frac{\delta_k}{\gamma_k}(g_k + \psi'(x_k)) \\
&= (1-\delta_k)\,\frac{\gamma_{k-1}}{\gamma_k}v_{k-1} + \frac{\mu\delta_k}{\gamma_k}y_{k-1} - \frac{\delta_k}{\gamma_k\eta_k}(y_{k-1} - x_k) \\
&= (1-\delta_k)\,\frac{\gamma_{k-1}}{\gamma_k\theta_{k-1}}\left(y_{k-1} - (1-\theta_{k-1})x_{k-1}\right) + \frac{\mu\delta_k}{\gamma_k}y_{k-1} - \frac{\delta_k}{\gamma_k\eta_k}(y_{k-1} - x_k) \\
&= (1-\delta_k)\,\frac{\gamma_{k-1}}{\gamma_k\theta_{k-1}}\left(y_{k-1} - (1-\theta_{k-1})x_{k-1}\right) + \frac{\mu\delta_k}{\gamma_k}y_{k-1} - \frac{1}{\delta_k}(y_{k-1} - x_k) \\
&= \left(\frac{(1-\delta_k)\,\gamma_{k-1}}{\gamma_k\theta_{k-1}} + \frac{\mu\delta_k}{\gamma_k} - \frac{1}{\delta_k}\right)y_{k-1} - \frac{(1-\delta_k)\,\gamma_{k-1}(1-\theta_{k-1})}{\gamma_k\theta_{k-1}}x_{k-1} + \frac{1}{\delta_k}x_k \\
&= \left(1 + \frac{(1-\delta_k)\,\gamma_{k-1}(1-\theta_{k-1})}{\gamma_k\theta_{k-1}} - \frac{1}{\delta_k}\right)y_{k-1} - \frac{(1-\delta_k)\,\gamma_{k-1}(1-\theta_{k-1})}{\gamma_k\theta_{k-1}}x_{k-1} + \frac{1}{\delta_k}x_k.
\end{aligned}$$

Then note that $\theta_{k-1} = \frac{\delta_k\gamma_{k-1}}{\gamma_{k-1}+\delta_k\mu}$ and thus, $\frac{\gamma_{k-1}(1-\theta_{k-1})}{\gamma_k\theta_{k-1}} = \frac{1}{\delta_k}$, and

$$v_k = x_{k-1} + \frac{1}{\delta_k}(x_k - x_{k-1}).$$

Then, we note that $x_k - x_{k-1} = \frac{\delta_k}{1-\delta_k}(v_k - x_k)$ and we are left with

$$y_k = x_k + \beta_k(x_k - x_{k-1}) = \frac{\beta_k\delta_k}{1-\delta_k}v_k + \left(1 - \frac{\beta_k\delta_k}{1-\delta_k}\right)x_k.$$

Then, it is easy to show that

$$\beta_k = \frac{(1-\delta_k)\,\delta_{k+1}\gamma_k}{\delta_k(\gamma_{k+1} + \delta_{k+1}\gamma_k)} = \frac{(1-\delta_k)\,\delta_{k+1}\gamma_k}{\delta_k(\gamma_k + \delta_{k+1}\mu)} = \frac{(1-\delta_k)\,\theta_k}{\delta_k},$$

which allows us to conclude that $y_k = (1-\theta_k)x_k + \theta_k v_k$ since the relation holds trivially for $k = 0$. □

2.D.8 Proof of Lemma 2.2

Proof.

$$\begin{aligned}
\mathbb{E}\left[F(x_k)\right] &= \mathbb{E}\left[f(x_k) + \psi(x_k)\right] \\
&\leq \mathbb{E}\left[f(y_{k-1}) + \nabla f(y_{k-1})^\top(x_k - y_{k-1}) + \frac{L}{2}\|x_k - y_{k-1}\|^2 + \psi(x_k)\right] \\
&= \mathbb{E}\left[f(y_{k-1}) + g_k^\top(x_k - y_{k-1}) + \frac{L}{2}\|x_k - y_{k-1}\|^2 + \psi(x_k)\right] \\
&\qquad + \mathbb{E}\left[(\nabla f(y_{k-1}) - g_k)^\top(x_k - y_{k-1})\right]
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\left[f(y_{k-1}) + g_k^\top(x_k - y_{k-1}) + \frac{L}{2}\|x_k - y_{k-1}\|^2 + \psi(x_k)\right] \\
&\qquad\qquad + \mathbb{E}\left[(\nabla f(y_{k-1}) - g_k)^\top x_k\right] \\
&= \mathbb{E}\left[f(y_{k-1}) + g_k^\top(x_k - y_{k-1}) + \frac{L}{2}\|x_k - y_{k-1}\|^2 + \psi(x_k)\right] \\
&\qquad\qquad + \mathbb{E}\left[(\nabla f(y_{k-1}) - g_k)^\top (x_k - w_k)\right] \\
&\leq \mathbb{E}\left[f(y_{k-1}) + g_k^\top(x_k - y_{k-1}) + \frac{L}{2}\|x_k - y_{k-1}\|^2 + \psi(x_k)\right] \\
&\qquad\qquad + \mathbb{E}\left[\|\nabla f(y_{k-1}) - g_k\|\|x_k - w_k\|\right] \\
&\leq \mathbb{E}\left[f(y_{k-1}) + g_k^\top(x_k - y_{k-1}) + \frac{L}{2}\|x_k - y_{k-1}\|^2 + \psi(x_k)\right] \\
&\qquad\qquad + \mathbb{E}\left[\eta_k\|\nabla f(y_{k-1}) - g_k\|^2\right] \\
&\leq \mathbb{E}\left[f(y_{k-1}) + g_k^\top(x_k - y_{k-1}) + \frac{L}{2}\|x_k - y_{k-1}\|^2 + \psi(x_k)\right] + \eta_k\omega_k^2,
\end{aligned}
$$

where $w_k = \mathrm{Prox}_{\eta_k\psi}[y_{k-1} - \eta_k\nabla f(y_{k-1})]$. The first inequality is due to the L-smoothness of f (Lemma 2.4); then, the next three relations exploit the fact that $\mathbb{E}\left[(\nabla f(y_{k-1}) - g_k)^\top z\right] = 0$ for all z that is deterministic with respect to the algebra $\mathcal{F}_{k-1}$; the third inequality uses the non-expansiveness of the proximal operator. Using the definition (2.28) for l_k, we proceed with

$$
\begin{aligned}
\mathbb{E}\left[F(x_k)\right] &\leq \mathbb{E}\left[f(y_{k-1}) + g_k^\top(x_k - y_{k-1}) + \frac{L}{2}\|x_k - y_{k-1}\|^2 + \psi(x_k)\right] + \eta_k\omega_k^2, \\
&= \mathbb{E}\left[l_k(y_{k-1}) + \tilde{g}_k^\top(x_k - y_{k-1}) + \frac{L}{2}\|x_k - y_{k-1}\|^2\right] + \eta_k\omega_k^2, \\
&\leq \mathbb{E}\left[l_k(y_{k-1})\right] + \left(\frac{L\eta_k^2}{2} - \eta_k\right)\mathbb{E}\left[\|\tilde{g}_k\|^2\right] + \eta_k\omega_k^2,
\end{aligned}
$$

where we use the fact that $x_k = y_{k-1} - \eta_k\tilde{g}_k$ and $\tilde{g}_k = g_k + \psi'(x_k)$. □

2.D.9 Proof of Corollary 2.9

Proof. The proof is similar to that of Corollary 2.16 for unaccelerated SGD. The first stage with constant step size requires $\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{F(x_0)-F^*}{\varepsilon}\right)\right)$ iterations. Then, we restart the optimization procedure, and assume that $\mathbb{E}\left[F(x_0) - F^* + \frac{\mu}{2}\|x^* - x_0\|^2\right] \leq \frac{2\sigma^2}{\sqrt{\mu L}}$. With the choice of parameters, we have $\gamma_k = \mu$ and $\delta_k = \sqrt{\gamma_k\eta_k} = \min\left(\sqrt{\frac{\mu}{L}}, \frac{2}{k+2}\right)$. We may then apply Theorem 2.7 where the value of Γ_k is given by Lemma 2.8. This yields for

$k \geq k_0 = \left\lceil 2\sqrt{\frac{L}{\mu}} - 2 \right\rceil$,

$$\begin{aligned}
\mathbb{E}\left[F(x_k) - F^*\right] &\leq \Gamma_k \left(\mathbb{E}\left[F(x_0) - F^* + \frac{\mu}{2}\left\|x_0 - x^*\right\|^2 \right] + \sigma^2 \sum_{t=1}^{k} \frac{\eta_t}{\Gamma_t} \right) \\
&\leq \Gamma_k \left(\frac{2\sigma^2}{\sqrt{\mu L}} + \frac{\sigma^2}{L} \sum_{t=1}^{k_0-1} \frac{1}{\Gamma_t} + \sigma^2 \sum_{t=k_0}^{k} \frac{4}{\Gamma_t \mu (t+2)^2} \right) \\
&= \frac{k_0(k_0+1)}{(k+1)(k+2)} \left(\Gamma_{k_0-1} \frac{2\sigma^2}{\sqrt{\mu L}} + \frac{\sigma^2}{L} \Gamma_{k_0-1} \sum_{t=1}^{k_0-1} \frac{1}{\Gamma_t} \right) + \sigma^2 \sum_{t=k_0}^{k} \frac{4\Gamma_k}{\Gamma_t \mu (t+2)^2} \\
&= \frac{k_0(k_0+1)}{(k+1)(k+2)} \left(\Gamma_{k_0-1} \frac{2\sigma^2}{\sqrt{\mu L}} + (1 - \Gamma_{k_0-1}) \frac{\sigma^2}{\sqrt{\mu L}} \right) + \sigma^2 \sum_{t=k_0}^{k} \frac{4\Gamma_k}{\Gamma_t \mu (t+2)^2} \\
&\leq \frac{k_0(k_0+1)}{(k+1)(k+2)} \frac{2\sigma^2}{\sqrt{\mu L}} + \sigma^2 \frac{1}{(k+1)(k+2)} \left(\sum_{t=k_0+1}^{k} \frac{4(t+1)(t+2)}{\mu (t+2)^2} \right) \\
&\leq \frac{k_0}{(k+1)(k+2)} \frac{4\sigma^2}{\mu} + \frac{4\sigma^2}{\mu(k+2)} \leq \frac{8\sigma^2}{\mu(k+2)},
\end{aligned}$$

where we use Lemmas 2.8 and 2.9. This leads to the desired iteration complexity. □

2.D.10 Proof of Corollary 2.10

Proof. Let us call x_0' the point obtained by running on step of iteration (A), which according to Theorem 2.1 satisfies, with $\gamma_0 = 1/\eta$,

$$\mathbb{E}\left[F(x_0') - F^* + \frac{1}{2\eta}\|x_0' - x^*\|^2 \right] \leq \frac{1}{2\eta}\left\|x_0 - x^*\right\|^2 + \eta\sigma^2.$$

Then, we note that according to Lemma 2.11, we have

$$\Gamma_k \leq \frac{4}{\left(2 + k\sqrt{\gamma_0 \eta}\right)^2} \leq \frac{4}{\gamma_0 \eta \left(1+k\right)^2},$$

and we apply Theorem 2.7 to obtain the relation

$$\begin{aligned}
\mathbb{E}\left[F(x_K) - F^*\right] &\leq \Gamma_K \mathbb{E}\left[F(x_0') - F^* + \frac{1}{2\eta}\|x_0' - x^*\|^2 \right] + \sigma^2 \eta \Gamma_K \sum_{t=1}^{K} \frac{1}{\Gamma_t} \\
&\leq \Gamma_K \left(\frac{\left\|x_0 - x^*\right\|^2}{2\eta} + \eta\sigma^2 \right) + \sigma^2 \eta K \\
&\leq \frac{2}{(1+K)^2 \eta}\left\|x_0 - x^*\right\|^2 + \sigma^2 \eta (K+1).
\end{aligned}$$

Optimizing with respect to η under the constraint $\eta \leq 1/L$ gives (2.30). □

2.D.11 Proof of Proposition 2.4

Proof.

$$\begin{aligned}
\omega_k^2 &= \mathbb{E}\left\|\frac{1}{q_{i_k}n}\left(\tilde{\nabla} f_{i_k}(y_{k-1}) - \tilde{\nabla} f_{i_k}(\tilde{x}_{k-1})\right) + \tilde{\nabla} f(\tilde{x}_{k-1}) - \nabla f(y_{k-1})\right\|^2 \\
&= \mathbb{E}\left\|\frac{1}{q_{i_k}n}\left(\nabla f_{i_k}(y_{k-1}) + \zeta_k - \zeta_k' - \nabla f_{i_k}(\tilde{x}_{k-1})\right) + \nabla f(\tilde{x}_{k-1}) + \bar{\zeta}_{k-1} - \nabla f(y_{k-1})\right\|^2, \\
&\leq \mathbb{E}\left\|\frac{1}{q_{i_k}n}\left(\nabla f_{i_k}(y_{k-1}) - \nabla f_{i_k}(\tilde{x}_{k-1})\right) + \nabla f(\tilde{x}_{k-1}) + \bar{\zeta}_{k-1} - \nabla f(y_{k-1})\right\|^2 + 2\rho_Q\tilde{\sigma}^2,
\end{aligned}$$

where ζ_k and ζ_k' are perturbations drawn at iteration k, and $\bar{\zeta}_{k-1}$ was drawn last time $\tilde{x}_{k-1}$ was updated. Then, by noticing that for any deterministic quantity Y and random variable X, we have $\mathbb{E}\left[\|X - \mathbb{E}[X] - Y\|^2\right] \leq \mathbb{E}\left[\|X\|^2\right] + \|Y\|^2$, taking expectation with respect to the index $i_k \sim Q$ and conditioning on $\mathcal{F}_{k-1}$, we have

$$\begin{aligned}
\omega_k^2 &\leq \mathbb{E}\left\|\frac{1}{q_{i_k}n}\left(\nabla f_{i_k}(y_{k-1}) - \nabla f_{i_k}(\tilde{x}_{k-1})\right)\right\|^2 + \mathbb{E}\left[\left\|\bar{\zeta}_{k-1}\right\|^2\right] + 2\rho_Q\tilde{\sigma}^2 \\
&\leq \frac{1}{n}\sum_{i=1}^n \frac{1}{q_i n}\mathbb{E}\left\|\nabla f_i(y_{k-1}) - \nabla f_i(\tilde{x}_{k-1})\right\|^2 + 3\rho_Q\tilde{\sigma}^2 \\
&\leq \frac{1}{n}\sum_{i=1}^n \frac{2L_i}{q_i n}\mathbb{E}\left[f_i(\tilde{x}_{k-1}) - f_i(y_{k-1}) - \nabla f_i(y_{k-1})^\top(\tilde{x}_{k-1} - y_{k-1})\right] + 3\rho_Q\tilde{\sigma}^2 \\
&\leq \frac{1}{n}\sum_{i=1}^n 2L_Q\mathbb{E}\left[f_i(\tilde{x}_{k-1}) - f_i(y_{k-1}) - \nabla f_i(y_{k-1})^\top(\tilde{x}_{k-1} - y_{k-1})\right] + 3\rho_Q\tilde{\sigma}^2 \\
&= 2L_Q\mathbb{E}\left[f(\tilde{x}_{k-1}) - f(y_{k-1}) - \nabla f(y_{k-1})^\top(\tilde{x}_{k-1} - y_{k-1})\right] + 3\rho_Q\tilde{\sigma}^2 \\
&= 2L_Q\mathbb{E}\left[f(\tilde{x}_{k-1}) - f(y_{k-1}) - g_k^\top(\tilde{x}_{k-1} - y_{k-1})\right] + 3\rho_Q\tilde{\sigma}^2, \qquad (2.50)
\end{aligned}$$

where the second inequality uses the upper-bound $\mathbb{E}\left[\|\bar{\zeta}\|^2\right] = \frac{\sigma^2}{n} \leq \rho_Q\sigma^2$, and the third one uses Theorem 2.1.5 in [Nesterov, 2014]. □

2.D.12 Proof of Lemma 2.3

Proof. We can show that Lemma 2.2 still holds and thus,

$$\begin{aligned}
\mathbb{E}\left[F(x_k)\right] &\leq \mathbb{E}\left[l_k(y_{k-1})\right] + \left(\frac{L\eta_k^2}{2} - \eta_k\right)\mathbb{E}\left[\|\tilde{g}_k\|^2\right] + \eta_k\omega_k^2. \\
&\leq \mathbb{E}\left[l_k(y_{k-1}) + a_k f(\tilde{x}_{k-1}) - a_k f(y_{k-1}) + a_k g_k^\top(y_{k-1} - \tilde{x}_{k-1})\right] \\
&\qquad + \mathbb{E}\left[\left(\frac{L\eta_k^2}{2} - \eta_k\right)\|\tilde{g}_k\|^2\right] + 3\rho_Q\eta_k\tilde{\sigma}^2,
\end{aligned}$$

Note also that

$$\begin{aligned} l_k(y_{k-1}) + f(\tilde{x}_{k-1}) - f(y_{k-1}) &= \psi(x_k) + \psi'(x_k)^\top (y_{k-1} - x_k) + f(\tilde{x}_{k-1}) \\ &\leq \psi(\tilde{x}_{k-1}) - \psi'(x_k)^\top (\tilde{x}_{k-1} - x_k) + \psi'(x_k)^\top (y_{k-1} - x_k) + f(\tilde{x}_{k-1}) \\ &= F(\tilde{x}_{k-1}) + \psi'(x_k)^\top (y_{k-1} - \tilde{x}_{k-1}). \end{aligned}$$

Therefore, by noting that $l_k(y_{k-1}) + a_k f(\tilde{x}_{k-1}) - a_k f(y_{k-1}) \leq (1-a_k) l_k(y_{k-1}) + a_k F(\tilde{x}_{k-1}) + a_k \psi'(x_k)^\top (y_{k-1} - \tilde{x}_{k-1})$, we obtain the desired result. □

2.D.13 Proof of Corollary 2.13

Proof. The proof is similar to that of Corollary 2.9 for accelerated SGD. The first stage with constant step size η requires $\mathcal{O}\left(\left(n + \sqrt{\frac{nL_Q}{\mu}}\right) \log\left(\frac{F(x_0)-F^*}{\varepsilon}\right)\right)$ iterations. Then, we restart the optimization procedure, and assume that $\mathbb{E}\left[F(x_0) - F^*\right] \leq B$ with $B = 3\rho_Q \tilde{\sigma}^2 \sqrt{\eta/\mu n}$.

With the choice of parameters, we have $\gamma_k = \mu$ and $\delta_k = \sqrt{\frac{5\mu\eta_k}{3n}} = \min\left(\sqrt{\frac{5\mu\eta}{3n}}, \frac{2}{k+2}\right)$. We may then apply Theorem 2.11 where the value of Γ_k is given by Lemma 2.8. This yields for $k \geq k_0 = \left\lceil \sqrt{\frac{12n}{5\mu\eta}} - 2 \right\rceil$,

$$\begin{aligned} \mathbb{E}\left[F(x_k) - F^*\right] &\leq \Gamma_k \left(\mathbb{E}\left[F(x_0) - F^* + \frac{\mu}{2}\left\|x_0 - x^*\right\|^2\right] + \frac{3\rho_Q\tilde{\sigma}^2}{n} \sum_{t=1}^{k} \frac{\eta_t}{\Gamma_t} \right) \\ &\leq \Gamma_k \left(2B + \frac{3\rho_Q\tilde{\sigma}^2\eta}{n} \sum_{t=1}^{k_0-1} \frac{1}{\Gamma_t} + \frac{3\rho_Q\tilde{\sigma}^2}{n} \sum_{t=k_0}^{k} \frac{12n}{5\Gamma_t \mu (t+2)^2} \right) \\ &= \frac{k_0(k_0+1)}{(k+1)(k+2)} \left(\Gamma_{k_0-1} 2B + \frac{3\rho_Q\tilde{\sigma}^2\eta}{n} \Gamma_{k_0-1} \sum_{t=1}^{k_0-1} \frac{1}{\Gamma_t} \right) + \frac{36\rho_Q\tilde{\sigma}^2}{5\mu} \sum_{t=k_0}^{k} \frac{\Gamma_k}{\Gamma_t (t+2)^2} \\ &= \frac{k_0(k_0+1)}{(k+1)(k+2)} \left(\Gamma_{k_0-1} 2B + (1-\Gamma_{k_0-1}) \frac{3\rho_Q\tilde{\sigma}^2\eta}{n\delta_{k_0}} \right) + \frac{36\rho_Q\tilde{\sigma}^2}{5\mu} \sum_{t=k_0}^{k} \frac{\Gamma_k}{\Gamma_t (t+2)^2} \\ &\leq \frac{2k_0(k_0+1)B}{(k+1)(k+2)} + \frac{8\rho_Q\tilde{\sigma}^2}{\mu(k+1)(k+2)} \left(\sum_{t=k_0+1}^{k} \frac{(t+1)(t+2)}{(t+2)^2} \right) \\ &\leq \frac{2k_0 B}{k+2} + \frac{8\rho_Q\tilde{\sigma}^2}{\mu(k+2)}, \end{aligned}$$

where we use Lemmas 2.8 and 2.9. Then, note that $k_0 B \leq 6\rho_Q\tilde{\sigma}^2/\mu$ and we obtain the right iteration complexity. □

2.D.14 Proof of Corollary 2.14

Proof. Let us call x_0' the point obtained by running one iteration of (A) with step size $\eta \leq \frac{1}{3L_Q}$ and gradient estimator $(1/n)\sum_{i=1}^n \tilde{\nabla} f_i(x_0)$, whose variance is $\tilde{\sigma}^2/n$. Following the proof of Corollary 2.6, the relation (2.49) holds. Then, we consider the main run of the

algorithm, and apply Theorem 2.11, replacing x_0 by x_0', which yields, combined with (2.49)

$$\begin{aligned}\mathbb{E}\left[F(x_k) - F^*\right] &\leq \Gamma_k \left(F(x_0') - F^* + \frac{1}{2\eta}\|x_0' - x^*\|^2 + \frac{3\rho_Q \tilde{\sigma}^2}{n} \sum_{t=1}^{k} \frac{\eta_t}{\Gamma_t} \right) \\ &\leq \Gamma_k \left(\frac{1}{2\eta}\|x_0' - x^*\|^2 + \eta \frac{\tilde{\sigma}^2}{n} + \frac{3\rho_Q \tilde{\sigma}^2}{n} \sum_{t=1}^{k} \frac{\eta_t}{\Gamma_t} \right).\end{aligned}$$

Then, we note that $\delta_k = \min\left(\sqrt{\frac{5\Gamma_k}{3n}}, \frac{1}{3n}\right)$ such that $\Gamma_k = \left(1 - \frac{1}{3n}\right)^k$ for $k \leq k_0$, where k_0 is the index such that $\left(1 - \frac{1}{3n}\right)^{k_0+1} \leq \frac{1}{15n} < \left(1 - \frac{1}{3n}\right)^{k_0}$, which gives us $(3n-1)\log(15n) \leq k_0 \leq 3n(\log(15n))$. For $k > k_0$, we are in a constant step size regime, and we may then use Lemma 2.11 to obtain

$$\Gamma_k = \Gamma_{k_0} \frac{4}{\left(2 + (k-k_0)\sqrt{\frac{5\gamma_{k_0}\eta}{3n}}\right)^2} \leq \Gamma_{k_0} \frac{4}{(k-k_0)^2 \frac{5\Gamma_{k_0}}{3n}} \leq \frac{3n}{(k-k_0)^2}.$$

Then, noticing that $K \geq 2k_0 + 1$, we have $K - k_0 \geq (K+1)/2$, and we conclude that

$$\mathbb{E}\left[F(x_K) - F^*\right] \leq \frac{3n\left\|x_0' - x^*\right\|^2}{2\eta(K-k_0)^2} + \frac{3\eta\rho_Q\tilde{\sigma}^2(K+1)}{n} \leq \frac{6n\|x_0' - x^*\|^2}{\eta(K+1)^2} + \frac{3\eta\rho_Q\tilde{\sigma}^2(K+1)}{n}.$$

Then, it remains to optimize with respect to η, under the constraint $\eta \leq 1/(3L_Q)$, which provides (2.32). □

Chapter 3

A Generic Acceleration for Stochastic Optimization

The convergence rate of a method is the central property of interest in many applications. Consequently, arises a natural need for *acceleration* of optimization methods. In this chapter, we consider acceleration in a *generic* fashion, meaning that one acceleration framework is simultaneously applicable to many methods with various characteristics. In contrast to direct acceleration approaches, which change existing methods from inside or just recreate them from scratch, the task of generic acceleration is to use an optimization method as a building block inside a meta-procedure that does not change the method from inside. We refer to such building blocks as *base* methods. The theoretical convergence bounds of such meta-procedures do not depend on the inner structure of base methods. Yet, acceleration meta-procedures may require these base methods to possess specific convergence properties in order to be able to accelerate them. For example, the Catalyst approach [Lin et al., 2015], developed for optimization of deterministic L-smooth, convex (or strongly convex) composite objectives, requires sub-optimal linear convergence of base methods in order to accelerate them. At the output, it produces an accelerated meta-algorithm with a convergence rate, which is essentially optimal up to factors which are logarithmic in the condition number. For the sake of brevity, we refer to such convergence as *near-optimal* in what follows.

In this chapter, we introduce various mechanisms of such generic acceleration that operate on first-order algorithms and generalize them to *stochastic* optimization problems. Specifically, we extend the Catalyst approach of deterministic optimization to the stochastic setting with L-smooth, convex (or strongly convex) composite objectives. The principal task here is to derive stochastic meta-algorithms with the optimal (up to a factor logarithmic in the condition number) convergence, meaning that we preserve the deterministic acceleration of the bias part (performed by the original Catalyst approach) and simultaneously achieve robustness to noise which comes from inexact oracle. As the base methods are essentially stochastic in our setting, the requirement of linear convergence imposed by Catalyst

has to be refined. In particular, we preserve the assumption of linear bias convergence of a base method, and additionally assume that its variance does not diverge. Under this rather mild requirement, we show that our multi-stage procedure results in meta-algorithms of near-optimal convergence rates. As the Catalyst approach is also applicable to incremental variance-reduced algorithms from Chapter 2 when $\sigma = 0$, we address their generic acceleration for $\sigma > 0$ as well.

This chapter is based on the following publication:

- A. Kulunchakov and J. Mairal. A generic acceleration framework for stochastic composite optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019

3.1 Introduction

In this chapter, we consider stochastic composite optimization problems of the form

$$\min_{x\in\mathbb{R}^p}\Big\{F(x) \triangleq f(x) + \psi(x)\Big\} \quad \text{with} \quad f(x) = \mathbb{E}_\rho[\tilde{f}(x,\rho)], \tag{3.1}$$

where f is convex and L-smooth, and we call μ its strong convexity modulus with respect to the Euclidean norm (if such μ is positive). The function ψ is convex lower semi-continuous and is not assumed to be necessarily differentiable, with possible practical examples of this penalty function given in Section 1.2.2. In addition, we separately consider the optimization setting of Chapter 2, being

$$\min_{x\in\mathbb{R}^p}\left\{F(x) \triangleq f(x) + \psi(x) \triangleq \frac{1}{n}\sum_{i=1}^n f_i(x) + \psi(x)\right\}, \tag{3.2}$$

where the terms f_i are L-smooth and convex. Moreover, we assume $(f_i)_{i=1}^n$ to be given in the form of expectation $f_i = \mathbb{E}_{\rho_i}\left[\tilde{f}_i(x,\rho_i)\right]$, so that one has access only to inexact gradient estimations for each f_i. Principally, in this chapter, we focus on solving optimization problems of the form (3.1). The statement (3.2) will be analyzed only in Section 3.3.2 where we consider acceleration of stochastic incremental approaches from Chapter 2, such as stochastic SVRG/SAGA/SDCA/Finito/MISO. The merits of consideration of problems (3.2) were given in Section 2.1, so that we focus further on the overview of methods solving (3.2).

We have mentioned earlier that the deterministic nature of $F(x)$ could drastically change performance guarantees of methods applied to minimize it. Yet, as noted in [Bottou and Bousquet, 2008], one is typically not interested in such minimization with high precision, but instead, one should focus on the expected risk (3.1) involving the true (unknown) data distribution. When one can draw an infinite number of samples from this distribution, this true risk may be minimized by using appropriate stochastic optimization techniques. Yet, standard non-accelerated methods of SA type admit optimal "slow" rates that are typically $\mathcal{O}\left(1/\sqrt{N}\right)$ for convex functions and $\mathcal{O}\left(1/N\right)$ for strongly convex ones without a simultaneous establishment of fast linear convergence of the initial

error. At the same time, as mentioned in Section 1.5, accelerated deterministic methods do not straightforwardly apply to the minimization of the true risk. That is why better understanding the gap between deterministic and stochastic optimization in terms of acceleration is one goal of this chapter.

Specifically, we are interested in Nesterov's acceleration of gradient-based approaches [Nesterov, 1983, 2014] Whereas no clear geometrical intuition seems to appear in the literature to explain why such acceleration occurs, there are now well established proof techniques to show accelerated convergence [Tseng, 2008, Beck and Teboulle, 2009, Nesterov, 2014] and extensions to a large class of other gradient-based algorithms [Nesterov, 2012, Chambolle and Pock, 2015, Shalev-Shwartz and Zhang, 2016, Allen-Zhu, 2017, Lin et al., 2018]. Yet, the effect of Nesterov's acceleration to stochastic objectives remains poorly understood since existing unaccelerated algorithms already achieve the optimal *asymptotic* rate. Nevertheless, several approaches such as [Hu et al., 2009, Xiao, 2010, Devolder et al., 2011, Ghadimi and Lan, 2012, 2013, Cohen et al., 2018, Aybat et al., 2019] have managed to show that acceleration may be useful in order to reach faster a region dominated by the noise of stochastic gradients. Then, "good" methods are expected to asymptotically converge with a rate exhibiting an optimal dependency in the noise variance, but with no dependency on the initialization. Therefore, there is a demand for accelerated stochastic methods that perform in both these regimes optimally. In this chapter, we address this task by developing a multi-stage procedure that takes non-accelerated methods—referred to as *base methods*—as input and outputs meta-algorithms with a near-optimal complexity.

Throughout the chapter, we denote each base method as $\mathcal{M}$ and assume it to satisfy the following property. Given an auxiliary strongly convex objective function h, a base method $\mathcal{M}$ is able to produce iterates $(z_t)_{t\geq 0}$ with expected linear convergence to a noise-dominated region—that is, such that

$$\mathbb{E}\left[h(z_t) - h^*\right] \leq C\left(1-\tau\right)^t \left(h(z_0) - h^*\right) + B\sigma^2, \tag{3.3}$$

where C, τ, B are positive, h^* is the minimum function value, and σ^2 is an upper bound on the variance of stochastic gradients accessed by $\mathcal{M}$, which we assume to be uniformly bounded, see Definition 1.10. Whereas this assumption has limitations, it remains the most standard one for stochastic optimization (see [Bottou et al., 2018, Nguyen et al., 2018] for more realistic settings in the smooth case) with the class of methods satisfying (3.3) being relatively large. For instance, when h is L-smooth, the stochastic gradient descent method (SGD) with constant step size $1/L$ and iterate averaging satisfies (3.3) with $\tau = \mu/L$, $B = 1/L$, and $C = 1$, see Corollary 2.15.

In order to build an acceleration framework for stochastic methods, we extend the Catalyst approach [Lin et al., 2018] originally developed for optimization problems in the deterministic setting. Let us briefly give a flavor of its scheme. In a nutshell, Catalyst is based on the inexact accelerated proximal point algorithm [Güler, 1992], which consists in solving approximately a sequence of sub-problems and updating two sequences $(x_k)_{k\geq 0}$ and $(y_k)_{k\geq 0}$ by

$$x_k \approx \operatorname*{argmin}_{x\in\mathbb{R}^p} \left\{ h_k(x) \triangleq F(x) + \frac{\kappa}{2}\left\|x - y_{k-1}\right\|^2 \right\} \quad \text{and} \quad y_k = x_k + \beta_k(x_k - x_{k-1}), \tag{3.4}$$

where β_k in $(0, 1)$ is obtained from Nesterov's acceleration principles [Nesterov, 2014], and $\kappa > 0$ is a well chosen regularization parameter. The base method $\mathcal{M}$ to be accelerated is used to obtain an approximate minimizer of h_k by using an appropriate computational budget. When $\mathcal{M}$ converges linearly (like in (3.3) when $\sigma = 0$), it may be shown that in the deterministic setting $\sigma = 0$ the resulting algorithm (3.4) enjoys a better worst-case complexity than if $\mathcal{M}$ was used directly on f. Since asymptotic linear convergence is out of reach when f is stochastic, a classical strategy consists in replacing $F(x)$ in (3.4) by its finite-sum approximation. Typically without Nesterov's acceleration (with $y_k = x_k$), this strategy is often called the stochastic proximal point algorithm [Asi and Duchi, 2019, Bertsekas, 2011, Kulis and Bartlett, 2010, Toulis et al., 2018, 2016].

The point of view we adopt in this chapter is different from stochastic PPA and is based on the minimization of surrogate functions h_k related to (3.4), which are more general and may take forms different from $F(x) + (\kappa/2)\left\|x - y_{k-1}\right\|^2$. Similarly to the original Catalyst method, given a base method $\mathcal{M}$ that satisfies the condition (3.3), our procedure is able to turn $\mathcal{M}$ into a converging algorithm $\mathcal{M}'$ with a near-optimal worst-case iteration complexity, being specifically

$$\mathbb{E}\left[F(x_N) - F^*\right] \leq \widetilde{\mathcal{O}}\left(\left(1 - \sqrt{\frac{\mu}{L}}\right)^N \left(F(x_0) - F^*\right)\right) + \widetilde{\mathcal{O}}\left(\frac{\sigma^2}{\mu N}\right), \quad \text{when } \mu > 0,$$

where N is the budget of iterations and $\widetilde{\mathcal{O}}\left(\cdot\right)$ may hide some logarithmic factors in the condition number L/μ. These rates are achieved essentially due to improved analysis and flexibility with respect to the choice of surrogate functions $h_k(x)$. The principal scheme of the procedure looks similar to the one of the original Catalyst with the only difference being the update for the extrapolated point y_k, which becomes

$$y_k = x_k^* + \beta_k(x_k^* - x_{k-1}) + \frac{(\mu + \kappa)(1 - \alpha_k)}{\kappa}(x_k - x_k^*) \tag{3.5}$$

where $\beta_k > 0$ and x_k^* is the minimizer of a properly chosen surrogate function h_k. When x_k^* is not available, the scheme is just the same as in the original Catalyst approach.

To illustrate the versatility of our approach, we consider the stochastic finite-sum problem (3.2) which was reviewed in details in Sections 1.4.4 and 2.1. Whereas it was shown in Chapter 2 that many variance-reduced algorithms, such as SVRG, SAGA, SDCA, Finito or MISO, can be made robust to noise, the analysis conducted there is only able to accelerate the SVRG approach. With our framework, all of the aforementioned incremental methods can be accelerated to a near-optimal convergence rate. As well as for the Catalyst approach, the price to pay compared to direct acceleration techniques of [Allen-Zhu, 2017, Lan and Zhou, 2018a] and Chapter 2 is a logarithmic factor.

In Table 3.1 we present the complete list of transformations of complexities performed by our framework. Note that even though in (3.3) we make assumptions about the behavior of $\mathcal{M}$ when applied to strongly convex sub-problems, we also effectively treat the setting where the objective is convex, but not strongly convex. In this case, base methods still need to be defined on strongly convex problems and enjoy the convergence (3.3), even though they are applied to a non-strongly convex objective. For the cases when a base method is not defined on strongly convex problems, we use a separate restart

Table 3.1 – Complexity transformations performed by our acceleration framework. The dash fields mean that the base method is an algorithm of one specific step. Here Δ_0 is an upper bound on the initial primal gap $F(x_0) - F^*$; Ω^2 is an upper bound on the initial approximation error $\|x_0 - x^*\|_2^2$; and ε is a targeted accuracy. All constants are omitted for a simplicity and generality.

Value of μ	Original complexity of $\mathcal{M}$	Resulting complexity
Deterministic case with $\sigma = 0$ and $n \not\gg 1$		
$\mu > 0$	$(L/\mu)\log(\Delta_0/\varepsilon)$	$\sqrt{L/\mu}\log(\Delta_0/\varepsilon)$
$\mu = 0$	$L\Omega^2/\varepsilon$	$\sqrt{L\Omega^2/\varepsilon}$
Deterministic case with $\sigma = 0$ and $n \gg 1$		
$\mu > 0$	$(n + L/\mu)\log(\Delta_0/\varepsilon)$	$\left(n + \sqrt{nL/\mu}\right)\log(\Delta_0/\varepsilon)$
Stochastic case with $\sigma > 0$ and $n \not\gg 1$		
$\mu > 0$	—	$\sqrt{L/\mu}\log(\Delta_0/\varepsilon) + \sigma^2/\mu\varepsilon$
$\mu = 0$	—	$\sqrt{L\Omega^2/\varepsilon} + (\sigma\Omega/\varepsilon)^2$
Stochastic case with $\sigma > 0$ and $n \gg 1$		
$\mu > 0$	$(n + L/\mu)\log(1/\varepsilon) + \sigma^2/\mu\varepsilon$ or even $(n + L/\mu)\log(1/\varepsilon)$ biased as $\varepsilon \geq \sigma^2/\mu$	$\left(n + \sqrt{nL/\mu}\right)\log(1/\varepsilon) + \sigma^2/\mu\varepsilon$

procedure, called *domain shrinking*, which was developed in [Ghadimi and Lan, 2013, Iouditski and Nesterov, 2014]. Specifically, this procedure is aiming to convert convex methods into strongly convex ones, thus changing sublinear bias convergence into the linear one. Many resulting complexities also require the use of a special restart procedure with exponentially increasing batches, which will be defined in the next section. The empty cells in Table 3.1 represent cases when the base methods consist of one step, so that the overall meta-procedure may be seen just as a stand-alone algorithm.

Finally, let us now briefly overview the explicit list of the contributions presented in the chapter.

3.1.1 Contributions of Chapter 3

- We extend the Catalyst approach [Lin et al., 2018] to minimization of stochastic objective functions. Under mild condition (3.3), our approach is able to turn a base method into a converging algorithm with one of the worst-case convergence rates (1.11) or (1.12), depending on the value of μ.
- Beyond the ability to deal with stochastic optimization problems, our procedure improves the Catalyst approach by allowing deterministic sub-problems of the form (3.3) with $\sigma = 0$ to be solved approximately *in expectation*, which is more realistic than the deterministic requirement made in [Lin et al., 2018] and which is also critical for stochastic optimization algorithms.

- The original Catalyst approach is able to accelerate various incremental variance-reduced methods applied to (3.2) in the deterministic case $\sigma = 0$. We generalize this acceleration to the stochastic setting using the generalized versions of these methods described in Chapter 2.
- A side contribution of this chapter is also a generic analysis that can handle *inexact* proximal operators, providing new insights about the robustness of stochastic algorithms when the proximal operator cannot be exactly computed. To the best of our knowledge, the stochastic setting with approximately computed proximal operators has never been analyzed in the literature.

The rest of the chapter is organized as follows. Section 3.2 overviews two basic multi-stage schemes, namely the domain shrinking and mini-batch restart procedures, which both improve the worst-case complexities of different stochastic optimization methods. In Section 3.3, we introduce the proposed acceleration framework along with its theoretical guarantees and examples of acceleration. Section 3.3.2 gives a slightly finer analysis to this framework, which allows to accelerate stochastic incremental variance-reduced algorithms. In Section 2.3, we present various experiments demonstrating the effectiveness of our procedure.

3.2 Preliminaries: Basic Multi-Stage Schemes

In this section, we overview two simple multi-stage procedures aimed to improve the worst-case iteration complexities of different stochastic optimization methods. Both of them are used by our framework in some cases in order to achieve near-optimal convergence rates for the accelerated methods.

Basic restart with mini-batching or decaying step sizes. Consider a base optimization method $\mathcal{M}$ with a convergence rate of the form (3.3) and assume that there exists a hyper-parameter to control a trade-off between the constant term $B\sigma^2$ and per-iteration computational complexity of $\mathcal{M}$. Specifically, we assume that the constant term can be reduced by an arbitrary factor $\eta < 1$, while paying the factor $1/\eta$ in terms of complexity per iteration (for instance, τ may become $\eta\tau < \tau$ thus slowing down the convergence). For example, this mechanism may be available in two cases:

— by using a mini-batch of size $1/\eta$ to sample gradients, which replaces σ^2 by $\eta\sigma^2$;
— if the method uses a step size proportional to η that can be chosen arbitrarily small.

Then, consider a target accuracy ε and define the sequences $\eta_k = 2^{-k}$ and $\varepsilon_k = 2B\eta_k\sigma^2$ for $k \geq 0$. We may now successively solve the problem of interest up to accuracy ε_k—*e.g.*, with a constant number $\mathcal{O}(1/\tau)$ of steps of $\mathcal{M}$ when using mini-batches of size $1/\eta_k = 2^k$—and by using the solution of iteration $k-1$ as a warm restart. This approach is expressed explicitly in Algorithm 3.1.

As shown in Appendix 3.B, this algorithm converges and the worst-case complexity to achieve the accuracy ε in expectation is

$$\mathcal{O}\left(\frac{1}{\tau}\log\left(\frac{C\left(F(x_0) - F^*\right)}{\varepsilon}\right) + \frac{B\sigma^2\log(2C)}{\tau\varepsilon}\right). \tag{3.6}$$

Algorithm 3.1 Mini-batch restart procedure

Input: objective F, optimization method $\mathcal{M}$ with convergence rate (3.3); an initial estimate x_0.
for $k = 1 \dots K$ **do**
— Set up an oracle with mini-batches of size $1/\eta_k = 2^k$;
— Launch the method $\mathcal{M}$ started from the last disposed solution x_{k-1} for $\mathcal{O}(1/\tau)$ iterations such that $\mathcal{M}$ reaches an (expected) accuracy $\mathbb{E}[F(x_k) - F^*] \leq 2B\sigma^2/2^k$.
end for
Output: x_K

This result opens doors for the following strategy. Let us consider the SGD algorithm with a constant step size η from Corollary 2.15 of Chapter 2, which has the following convergence rate

$$\mathbb{E}[F(\hat{x}_k) - F^*] \leq 2\left(1 - \frac{\mu}{L}\right)^k (F(\hat{x}_0) - F^*) + \frac{\sigma^2}{L},$$

where $\hat{x}_0 = x_0$ and $\hat{x}_k$ is the estimate recursively obtained as $\hat{x}_k = (1 - \delta_k)\hat{x}_{k-1} + \delta_k x_k$ (we have neglected all absolute factors for simplicity). This expression is of the form (3.3) with $B = 1/L$, $C = 2$, and $\tau = \mu/L$. Then, when used as a base method in Algorithm 3.1, it obtains the convergence guarantee (3.6) with the left term being the classical complexity $\mathcal{O}((L/\mu)\log(1/\varepsilon))$ of the (unaccelerated) gradient descent algorithm for deterministic objectives, whereas the right term is the optimal complexity for stochastic optimization being $\mathcal{O}(\sigma^2/\mu\varepsilon)$. Similar restart principles appear for instance in [Aybat et al., 2019] in the design of a multi-stage accelerated SGD algorithm.

Domain shrinking: from sub-linear to linear rate with strong convexity. A natural question is whether asking for a linear rate in (3.3) for strongly convex problems is a strong requirement. Here, we show that a sublinear rate is in fact sufficient for our needs by using the technique of *domain shrinking* developed in [Ghadimi and Lan, 2013, Iouditski and Nesterov, 2014]. Specifically, consider an algorithm $\mathcal{M}$ for stochastic optimization with the following convergence rate

$$\mathbb{E}[h(z_t) - h^*] \leq \frac{D\|z_0 - z^*\|^2}{2t^d} + \frac{B\sigma^2}{2}, \tag{3.7}$$

where $D, d > 0$ and z^* is a minimizer of h. Assume now that h is μ-strongly convex with $D \geq \mu$ and consider restarting s times the method $\mathcal{M}$, each time running it for $t' = \left\lceil (2D/\mu)^{1/d} \right\rceil$ iterations. This scheme is expressed explicitly in Algorithm 3.2.

It is shown in Appendix 3.B that the resulted algorithm enjoys the following convergence rate

$$\mathbb{E}[h(z_t) - h^*] \leq (h(z_0) - h^*)\left(1 - \frac{1}{2}\left(\frac{\mu}{2D}\right)^{1/d}\right)^t + B\sigma^2,$$

which is essentially the desired relation (3.3) with $C = 1$ and $\tau = 1/(2t') = \frac{1}{2}(\mu/2D)^{1/d}$. Therefore, if a mini-batch or step size mechanism is available, we may then proceed with Algorithm 3.1 and finally obtain a converging scheme with the complexity (3.6).

Algorithm 3.2 Restart with domain shrinking

Input: optimization method $\mathcal{M}$; an initial point x_0 in $\mathbb{R}^p$; target accuracy ε; parameters μ and D, B, σ from (3.7).
for $k = 1 \dots K = \lceil \log_2 (2B\sigma^2/\varepsilon) \rceil$ **do**
produce x_k by running $\mathcal{M}$ with

$$t' = \left\lceil (2D/\mu)^{1/d} \right\rceil \tag{3.8}$$

iterations, being initialized at x_{k-1}.
end for
Output: x_K

Table 3.1 – Complexity transformations performed by the domain shrinking restart procedure from Algorithm 3.2, when a base method $\mathcal{M}$ is defined only for non-strongly convex problems. Here Ω^2 is an upper bound on the initial approximation error $\|z_0 - z^*\|_2^2$; and ε is a targeted accuracy. All constants are omitted for a simplicity and generality.

B	Original complexity of $\mathcal{M}$	Resulting complexity	Use Algorithm 3.1
Procedure from Algorithm 3.2			
$1/L$	$L\Omega^2/\varepsilon$ with $\varepsilon \geq \sigma^2/2L$	$(L/\mu)\log(\Delta_0/\varepsilon)$ with $\varepsilon \geq \sigma^2/L$	no
$1/L$	$L\Omega^2/\varepsilon$ with $\varepsilon \geq \sigma^2/2L$	$(L/\mu)\log(\Delta_0/\varepsilon) + \sigma^2/\mu\varepsilon$	yes
$1/\sqrt{\mu L}$	$\sqrt{L\Omega^2/\varepsilon}$ with $\varepsilon \geq \sigma^2/2\sqrt{\mu L}$	$\sqrt{L/\mu}\log(\Delta_0/\varepsilon)$ with $\varepsilon \geq \sigma^2/\sqrt{\mu L}$	no
$1/\sqrt{\mu L}$	$\sqrt{L\Omega^2/\varepsilon}$ with $\varepsilon \geq \sigma^2/2\sqrt{\mu L}$	$\sqrt{L/\mu}\log(\Delta_0/\varepsilon) + \sigma^2/\mu\varepsilon$	yes
Modified procedure from Algorithm 3.2 according to [Ghadimi and Lan, 2013]			
—	$L\Omega^2/\varepsilon + (\Omega\sigma/\varepsilon)^2$	$(L/\mu)\log(\Delta_0/\varepsilon) + \sigma^2/\mu\varepsilon$	no
—	$\sqrt{L\Omega^2/\varepsilon} + (\Omega\sigma/\varepsilon)^2$	$\sqrt{L/\mu}\log(\Delta_0/\varepsilon) + \sigma^2/\mu\varepsilon$	no

According to [Ghadimi and Lan, 2013], it can also be shown that a slightly modified Algorithm 3.2 is also applicable to methods with convergence rates of the form

$$\mathbb{E}\left[h(z_t) - h^*\right] \leq \frac{D\left\|z_0 - z^*\right\|^2}{2t^d} + \frac{\left\|z_0 - z^*\right\|\sigma}{\sqrt{t}} \quad \text{with} \quad d = 1 \text{ or } 2, \tag{3.9}$$

having thus a typical rate for convex optimization problems. For this purpose, we need only to change the expression for the number of iterations (3.8) which becomes slightly more complicated, so that we do not present it here and refer the reader to Equation (3.7) of [Ghadimi and Lan, 2013]. Then, when a base method with convergence of the type (3.9) is wrapped by the modified Algorithm 3.2, it will eventually converge with the following rate

$$\mathbb{E}\left[h(z_t) - h^*\right] \leq \left(h(z_0) - h^*\right)\left(1 - \left(\frac{\mu}{D}\right)^{1/d}\right)^t + \frac{\sigma^2}{\mu t},$$

with absolute constants omitted for simplicity.

For convenience, we present all complexity transformations performed by Algorithm 3.2 in Table 3.1. The bottom row demonstrates the transformation which is essentially the one made in [Ghadimi and Lan, 2013].

3.3 Generic Multi-Stage Approaches with Acceleration

We are now in shape to present the main contribution of this chapter, namely a generic acceleration framework that generalizes the Catalyst approach. First, we introduce the concept of *surrogate functions* which is the key for this procedure.

Assumption 3.1 (Surrogate function)**.** *Assume that we have a base optimization method $\mathcal{M}$ with convergence (3.3). Consider the k-th stage of some multi-stage procedure that forms two two sequences $(x_k)_{k\geq 0}$ and $(y_k)_{k\geq 0}$. Given some parameters $\kappa > 0$ and $\delta_k \geq 0$, we consider a surrogate function h_k that satisfies the following properties:*

($\mathcal{H}_1$) h_k is $(\mu+\kappa)$-strongly convex, where μ is the strong convexity parameter of f;

($\mathcal{H}_2$) $\mathbb{E}\left[h_k(x)|\mathcal{F}_{k-1}\right] \leq F(x) + (\kappa/2)\left\|x - y_{k-1}\right\|^2$ for $x = \alpha_{k-1}x^* + (1-\alpha_{k-1})x_{k-1}$, which is deterministic given the past information $\mathcal{F}_{k-1}$ up to iteration $k-1$ and α_{k-1} is given further in Algorithm 3.3;

($\mathcal{H}_3$) $\mathcal{M}$ can provide the exact minimizer x_k^* of h_k and a point x_k (possibly equal to x_k^*) such that $\mathbb{E}\left[F(x_k)\right] \leq \mathbb{E}\left[h_k^*\right] + \delta_k$ where $h_k^* = \min_x h_k(x) = h_k(x_k^*)$.

The parameter κ from Definition 3.1 has the same meaning as the one from the Catalyst approach, being essentially the constant smoothing parameter for the sub-problems. Note that the conditions imposed on h_k bear similarities with estimate sequences introduced by [Nesterov, 2014]. Indeed, ($\mathcal{H}_3$) is a direct generalization of (2.2.2) from [Nesterov, 2014] and ($\mathcal{H}_2$) resembles (2.2.1). However, the choices of h_k and our proof technique are significantly different, as we will see with various examples below. Note also that ($\mathcal{H}_3$) is rather a definition for the parameter δ_k than a condition imposed on $h_k(x)$. In other words, one may implicitly state that $\delta_k = \mathbb{E}\left[F(x_k)\right] - \mathbb{E}\left[h_k^*\right]$. At the moment, we assume that the exact minimizer x_k^* of h_k is available, which differs from the original Catalyst framework [Lin et al., 2018]. The case with approximate minimization will be presented further in Section 3.3.1.

Now, we are able to present the scheme of our generic acceleration framework in Algorithm 3.3. Note that we do not need to know all of the surrogate functions $(h_k)_{k=1}^K$ in advance. We can construct them one per stage on the fly.

Note that the update rule (3.10) for the extrapolation point becomes $y_k = x_k + \beta_k(x_k - x_{k-1})$ when $x_k^* = x_k$, being the update of the deterministic Catalyst (3.4) approach and the standard Nesterov's extrapolation rule. In this sense, Algorithm 3.3 can be seen as a generalization of Nesterov acceleration, where gradient steps are replaced with full launches of the base input method $\mathcal{M}$. The convergence rate of Algorithm 3.3 is expressed in the following

Algorithm 3.3 Generic Acceleration Framework with Exact Minimization of h_k

Input: x_0 (initial estimate); $\mathcal{M}$ (optimization method); μ (strong convexity constant); κ (parameter for h_k); K (number of iterations); $(\delta_k)_{k\geq 0}$ (approximation errors);
Initialization: $y_0 = x_0$; $q = \frac{\mu}{\mu+\kappa}$; $\alpha_0 = 1$ if $\mu = 0$ or $\alpha_0 = \sqrt{q}$ if $\mu \neq 0$;
for $k = 1, \ldots, K$ **do**
    Consider a surrogate h_k satisfying $(\mathcal{H}_1)$, $(\mathcal{H}_2)$ and obtain x_k, x_k^* using $\mathcal{M}$ satisfying $(\mathcal{H}_3)$;
    Compute α_k in $(0,1)$ by solving the equation $\alpha_k^2 = (1-\alpha_k)\alpha_{k-1}^2 + q\alpha_k$.
    Update the extrapolated sequence

$$y_k = x_k^* + \beta_k(x_k^* - x_{k-1}) + \frac{(\mu+\kappa)(1-\alpha_k)}{\kappa}(x_k - x_k^*) \quad \text{with} \quad \beta_k = \frac{\alpha_{k-1}(1-\alpha_{k-1})}{\alpha_{k-1}^2 + \alpha_k}. \tag{3.10}$$

end for
Output: x_k (final estimate).

Proposition 3.1 (Convergence analysis for Algorithm 3.3)**.** *Consider the optimization problem 3.2. Given a base optimization method with a convergence rate of the type (3.3), the scheme expressed in Algorithm 3.3 has the following convergence rate*

$$\mathbb{E}\left[F(x_k) - F^*\right] \leq \begin{cases} \left(1-\sqrt{q}\right)^k \left(2\left(F(x_0) - F^*\right) + \sum_{j=1}^k \left(1-\sqrt{q}\right)^{-j} \delta_j\right) & \text{if } \mu \neq 0 \\ \frac{2}{(k+1)^2}\left(\kappa \left\|x_0 - x^*\right\|^2 + \sum_{j=1}^k \delta_j (j+1)^2\right) & \text{otherwise} \end{cases}. \tag{3.11}$$

The proof of the proposition is given in Appendix 3.C. This proof is based on an extension of the analysis of Catalyst [Lin et al., 2018]. As we see, these bounds depend on values of $(\delta_j)_{j=1}^k$, which essentially express the "tightness" of the surrogate functions $(h_j(x))_{j=1}^k$ to the true objective function $F(x)$.

Let us now present various application cases leading to algorithms with acceleration.

Accelerated proximal gradient method. Assume that f is deterministic and there is an exact oracle for the true gradient $\nabla f(x)$, and the proximal operator (1.9) of ψ can be computed in closed form. Then, choose $\kappa = L - \mu$ and define

$$h_k(x) \triangleq f(y_{k-1}) + \nabla f(y_{k-1})^\top (x - y_{k-1}) + \frac{L}{2}\left\|x - y_{k-1}\right\|^2 + \psi(x). \tag{3.12}$$

Consider $\mathcal{M}$ that minimizes h_k in closed form: $x_k = x_k^* = \mathrm{Prox}_{\psi/L}\left[y_{k-1} - \frac{1}{L}\nabla f(y_{k-1})\right]$. Then, $(\mathcal{H}_1)$ is obvious; $(\mathcal{H}_2)$ holds from the convexity of f, and $(\mathcal{H}_3)$ with $\delta_k = 0$ follows from classical inequalities for L-smooth functions [Nesterov, 2014]. Finally, we recover accelerated convergence rates [Beck and Teboulle, 2009, Nesterov, 2014]

$$\mathbb{E}\left[F(x_k) - F^*\right] \leq \begin{cases} 2\left(1-\sqrt{\mu/L}\right)^k \left(F(x_0) - F^*\right) & \text{if } \mu \neq 0 \\ 2L\left\|x_0 - x^*\right\|^2/(k+1)^2 & \text{otherwise} \end{cases}. \tag{3.13}$$

Accelerated proximal point algorithm. We consider h_k given in (3.4) with exact minimization that is performed by a base method $\mathcal{M}$ in $N_0 = \mathcal{O}(1)$ iterations. This is an unrealistic setting, but still conceptually interesting. Given $\kappa = L - \mu$, the assumptions $(\mathcal{H}_1)$, $(\mathcal{H}_2)$, and $(\mathcal{H}_3)$ are satisfied with $\delta_k = 0$ and we recover the accelerated rates of [Güler, 1992].

$$\mathbb{E}\left[F(x_k) - F^*\right] \leq \begin{cases} 2\left(1 - \sqrt{\mu/L}\right)^{k/N_0} (F(x_0) - F^*) & \text{if } \mu \neq 0 \\ 2L \left\|x_0 - x^*\right\|^2 / (k/N_0 + 1)^2 & \text{otherwise} \end{cases}. \tag{3.14}$$

Accelerated stochastic gradient descent with prox. A more interesting choice of surrogate is

$$h_k(x) \triangleq f(y_{k-1}) + g_k^\top (x - y_{k-1}) + \frac{\mu + \kappa}{2} \left\|x - y_{k-1}\right\|^2 + \psi(x), \tag{3.15}$$

where $\kappa \geq L - \mu$ and g_k is an unbiased estimate of $\nabla f(y_{k-1})$ that satisfies

$$\mathbb{E}\left[g_k | \mathcal{F}_{k-1}\right] = \nabla f(y_{k-1}) \quad \text{and} \quad \mathbb{E}\left[\left\|g_k - \nabla f(y_{k-1})\right\|^2 | \mathcal{F}_{k-1}\right] \leq \sigma^2$$

following classical assumptions from the stochastic optimization literature [Hu et al., 2009, Ghadimi and Lan, 2012, 2013]. Then, $(\mathcal{H}_1)$ and $(\mathcal{H}_2)$ are satisfied given that f is convex. To characterize $(\mathcal{H}_3)$, consider a base method $\mathcal{M}$ that minimizes h_k in closed form in one step:

$$x_k = x_k^* = \mathrm{Prox}_{\psi/(\mu+\kappa)}\left[y_{k-1} - \frac{1}{\mu + \kappa} g_k\right],$$

and define

$$u_{k-1} \triangleq \mathrm{Prox}_{\psi/(\mu+\kappa)}\left[y_{k-1} - \frac{1}{\mu + \kappa} \nabla f(y_{k-1})\right],$$

which is deterministic given $\mathcal{F}_{k-1}$. Then, from (3.15), we have

$$\begin{aligned} f(x_k) &\leq h_k(x_k) + (\nabla f(y_{k-1}) - g_k)^\top (x_k - y_{k-1}) \qquad \text{(from } L\text{-smoothness of } f) \\ &= h_k^* + (\nabla f(y_{k-1}) - g_k)^\top (x_k - u_{k-1}) + (\nabla f(y_{k-1}) - g_k)^\top (u_{k-1} - y_{k-1}). \end{aligned}$$

When taking expectations, the last term on the right disappears since $\mathbb{E}\left[g_k | \mathcal{F}_{k-1}\right] = \nabla f(y_{k-1})$:

$$\mathbb{E}\left[f(x_k)\right] \leq \mathbb{E}\left[h_k^*\right] + \mathbb{E}\left[\left\|g_k - \nabla f(y_{k-1})\right\| \left\|x_k - u_{k-1}\right\|\right] \tag{3.16}$$

$$\leq \mathbb{E}\left[h_k^*\right] + \frac{1}{\mu + \kappa} \mathbb{E}\left[\left\|g_k - \nabla f(y_{k-1})\right\|^2\right] \leq \mathbb{E}\left[h_k^*\right] + \frac{\sigma^2}{\mu + \kappa}, \tag{3.17}$$

where we used non-expansiveness of the proximal operator [Moreau, 1965]. Therefore, $(\mathcal{H}_3)$ holds with $\delta_k = \sigma^2/(\mu + \kappa)$. The resulting algorithm is similar to Algorithm C from Chapter 2 and offers the same guarantees according to Corollaries 3.1 and 3.2. Compared to Algorithm C, the novelty of our approach is a unified convergence proof for the deterministic and stochastic cases.

Corollary 3.1 (Complexity of Algorithm 3.3 when $\mu > 0$)**.** *Consider the setting of Proposition 3.1 with h_k defined in (3.15). When f is μ-strongly convex, choose $\kappa = L - \mu$. Then, the Algorithm 3.3 has the following convergence rate*

$$\mathbb{E}\left[F(x_k) - F^*\right] \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)^k \left(F(x_0) - F^*\right) + \frac{\sigma^2}{\sqrt{\mu L}},$$

which is of the form (3.3) with $\tau = \sqrt{\mu/L}$ and $B = \sigma^2/\sqrt{\mu L}$. Therefore, the optimal complexity $\mathcal{O}\left(\sqrt{L/\mu}\log((F(x_0) - F^*)/\varepsilon) + \sigma^2/\mu\varepsilon\right)$ can be obtained by using the restart strategy in Algorithm 3.1 either with exponentially increasing mini-batches or decreasing step sizes.

When the objective is convex, but not strongly convex, Proposition 3.1 gives a constant bias term $\mathcal{O}\left(\sigma^2 k/\kappa\right)$ that increases linearly with k. Yet, the following corollary exhibits an optimal rate for the finite horizon setting, when both σ^2 and an upper-bound on $\|x_0 - x^*\|^2$ are available. Even though non-practical, the result shows that our analysis recovers the optimal dependency in the noise level, as in [Ghadimi and Lan, 2013], Chapter 2 and others.

Corollary 3.2 (Complexity of Algorithm 3.3 when $\mu = 0$)**.** *Consider the setting of Proposition 3.1. Assume that we have a fixed budget K of iterations of Algorithm 3.3 with h_k defined in (3.15). When $\kappa = \max(L, \sigma(K+1)^{3/2}/\|x_0 - x^*\|)$, the procedure has the following convergence rate*

$$\mathbb{E}\left[F(x_K) - F^*\right] \leq \frac{2L\|x_0 - x^*\|^2}{(K+1)^2} + \frac{3\sigma\|x_0 - x^*\|}{\sqrt{K+1}}.$$

While all the previous examples use the choice $x_k = x_k^*$, we demonstrate in Section 3.3.2 a situation where we need to choose $x_k \neq x_k^*$ even though x_k^* is available. Now, we introduce a variant of acceleration framework when we choose $x_k \neq x_k^*$, because x_k^* is not known.

3.3.1 Variant with Inexact Minimization

In this variant, the minimizer x_k^* is not available and we impose an additional assumption on $\mathcal{M}$ to satisfy the following condition

($\mathcal{H}_4$) the estimate x_k provided by $\mathcal{M}$ in the condition ($\mathcal{H}_3$) satisfies $\mathbb{E}\left[h_k(x_k) - h_k^*\right] \leq \varepsilon_k$ for some $\varepsilon_k \geq 0$.

Similarly to ($\mathcal{H}_3$), it is rather a definition for the parameter ε_k than a condition imposed on $h_k(x)$. In other words, one may implicitly state that $\varepsilon_k = \mathbb{E}\left[h_k(x_k) - h_k^*\right]$. The modified version of the framework is presented in Algorithm 3.4.

The next proposition gives us some insight on how to achieve acceleration, when x_k^* is not available. The proof may be found in Appendix 3.C.

Proposition 3.2 (Convergence analysis for Algorithm 3.4)**.** *Consider the optimization problem 3.2. Given a base optimization method with a convergence rate of the type (3.3),*

Algorithm 3.4 Generic Acceleration Framework with Inexact Minimization of h_k

Input: same as Algorithm 3.4;
Initialization: $y_0 = x_0$; $q = \frac{\mu}{\mu+\kappa}$; $\alpha_0 = 1$ if $\mu = 0$ or $\alpha_0 = \sqrt{q}$ if $\mu \neq 0$;
for $k = 1, \ldots, K$ **do**
    Consider a surrogate h_k satisfying $(\mathcal{H}_1)$, $(\mathcal{H}_2)$ and obtain x_k satisfying $(\mathcal{H}_3)$ and $(\mathcal{H}_4)$;
    Compute α_k in $(0, 1)$ by solving the equation $\alpha_k^2 = (1 - \alpha_k)\alpha_{k-1}^2 + q\alpha_k$.
    Update the extrapolated sequence $y_k = x_k + \beta_k(x_k - x_{k-1})$ with β_k defined in (3.10);
end for
Output: x_k (final estimate).

the scheme expressed in Algorithm 3.4 has the following convergence rate for any $\gamma \in (0, 1]$,

$$\mathbb{E}\left[F(x_k) - F^*\right] \leq \begin{cases} \left(1 - \frac{\sqrt{q}}{2}\right)^k \left(2\left(F(x_0) - F^*\right) + 4\sum_{j=1}^k \left(1 - \frac{\sqrt{q}}{2}\right)^{-j} \left(\delta_j + \frac{\varepsilon_j}{\sqrt{q}}\right)\right) & \text{if } \mu \neq 0 \\ \frac{2e^{1+\gamma}}{(k+1)^2}\left(\kappa\left\|x_0 - x^*\right\|^2 + \sum_{j=1}^k (j+1)^2\delta_j + (j+1)^{3+\gamma}\gamma^{-1}\varepsilon_j\right) & \text{if } \mu = 0. \end{cases}$$

From this result, we see that in order to maintain the accelerated rate, the sequence $(\delta_k)_{k\geq 0}$ needs to converge at a similar speed as in Proposition 3.1, but the dependency in ε_k is slightly worse. Specifically, when f is μ-strongly convex, we may have both $(\varepsilon_k)_{k\geq 0}$ and $(\delta_k)_{k\geq 0}$ decreasing at a rate $\mathcal{O}\left((1-\rho)^k\right)$ with $\rho > \sqrt{q}/2$, but we pay a factor $(1/\sqrt{q})$ compared to (3.11). When $\mu = 0$, the accelerated $\mathcal{O}\left(1/k^2\right)$ rate is preserved whenever $\varepsilon_k = \mathcal{O}\left(1/k^{4+2\gamma}\right)$ and $\delta_k = \mathcal{O}\left(1/k^{3+\gamma}\right)$, but we pay a factor $\mathcal{O}\left(1/\gamma\right)$ compared to (3.11). This is the price both for stochasticity in f and for not having $(x_k^*)_{k\geq 1}$ available.

Catalyst [Lin et al., 2018]. Let us show that in the deterministic case with $\sigma = 0$ we recover the convergence rates of the original Catalyst approach, when using a surrogate function $h_k = F(x) + (\kappa/2)\left\|x - y_{k-1}\right\|^2$ for $k \geq 1$. First, in such a case we can set up $\delta_k = \varepsilon_k$ since $\mathbb{E}\left[F(x_k)\right] \leq \mathbb{E}\left[h_k(x_k)\right] \leq \mathbb{E}\left[h_k^*\right] + \delta_k$. Second, we demand that at the k-th stage of Algorithm 3.4 the base method $\mathcal{M}$ is initialized with the previous solution x_{k-1}. This is essentially the warm start strategy from [Lin et al., 2018], which allows to explicitly express the number of iterations required by $\mathcal{M}$ in order to minimize $h_k(x)$ up to accuracy ε_k *sufficiently* quickly. For this purpose we introduce the following proposition, which is essentially a generalization of Proposition 12 from [Lin et al., 2018].

Proposition 3.3 (Warm restart for Catalyst)**.** *Consider Algorithm 3.4 with* h_k *defined in (3.4). Set up* $x_{-1} = x_0$*. Then, for* $k \geq 2$,

$$\mathbb{E}\left[h_k(x_{k-1}) - h_k^*\right] \leq \frac{3\varepsilon_{k-1}}{2} + 54\kappa \max\left\{\left\|x_{k-1} - x^*\right\|^2, \left\|x_{k-2} - x^*\right\|^2, \left\|x_{k-3} - x^*\right\|^2\right\},$$

The proof may be found in Appendix 3.C. Following [Lin et al., 2018], we may now analyze the global complexity of Algorithm 3.4 when the base method $\mathcal{M}$ behaves as (3.3). For instance, when f is μ-strongly convex, we may choose $\varepsilon_k = \mathcal{O}\left((1-\rho)^k\left(F(x_0) - F^*\right)\right)$ with $\rho = \sqrt{q}/3$. Then, it is possible to show that Proposition 3.2 yields $\mathbb{E}\left[F(x_k) - F^*\right] = \mathcal{O}\left(\varepsilon_k/q\right)$. Therefore, from the inequality $(\mu/2)\left\|x_k - x^*\right\|^2 \leq F(x_k) - F^*$ and Corollary 3.3,

we have $\mathbb{E}\left[h_k(x_{k-1}) - h_k^*\right] = \mathcal{O}\left(\kappa\varepsilon_{k-1}/(\mu q)\right) = \mathcal{O}\left(\varepsilon_{k-1}/q^2\right)$. When $\sigma = 0$, the estimate x_k is obtained by $\mathcal{M}$ in $\mathcal{O}\left(\log(1/q)/\tau\right) = \widetilde{\mathcal{O}}\left(1/\tau\right)$ iterations. This yields the global complexity $\widetilde{\mathcal{O}}\left((1/(\tau\sqrt{q}))\log(1/\varepsilon)\right)$. For example, when $\mathcal{M}$ is the proximal gradient descent method, we have $\tau = (\mu+\kappa)/(L+\kappa)$ and choosing $\kappa = L$ we obtain the global complexity $\widetilde{\mathcal{O}}\left(\sqrt{L/\mu}\log(1/\varepsilon)\right)$ of an accelerated method. Recall that $\widetilde{\mathcal{O}}\left(\cdot\right)$ notation may hide some terms logarithmic in L and μ.

In this deterministic case $\sigma = 0$, our results improve upon Catalyst [Lin et al., 2018] in two aspects that are crucial for stochastic optimization: (i) we allow the sub-problems to be solved in expectation, whereas Catalyst requires the stronger condition $h_k(x_k) - h_k^* \leq \varepsilon_k$; (ii) Proposition 3.3 removes the requirement of [Lin et al., 2018] to perform a full gradient step for initializing the method $\mathcal{M}$ in the composite case (see Proposition 12 in [Lin et al., 2018]).

Proximal gradient descent with inexact prox [Schmidt et al., 2011]. Assume that we have the surrogate (3.12), but the proximal operator can not be computed exactly, notwithstanding that $\sigma = 0$. In this case, it can be treated in the same way as Catalyst above, which provides a unified proof for both problems. Then, we recover the results of [Schmidt et al., 2011], while allowing inexact minimization to be performed in expectation.

Stochastic Catalyst. With Proposition 3.3 at hand, we are in shape to consider stochastic problems with $\sigma^2 \neq 0$ when using a method $\mathcal{M}$ that converges as (3.3). As in Section 3.2, we also assume that there exists a mini-batch/step size parameter η that can reduce the constant bias by a factor $\eta < 1$ while paying a factor $1/\eta$ in terms of complexity per stage. As above, we discuss the strongly convex case and choose the same sequence $(\varepsilon_k)_{k\geq 0}$ with $\varepsilon_k = \mathcal{O}\left((1-\rho)^k\left(F(x_0) - F^*\right)\right)$. In order to minimize h_k up to accuracy ε_k, we set $\eta_k = \min(1, \varepsilon_k/(2B\sigma^2))$ such that $\eta_k B\sigma^2 \leq \varepsilon_k/2$. Then, the complexity to minimize h_k with $\mathcal{M}$ when using the initialization x_{k-1} becomes $\widetilde{\mathcal{O}}\left(1/(\eta_k\tau)\right)$, leading to the global complexity for minimization of $F(x)$ up to accuracy ε

$$\widetilde{\mathcal{O}}\left(\frac{1}{\tau\sqrt{q}}\log\left(\frac{F(x_0) - F^*}{\varepsilon}\right)\right) + \frac{B\sigma^2}{q^{3/2}\tau\varepsilon}. \tag{3.18}$$

Details about the derivation are given in Appendix 3.B. The left term corresponds to the Catalyst accelerated rate, but it may be shown that the term on the right is sub-optimal. Indeed, consider $\mathcal{M}$ to be ISTA with $\kappa = L - \mu$. Then, $B = 1/L$, $\tau = \mathcal{O}\left(1\right)$, and the right term becomes $\widetilde{\mathcal{O}}\left(\left(\sqrt{L/\mu}\right)\sigma^2/(\mu\varepsilon)\right)$, which is sub-optimal by a factor $\sqrt{L/\mu}$. Whereas this result is a negative one, suggesting that Catalyst is not robust to noise, we show in Section 3.3.2 how to circumvent this for a large class of algorithms.

Accelerated stochastic proximal gradient descent with inexact prox. Finally, consider h_k defined in (3.15) with $\sigma \neq 0$ and let the proximal operator there be computed approximately, which, to our knowledge, has never been analyzed in the stochastic context. Then, it is shown in Appendix 3.B that Proposition 3.2 holds with $\delta_k = 2\varepsilon_k + 3\sigma^2/(2(\mu+\kappa))$. Then, an interesting question is how small should ε_k be to guarantee the optimal (or

near-optimal) dependency with respect to σ^2 as in Corollary 3.1. In the strongly convex case, Proposition 3.2 simply gives $\varepsilon_k = \mathcal{O}\left(\sqrt{q}\sigma^2/(\mu+\kappa)\right)$ such that $\delta_k \approx \varepsilon_k/\sqrt{q}$.

3.3.2 Exploiting methods $\mathcal{M}$ providing strongly convex surrogates

Among various aforementioned application cases, we have seen an extension of Catalyst to stochastic problems. To achieve convergence, the strategy was to use a specific mechanism to reduce the bias $B\sigma^2$ in (3.3), *e.g.*, by using mini-batches or decreasing step sizes described in Section 3.2. Yet, the approach suffers from two issues: (i) some of the parameters are based on unknown quantities such as σ^2; (ii) the worst-case complexity exhibits a sub-optimal dependency in σ^2, typically with an extra factor $1/\sqrt{q}$ when $\mu > 0$. Whereas practical workarounds for the first point are discussed in Section 3.4, we now show how to solve the second one in many cases by using Algorithm 3.3.

Denote a sequence of surrogate functions $(H_k)_{k\geq 1}$ with $H_k(x) = F(x) + (\kappa/2)\left\|x - y_{k-1}\right\|^2$ for some positive smoothing parameter $\kappa > 0$. Consider a base method $\mathcal{M}$ with convergence (3.3), which is able, after T steps, to produce a point x_k such that for some $\xi_{k-1} > 0$ the following bound holds

$$\mathbb{E}\left[H_k(x_k) - h_k^*\right] \leq C\left(1-\tau\right)^T \left(H_k(x_{k-1}) - H_k^* + \xi_{k-1}\right) + B\sigma^2. \tag{3.19}$$

where h_k is a function satisfying $(\mathcal{H}_1)$, $(\mathcal{H}_2)$ that can be minimized in closed form. In this sense, h_k is a surrogate function for $F(x)$, because, given (3.19), $(\mathcal{H}_3)$ is immediately satisfied with δ_k being the right side of (3.19)

$$\delta_k = C\left(1-\tau\right)^T \left(H_k(x_{k-1}) - H_k^* + \xi_{k-1}\right) + B\sigma^2.$$

The value of ξ_{k-1} is typically chosen as $\xi_{k-1} = \mathcal{O}\left(\mathbb{E}\left[F(x_{k-1}] - F^*\right)\right)$. In this section, the functions $(h_k)_{k\geq 1}$ are assumed to be constructed explicitly by the base optimization method $\mathcal{M}$. Specifically, we will associate $(h_k)_{k\geq 1}$ with the estimate sequences $(d_k)_{k\geq 1}$ of Chapter 2 when accelerating the variance-reduced algorithms generalized there to the stochastic setting. In a sense, h_k may be seen as a supporting model of the function $H_k(x)$ with a "simpler form".

Let us now explain the acceleration procedure of this section in more details. In a nutshell, we apply Algorithm 3.3, where the base method $\mathcal{M}$ is used to perform *approximate minimization* of $H_k = F(x) + (\kappa/2)\left\|x - y_{k-1}\right\|^2$. However, at the post-processing step we refer to *another surrogate*, namely h_k, with closed-form minimizer. This function may be provided by an oracle or constructed by $\mathcal{M}$ during the solving of the k-th sub-problem. As the minimizer of h_k is available, we are able to apply the update (3.10) and finally benefit from Algorithm 3.3, which has better convergence guarantees than Algorithm 3.4.

As shown in Appendix 3.D, even though the condition (3.19) looks technical, a large class of optimization techniques are able to provide a solution x_k and surrogate $h_k(x)$ satisfying it. The examples include many variants of proximal stochastic gradient descent methods with variance reduction such as SAGA [Defazio et al., 2014a], MISO [Mairal, 2015], SDCA [Shalev-Shwartz and Zhang, 2016], or SVRG [Xiao and Zhang, 2014] due to the theoretical results of Chapter 2. In this sense, the constructed surrogates $(h_k)_{k\geq 1}$ may be set up as the estimate sequences $(d_k)_{k\geq 1}$ from relation (2.8) of Chapter 2.

Method $\mathcal{M}$	κ	τ	B	q	Complexity after Catalyst
prox-SGD	$L-\mu$	$\frac{1}{2}$	$\frac{1}{L}$	$\frac{\mu}{L}$	$\widetilde{\mathcal{O}}\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{F_0}{\varepsilon}\right)+\frac{\sigma^2}{\mu\varepsilon}\right)$
SVRG/SAGA/MISO etc. of Chapter 2 with $\frac{L}{n}\geq\mu$	$\frac{L}{n}-\mu$	$\frac{1}{n}$	$\frac{1}{L}$	$\frac{\mu n}{L}$	$\widetilde{\mathcal{O}}\left(\sqrt{n\frac{L}{\mu}}\log\left(\frac{F_0}{\varepsilon}\right)+\frac{\tilde{\sigma}^2}{\mu\varepsilon}\right)$

Table 3.1 – Meta-parameters τ and B describing convergences of different methods according to (3.3) along with the practical choices of κ and corresponding values of q. This parameter description is accompanied with the final complexity of the methods being wrapped by our accelerating framework. The value of F_0 denotes the initialization error $F(x_0)-F^*$, the values of σ^2 and $\tilde{\sigma}^2$ are defined in (1.10) and (2.18) respectively. Once again, $\widetilde{\mathcal{O}}(\cdot)$ may hide some logarithmic factors.

Now, whereas (3.19) seems to be a minor modification of (3.3), an important consequence is that it will allow us to gain a factor $1/\sqrt{q}$ in the variance complexity when $\mu>0$, corresponding precisely to the sub-optimality factor. Therefore, even though the surrogate H_k needs only to be minimized approximately, the condition (3.19) allows us to use Algorithm 3.3 instead of Algorithm 3.4. Given that the dependency with respect to δ_k is better than ε_k (by $1/\sqrt{q}$), we have then the following result:

Proposition 3.4 (Stochastic Catalyst with Optimality Gaps, $\mu>0$)**.** *Consider Algorithm 3.3 with a method $\mathcal{M}$ and surrogates h_k satisfying (3.19) when $\mathcal{M}$ is used to minimize H_k by using x_{k-1} as a warm restart. Assume that f is μ-strongly convex and that there exists a parameter η that can reduce the bias $B\sigma^2$ by a factor $\eta<1$ while paying the same factor in terms of complexity per stage.*

Choose $\delta_k=\mathcal{O}\left(\left(1-\sqrt{q}/2\right)^k\left(F(x_0)-F^*\right)\right)$ and $\eta_k=\min\{1,\delta_k/2B\sigma^2\}$. Then, the complexity to minimize approximately $H_k(x)$ and compute x_k satisfying (3.19) is $\widetilde{\mathcal{O}}(1/\eta_k\tau)$, so that the global complexity to obtain $\mathbb{E}[F(x_k)-F^*]\leq\varepsilon$ is

$$\widetilde{\mathcal{O}}\left(\frac{1}{\tau\sqrt{q}}\log\left(\frac{F(x_0)-F^*}{\varepsilon}\right)+\frac{B\sigma^2}{q\tau\varepsilon}\right).$$

The term on the left is the accelerated rate of the Catalyst approach for deterministic problems, whereas the term on the right is potentially near-optimal for strongly convex problems, as illustrated in Table 3.1. We provide indeed practical choices for the parameters κ, leading to various values of B,τ,q, for the proximal stochastic gradient descent method with iterate averaging as well as for stochastic variants of SVRG, SAGA, MISO, SDCA, Finito from Chapter 2. All the values below are given up to universal constants to simplify the presentation.

The aforementioned incremental algorithms are applied to the stochastic finite-sum optimization problem with n L-smooth functions (3.2). The corresponding values of τ and B for them and for the proximal variants of SGD follow from Corollaries 2.15 and 2.3 respectively. As well as in the deterministic case considered by [Lin et al., 2018], we note that when $L/n\leq\mu$, there is no acceleration for the incremental algorithms

since the complexity of the unaccelerated method $\mathcal{M}$ is $\widetilde{\mathcal{O}}\left(n\log\left(F_0/\varepsilon\right)+\sigma^2/\mu\varepsilon\right)$, which is independent of the condition number and already near-optimal, see Corollary 2.12. In comparison, the logarithmic terms in L,μ that are hidden in the notation $\tilde{O}$ do not appear for a variant of the SVRG method with direct acceleration introduced in Chapter 2. However, while we managed there to directly accelerate only the SVRG approach, our approach developed in the current chapter is more generic. Note also that σ^2 for prox-SGD is typically much larger than $\tilde{\sigma}^2$ for the incremental algorithms since the source of randomness is larger for prox-SGD, see comparison of 1.28 with 1.12 in Section 1.4.5.

3.4 Experiments

In this section, we perform numerical evaluations by following the steps of Section 2.5, where we were able to make SVRG and SAGA robust to stochastic noise, and, in addition, accelerate SVRG.

Formulation We consider ℓ_2-logistic regression and support vector machine with the squared hinge loss. More specifically, given training data $(a_i,b_i)_{i=1,\dots,n}$, with a_i in $\mathbb{R}^p$ and b_i in $\{-1,+1\}$, we consider the optimization problem

$$\min_{x\in\mathbb{R}^p}\frac{1}{n}\sum_{i=1}^{n}\phi(b_ia_i^\top x)+\frac{\lambda}{2}\left\|x\right\|^2,$$

where ϕ is either the logistic loss $\phi(u)=\log(1+e^{-u})$, or the squared hinge loss $\phi(u)=\frac{1}{2}\max(0,1-u)^2$, which are both L-smooth, with $L=0.25$ for logistic and $L=1$ for the squared hinge loss. The scalar λ is a regularization parameter that acts as a lower bound on the strong convexity constant μ of the problem. It is chosen among the smallest values one would try when performing parameter search, *e.g.*, by cross validation. Specifically, we consider $\lambda=1/10n$ and $\lambda=1/100n$, where n is the number of training points; we also try $\lambda=1/1000n$ in order to evaluate the numerical stability of methods on very ill-conditioned problems. Studying the squared hinge loss is interesting since its gradients are unbounded on the optimization domain, which may break the bounded noise assumption. Following [Bietti and Mairal, 2017, Zheng and Kwok, 2018] and Section 2.5, we consider DropOut perturbations [Srivastava et al., 2014]. DropOut consists of randomly setting to zero each entry of a data point with probability δ, leading to the optimization problem

$$\min_{x\in\mathbb{R}^p}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_\rho\left[\phi(b_i(\rho\circ a_i)^\top x)\right]+\frac{\lambda}{2}\left\|x\right\|^2, \tag{3.20}$$

where ρ is a binary vector in $\{0,1\}^p$ with i.i.d. Bernoulli entries that are set up according to probability δ, and $\circ$ denotes the element-wise multiplication between two vectors. We consider DropOut perturbations with rates $\delta=0$ (no noise), $\delta=0.01$ and $\delta=0.1$. The purpose of using such perturbation is to manually emulate inexactness in the oracle.

Datasets We consider the same datasets used in Section 2.5 with all points being normalized to have unit ℓ_2-norm.

— alpha is from the Pascal Large Scale Learning Challenge website [1] and contains

1. http://largescale.ml.tu-berlin.de/

$n = 250\,000$ points in dimension $p = 500$.

— gene consists of gene expression data and the binary labels b_i characterize two different types of breast cancer. This is a small dataset with $n = 295$ and $p = 8\,141$.

— ckn-cifar is an image classification task where each image from the CIFAR-10 dataset [2] is represented by using a two-layer unsupervised convolutional neural network [Mairal, 2016]. We consider here the binary classification task consisting of predicting the class 1 vs. other classes. The dataset contains $n = 50\,000$ images and the dimension of the representation is $p = 9\,216$.

Methods We consider the variants of SVRG and SAGA of Chapter 2, which use decreasing step sizes when $\delta > 0$ (otherwise, they do not converge). We use the suffix "-d" each time decreasing step sizes are used in order to designate curves on plots. We also consider Katyusha [Allen-Zhu, 2017] when $\delta = 0$, and the accelerated SVRG method of Section 2.4 from Chapter 2, denoted by acc-SVRG. Finally, SVRG-d, SAGA-d, acc-SVRG-d are used with the step size strategies described in Chapter 2.

Practical questions and implementation. In all setups, we choose the parameter κ according to the theory, which are described in the previous section, following the Catalyst approach. For composite problems, Proposition 3.3 suggests to use x_{k-1} as a warm start for the sub-problems. For the smooth ones, [Lin et al., 2018] shows that, in fact, other choices such as y_{k-1} are appropriate and lead to similar complexity results. In our experiments with smooth losses we use y_{k-1}, which has shown to perform consistently better.

The strategy for η_k discussed in Proposition 3.4 suggests to use constant step sizes in the first stages, typically of order $1/(\kappa + L)$ for the methods we consider, before using an exponentially decreasing schedule. Unfortunately, even though theory suggests a rate of decay in $(1 - \sqrt{q}/2)^k$, it does not provide useful insight on when decaying should start since it requires knowledge of σ^2. A similar issue arises in stochastic optimization techniques involving iterate averaging [Bottou et al., 2018]. We adopt a similar heuristic as in this literature and start step size decaying after k_0 epochs, with $k_0 = 30$. Finally, we discuss the number of iterations of $\mathcal{M}$ to perform per stage. When $\eta_k = 1$, the theoretical value is of order $\tilde{O}(1/\tau) = \tilde{O}(n)$, and we choose exactly n iterations (one epoch), as in Catalyst [Lin et al., 2018]. After starting decaying the step sizes ($\eta_k < 1$), we use $\lceil n/\eta_k \rceil$, according to the theory.

Making plots. We run each experiment five times and average the outputs. We display plots on a logarithmic scale for the primal gap $F(x_k) - F^*$ (with F^* estimated as the minimum value observed from all runs), and the x-axis denotes the number of epochs. Note that for SVRG, one iteration is considered to perform two epochs since it requires accessing the full dataset every n iterations on average. The colored tubes around each curve denote one standard deviation across 5 runs.

Acceleration with no noise, $\delta = 0$. We start evaluating the acceleration approach when there is no noise. This is essentially evaluating the Catalyst method [Lin et al., 2018]

2. `https://www.cs.toronto.edu/~kriz/cifar.html`

in a deterministic setup in order to obtain a baseline comparison when $\delta = 0$. The results are presented in Figures 3.1 and 3.2 for the logistic regression problem. As predicted by theory, acceleration is more important when conditioning is low (bottom curves).

Stochastic acceleration with no noise, $\delta = 0.01$ and $\delta = 0.1$. Then, we perform a similar experiments by adding noise and report the results in Figures 3.3, 3.4, 3.5, 3.6, 3.7. In general, the stochastic Catalyst approach seems to perform on par with the accelerated SVRG approach of Chapter 2 and even better in one case, showing that generic acceleration may be useful even in the stochastic optimization regime, consistently with Section 2.5.

Evaluating the square hinge loss. In Figure 3.8, we perform experiments using the square hinge loss, where the methods perform similarly as for the logistic regression case, despite the fact that the bounded noise assumption does not necessarily hold on the optimization domain for the square hinge loss.

Evaluating ill-conditioned problems. In Figure 3.10, we study the methods behavior on badly conditioned problems. There, acceleration seems to work on ckn-cifar, but fails on gene and alpha, suggestions that acceleration is difficult to achieve when the condition number is extremely low. However, we should note that the ill-conditioning obtained by choosing $\lambda = 1/1000n$ is unrealistic in the context of empirical risk minimization

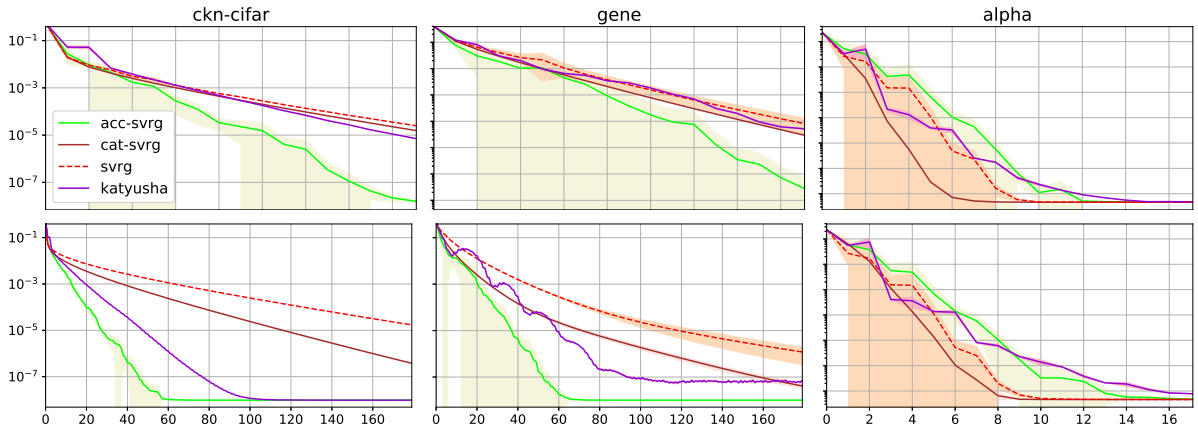


Figure 3.1 – Accelerating SVRG-like methods for ℓ_2-logistic regression with $\lambda = 1/(10n)$ (top) and $\lambda = 1/(100n)$ (bottom) for $\delta = 0$.

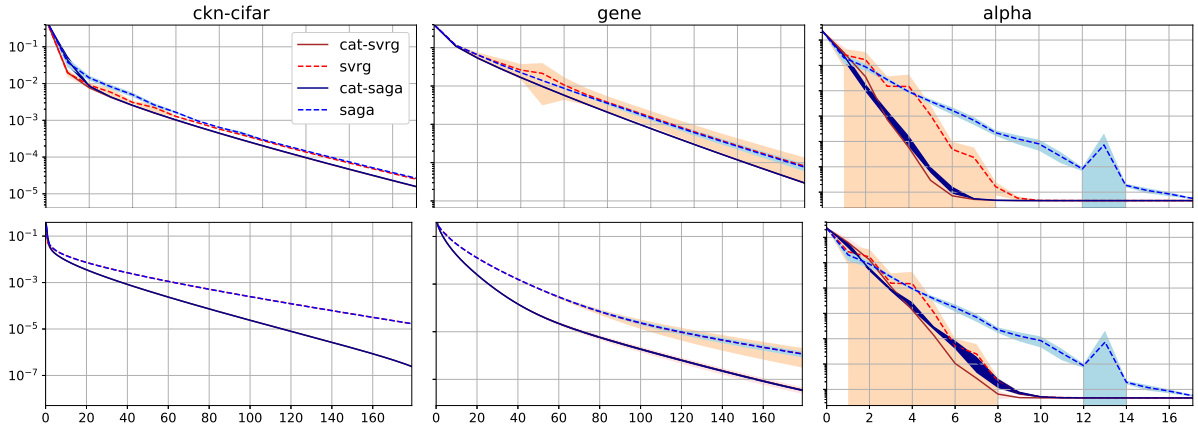


Figure 3.2 – Same plots as in Figure 3.1 when comparing SVRG and SAGA, with no noise ($\delta = 0$) with $\lambda = 1/(10n)$ (top) and $\lambda = 1/(100n)$ (bottom) .

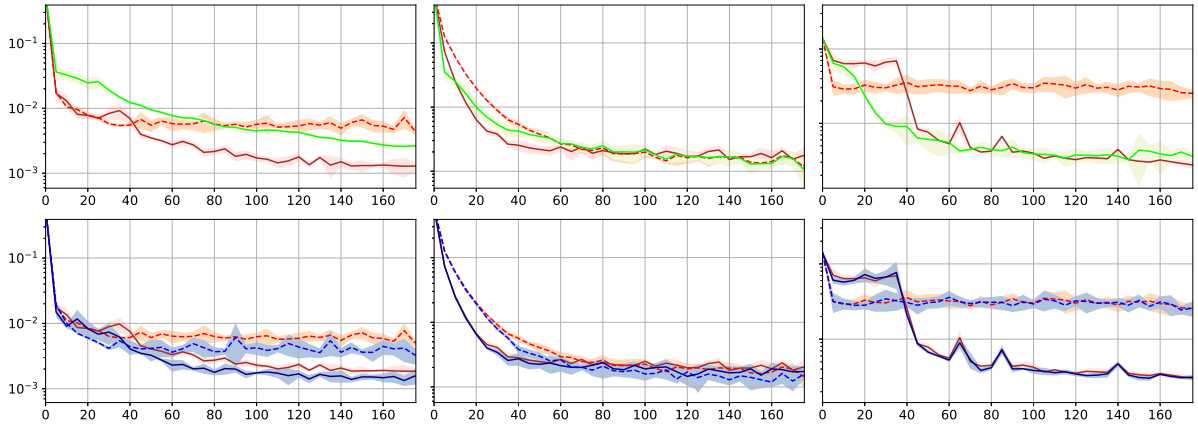


Figure 3.3 – Accelerating SVRG-like (top) and SAGA (bottom) methods for ℓ_2-logistic regression with $\lambda = 1/(100n)$ (bottom) for $\delta = 0.1$.

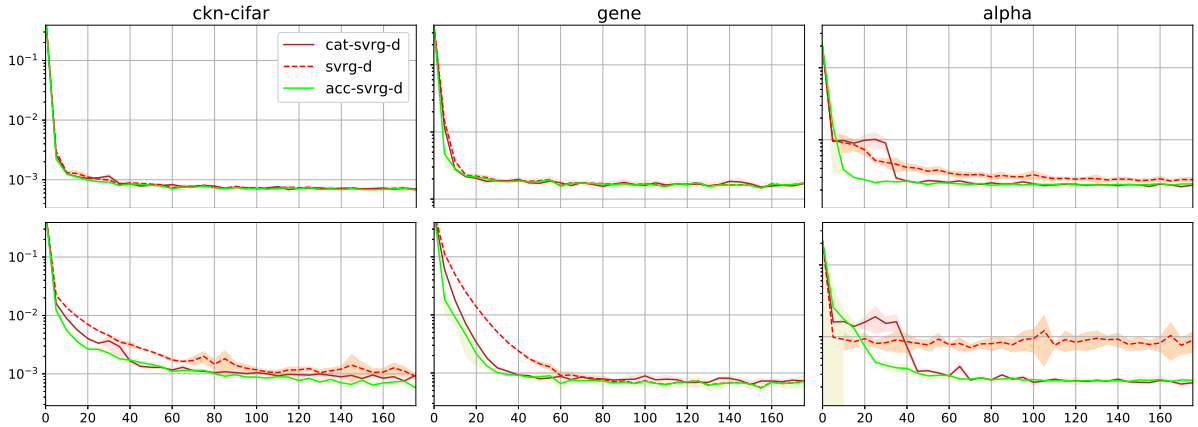


Figure 3.4 – Same plots as in Figure 3.1 for $\delta = 0.01$ with $\lambda = 1/(10n)$ (top) and $\lambda = 1/(100n)$ (bottom).

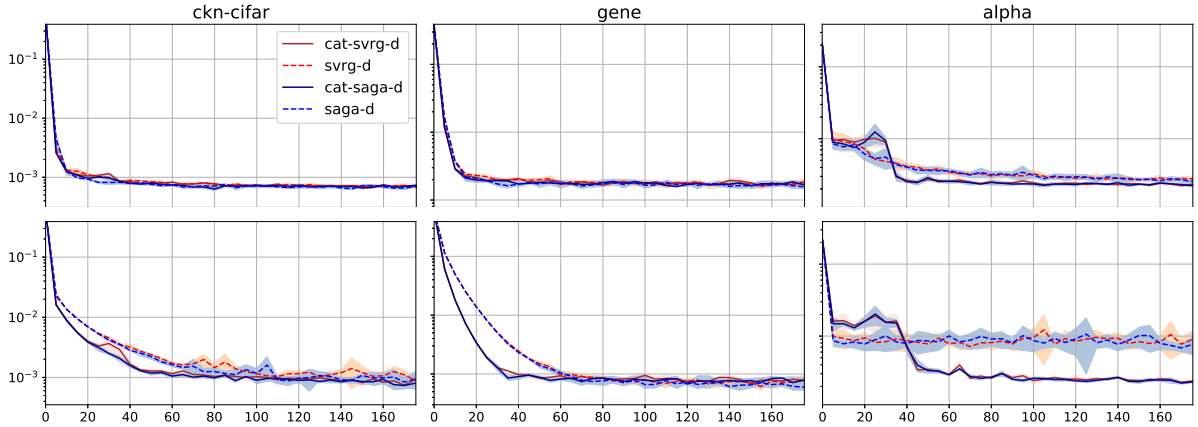


Figure 3.5 – Same plots as in Figure 3.2 for $\delta = 0.01$ with $\lambda = 1/(10n)$ (top) and $\lambda = 1/(100n)$ (bottom).

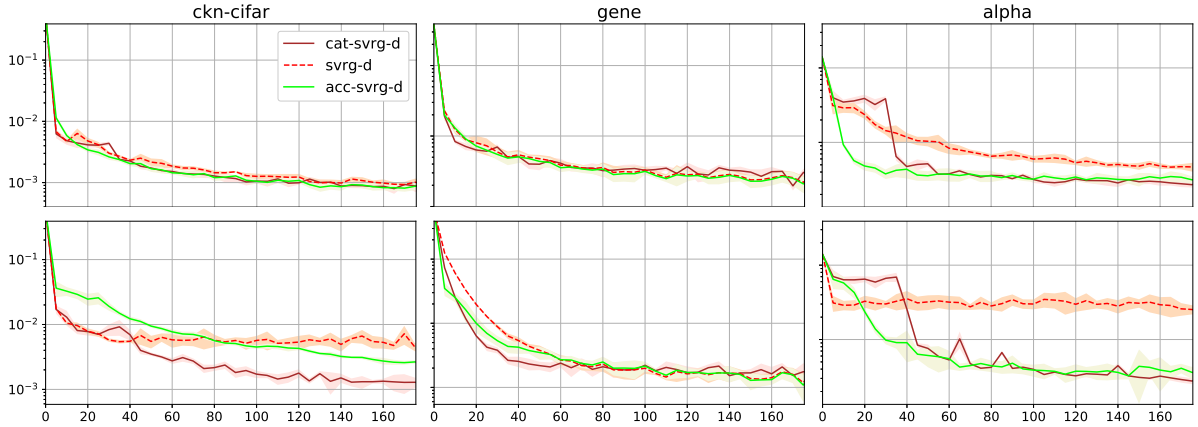


Figure 3.6 – Same plots as in Figure 3.1 for $\delta = 0.1$ with $\lambda = 1/(10n)$ (top) and $\lambda = 1/(100n)$ (bottom).

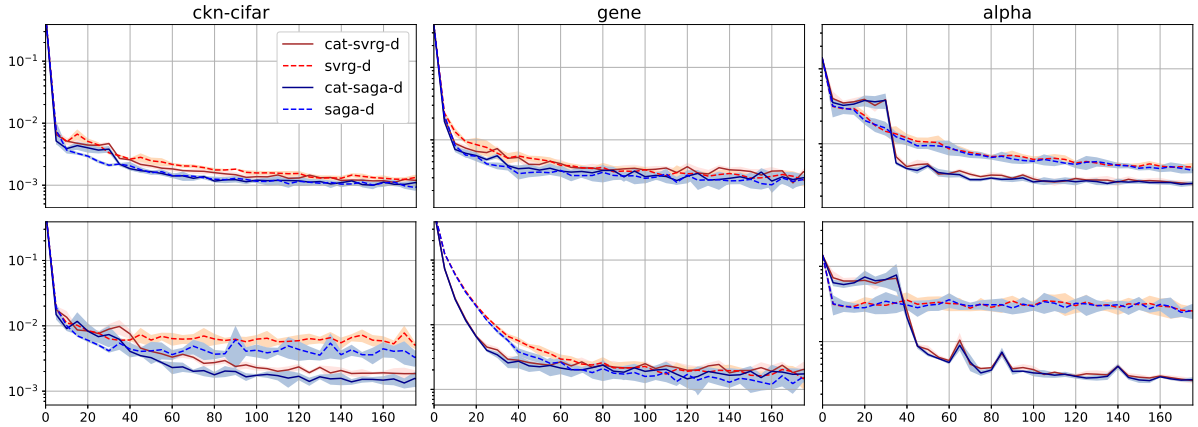


Figure 3.7 – Same plots as in Figure 3.2 for $\delta = 0.1$ with $\lambda = 1/(10n)$ (top) and $\lambda = 1/(100n)$ (bottom).

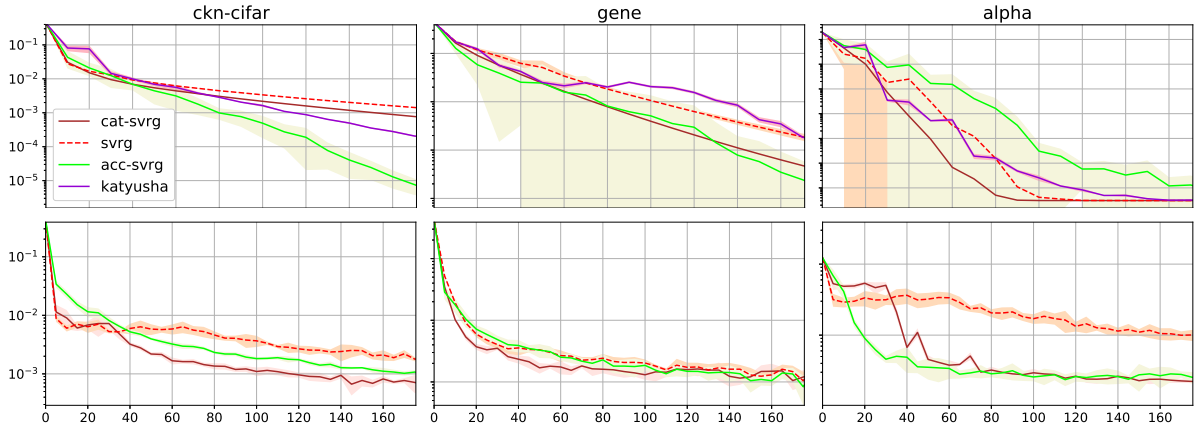


Figure 3.8 – Accelerating SVRG-like methods when using the squared hinge loss instead of the logistic for $\delta = 0$ (top) and $\delta = 0.1$, both with $\lambda = 1/(10n)$.

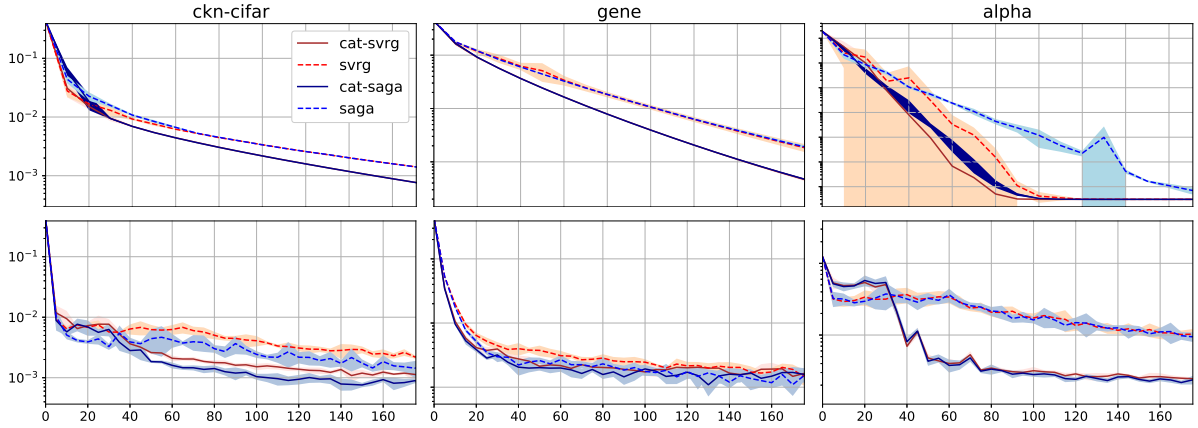


Figure 3.9 – Same plots as in Figure 3.8 for SVRG and SAGA, with $\delta = 0$ (top) and $\delta = 0.1$ for $\lambda = 1/(10n)$.

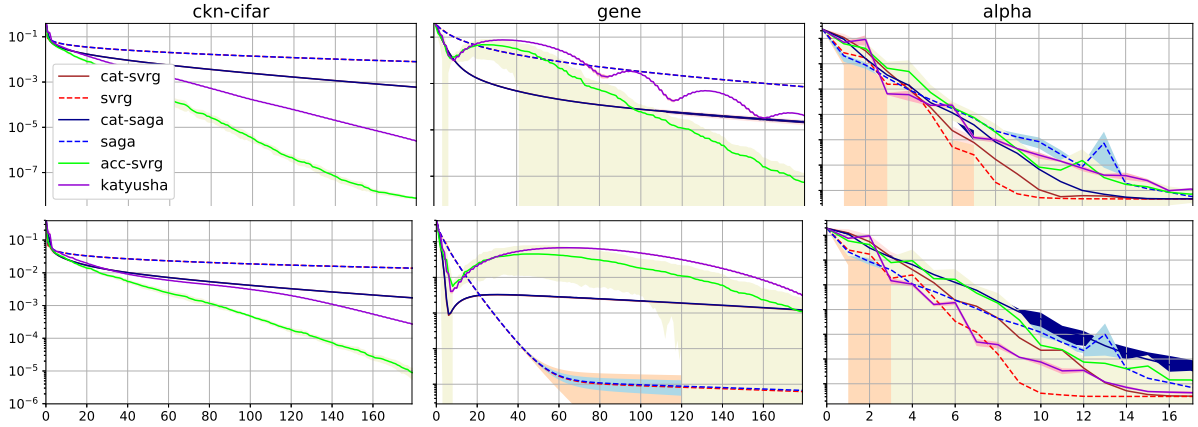


Figure 3.10 – Illustration of potential numerical instabilities problems when the problem is very ill-conditioned. We use $\lambda = 1/(1000n)$ with $\delta = 0$ for the logistic loss (top) and squared hinge (bottom).

Appendix

3.A Useful Results and Definitions

In this section, we present auxiliary results that are used in the subsequent proofs.

Lemma 3.1 (Convergence rate of the sequences $(\alpha_k)_{k\geq 0}$ and $(A_k)_{k\geq 0}$)**.** *Consider the sequence in* $(0,1)$ *defined by the recursion*

$$\alpha_k^2 = (1-\alpha_k)\,\alpha_{k-1}^2 + q\alpha_k \quad \textit{with} \quad 0 \leq q < 1,$$

and define $A_k = \prod_{t=1}^{k}(1-\alpha_t)$*. Then,*

— *if* $q = 0$ *and* $\alpha_0 = 1$*, then, for all* $k \geq 1$*,*

$$\frac{2}{(k+2)^2} \leq A_k = \alpha_k^2 \leq \frac{4}{(k+2)^2}.$$

— *if* $\alpha_0 = \sqrt{q}$*, then for all* $k \geq 1$*,*

$$A_k = (1-\sqrt{q})^k \quad \textit{and} \quad \alpha_k = \sqrt{q}.$$

— *if* $\alpha_0 = 1$*, then for all* $k \geq 1$*,*

$$A_k \leq \min\left((1-\sqrt{q})^k, \frac{4}{(k+2)^2}\right) \quad \textit{and} \quad \alpha_k \geq \max\left(\sqrt{q}, \frac{\sqrt{2}}{k+2}\right).$$

Proof. We prove the three points, one by one.

First point. Let us prove the first point when $q = 0$ and $\alpha_0 = 1$. The relation $A_k = \alpha_k^2$ is obvious for all $k \geq 1$ and the relation $\alpha_k^2 \leq \frac{4}{(k+2)^2}$ holds for $k = 0$. By induction, let us assume that we have the relation $\alpha_{k-1}^2 \leq \frac{4}{(k+1)^2}$ and let us show that it propagates for α_k^2. Assume, by contradiction, that $\alpha_k^2 > \frac{4}{(k+2)^2}$, meaning that $\alpha_k > \frac{2}{(k+2)}$. Then,

$$\alpha_k^2 = (1-\alpha_k)\alpha_{k-1}^2 \leq (1-\alpha_k)\frac{4}{(k+1)^2} < \frac{4k}{(k+2)(k+1)^2} = \frac{4}{(k+2)(k+2+\frac{1}{k})} < \frac{4}{(k+2)^2},$$

and we obtain a contradiction. Therefore, $\alpha_k^2 \leq \frac{4}{(k+2)^2}$ and the induction hypothesis allows us to conclude for all $k \geq 1$. Then, note [Paquette et al., 2018] that we also have for all $k \geq 1$,

$$A_k = \prod_{t=1}^{k}(1-\alpha_t) \geq \prod_{t=1}^{k}\left(1-\frac{2}{t+2}\right) = \frac{2}{(k+1)(k+2)} \geq \frac{2}{(k+2)^2}.$$

Second point. The second point is obvious by induction.

Third point. For the third point, we simply assume $\alpha_0 = 1$ such that $\alpha_0 \geq \sqrt{q}$. Then, the relation $\alpha_k \geq \sqrt{q}$ and therefore $A_k \leq \left(1-\sqrt{q}\right)^k$ are easy to show by induction. Then, consider the sequence defined recursively by $u_k^2 = (1-u_k)u_{k-1}^2$ with $u_0 = 1$. From the first point, we have that $\frac{\sqrt{2}}{k+2} \leq u_k \leq \frac{2}{k+2}$. We will show that $\alpha_k \geq u_k$ for all $k \geq 0$, which will be sufficient to conclude since then we would have $A_k \leq \prod_{t=1}^{k}(1-u_t) \leq \frac{4}{(k+2)^2}$. First, we note that $\alpha_0 = u_0$; then, assume that $\alpha_{k-1} \geq u_{k-1}$ and also assume by contradiction that $\alpha_k > u_k$. This implies that

$$u_k^2 = (1-u_k)u_{k-1}^2 \leq (1-u_k)\alpha_{k-1}^2 < (1-\alpha_k)\alpha_{k-1}^2 \leq \alpha_k^2,$$

which contradicts the assumption $\alpha_k > u_k$. This allows us to conclude by induction. □

Lemma 3.2 (Convergence rate of sequences $\Theta_k = \prod_{i=1}^{k}(1-\theta_i)$)**.** *Consider the sequence* $\theta_j = \frac{\gamma}{(1+j)^{1+\gamma}}$ *with* γ *in* $(0,1]$*. Then,*

$$e^{-(1+\gamma)} \leq \Theta_k \leq 1. \tag{3.21}$$

Proof. We use the classical inequality $\log(1+u) \geq \frac{u}{1+u}$ for all $u > -1$:

$$-\log(\Theta_k) = -\sum_{j=1}^{k}\log\left(1-\frac{\gamma}{(1+j)^{1+\gamma}}\right) \leq \sum_{j=1}^{k}\frac{\gamma}{(1+j)^{1+\gamma}-\gamma} \leq \sum_{j=1}^{k}\frac{\gamma}{j^{1+\gamma}},$$

when noting that the function $g(x) = (1+x)^{1+\gamma} - x^{1+\gamma}$ is greater than γ for all $x \geq 1$, since $g(1) \geq 1 \geq \gamma$ and g is non-decreasing. Then,

$$-\log(\Theta_k) \leq \sum_{j=1}^{k}\frac{\gamma}{j^{1+\gamma}} \leq \gamma + \gamma\int_{x=1}^{k}\frac{1}{x^{1+\gamma}}dx = \gamma + 1 - \frac{1}{k^\gamma} \leq \gamma + 1.$$

Then, we immediately obtain (3.21).

□

3.B Details about Complexity Results

3.B.1 Details about (3.6)

Consider the complexity (3.3) with $h = f$. To achieve the accuracy $2B\sigma^2$, it is sufficient to run the method $\mathcal{M}$ for t_0 iterations, such that

$$C(1-\tau)^{t_0}\left(F(x_0) - F^*\right) \leq B\sigma^2.$$

It is then easy to see that this inequality is satisfied as soon as t_0 is greater than $\frac{1}{\tau}\log(C\left(F(x_0)-F^*\right)/B\sigma^2)$. Since $\varepsilon \leq B\sigma^2$ and using the concavity of the logarithm function, it is also sufficient to choose $t_0 = \frac{1}{\tau}\log(C\left(F(x_0)-F^*\right)/\varepsilon)$.

Then, we perform K restart stages such that $\varepsilon_K \leq \varepsilon$. Each stage is initialized with a point x_k satisfying $\mathbb{E}\left[F(x_k)-F^*\right] \leq \varepsilon_{k-1}$, and the goal of each stage is to reduce the error by a factor $1/2$. Given that η_k increases the computational cost, the complexity of the k-th stage is then upper-bounded by $\frac{2^k}{\tau}\log(2C)$, leading to the global complexity

$$\mathcal{O}\left(\frac{1}{\tau}\log\left(\frac{C\left(F(x_0)-F^*\right)}{\varepsilon}\right)+\sum_{k=1}^{K}\frac{2^k}{\tau}\log\left(2C\right)\right) \quad \text{with} \quad K=\left\lceil \log_2\left(\frac{2B\sigma^2}{\varepsilon}\right)\right\rceil,$$

and (3.6) follows by elementary calculations.

3.B.2 Obtaining (3.6) from (3.7)

Since h is μ-strongly convex, we notice that (3.7) implies the rate

$$\mathbb{E}\left[h(z_t)-h^*\right] \leq \frac{D(h(z_0)-h^*)}{\mu t^d}+\frac{B\sigma^2}{2},$$

by using the strong convexity inequality $h(z_0) \geq h^* + \frac{\mu}{2}\|z_0 - z^*\|^2$. After running the algorithm for $t' = \lceil (2D/\mu)^{1/d} \rceil$ iterations, we can show that

$$\mathbb{E}\left[h(z_{t'})-h^*\right] \leq \frac{h(z_0)-h^*}{2}+\frac{B\sigma^2}{2}.$$

Then, when restarting the procedure s times (using the solution of the previous iteration as initialization), and denoting by $h_{st'}$ the last iterate, it is easy to show that

$$\mathbb{E}\left[h(x_{st'})-h^*\right] \leq \frac{h(x_0)-h^*}{2^s}+\frac{B\sigma^2}{2}\left(\sum_{i=0}^{s-1}\frac{1}{2^i}\right) \leq \frac{h(z_0)-h^*}{2^s}+B\sigma^2.$$

Then, calling $t = st'$, we can use the inequality $2^{-u} \leq 1-\frac{u}{2}$ for u in $[0,1]$, due to convexity, and

$$\mathbb{E}\left[h(z_t)-h^*\right] \leq \left(h(z_0)-h^*\right)\left(2^{-1/t'}\right)^t + B\sigma^2 = \left(h(z_0)-h^*\right)\left(1-\frac{1}{2t'}\right)^t + B\sigma^2,$$

which gives us (3.3) with $C=1$ and $\tau = \frac{1}{2t'}$. It is then easy to obtain (3.6) by following similar steps as in Section 3.B.1, by noticing that the restart frequency is of the same order $\mathcal{O}\left(1/\tau\right)$.

3.B.3 Details about (3.18)

Inner-loop complexity. Since η_k is chosen such that the bias $\eta_k B\sigma^2$ is smaller than ε_k, the number of iterations of $\mathcal{M}$ to solve the sub-problem is $\mathcal{O}\left(1/\tau\right) = \mathcal{O}\left(\log\left(1/q\right)/\tau\right)$, as in the deterministic case, and the complexity is thus $\mathcal{O}\left(1/(\eta_k\tau)\right)$.

Outer-loop complexity. Since

$$\mathbb{E}\left[F(x_k) - F^*\right] \leq \mathcal{O}\left(\left(1 - \sqrt{q}/3\right)^k \left(F(x_0) - F^*\right)/q\right)$$

according to Proposition 3.2, it suffices to choose

$$K = \mathcal{O}\left(\frac{1}{\sqrt{q}} \log\left(\frac{F(x_0) - F^*}{q\varepsilon}\right)\right)$$

iterations to guarantee

$$\mathbb{E}\left[F(x_K) - F^*\right] \leq \varepsilon = \mathcal{O}\left(\varepsilon_K/q\right) = \mathcal{O}\left(\left(1 - \sqrt{q}/3\right)^K \left(F(x_0) - F^*\right)/q\right)$$

Global complexity. The total complexity to guarantee $\mathbb{E}\left[F(x_k) - F^*\right] \leq \varepsilon$ is then

$$\begin{aligned}
C &= \sum_{k=1}^{K} \widetilde{\mathcal{O}}\left(\frac{1}{\eta_k \tau}\right) \\
&\leq \widetilde{\mathcal{O}}\left(\sum_{k=1}^{K} \frac{1}{\tau} + \sum_{k=1}^{K} \frac{B\sigma^2}{\varepsilon_k \tau}\right) \\
&= \widetilde{\mathcal{O}}\left(\sum_{k=1}^{K} \frac{1}{\tau} + \sum_{k=1}^{K} \frac{B\sigma^2}{\tau\left(1 - \frac{\sqrt{q}}{3}\right)^k \left(F(x_0) - F^*\right)}\right) \\
&= \widetilde{\mathcal{O}}\left(\frac{1}{\tau\sqrt{q}} \log\left(\frac{F(x_0) - F^*}{\varepsilon}\right) + \frac{B\sigma^2}{\tau\sqrt{q}\left(1 - \frac{\sqrt{q}}{3}\right)^{K+1} \left(F(x_0) - F^*\right)}\right) \\
&= \widetilde{\mathcal{O}}\left(\frac{1}{\tau\sqrt{q}} \log\left(\frac{F(x_0) - F^*}{\varepsilon}\right) + \frac{B\sigma^2}{q^{3/2}\varepsilon\tau}\right),
\end{aligned}$$

where the last relation uses the fact that $\varepsilon = \mathcal{O}\left(\varepsilon_K/q\right) = \mathcal{O}\left(\left(1 - \sqrt{q}/3\right)^K \left(F(x_0) - F^*\right)/q\right)$.

3.B.4 Complexity of accelerated stochastic proximal gradient descent with inexact prox

Assume that $h_k(x_k) - h_k^* \leq \varepsilon_k$. Then, following similar steps as in (3.17),

$$\begin{aligned}
\mathbb{E}\left[F(x_k)\right] &\leq \mathbb{E}\left[h_k(x_k)\right] + \mathbb{E}\left[\left(g_k - \nabla f(y_{k-1})\right)^\top (x_k - y_{k-1})\right] \\
&= \mathbb{E}\left[h_k(x_k)\right] + \mathbb{E}\left[\left(g_k - \nabla f(y_{k-1})\right)^\top (x_k - u_{k-1})\right] \\
&= \mathbb{E}\left[h_k(x_k)\right] + \mathbb{E}\left[\left(g_k - \nabla f(y_{k-1})\right)^\top (x_k - x_k^*)\right] + \mathbb{E}\left[\left(g_k - \nabla f(y_{k-1})\right)^\top (x_k^* - u_{k-1})\right] \\
&\leq \mathbb{E}\left[h_k(x_k)\right] + \mathbb{E}\left[\left(g_k - \nabla f(y_{k-1})\right)^\top (x_k - x_k^*)\right] + \frac{\sigma^2}{\mu + \kappa}
\end{aligned}$$

$$\leq \mathbb{E}\left[h_k(x_k)\right] + \frac{\mathbb{E}\left[\|g_k - \nabla f(y_{k-1})\|^2\right]}{2(\mu+\kappa)} + \frac{(\mu+\kappa)\mathbb{E}\left[\|x_k - x_k^*\|^2\right]}{2} + \frac{\sigma^2}{\mu+\kappa}$$
$$\leq \mathbb{E}\left[h_k(x_k)\right] + \mathbb{E}\left[h_k(x_k) - h_k^*\right] + \frac{3\sigma^2}{2(\mu+\kappa)}$$
$$\leq \mathbb{E}\left[h_k^*\right] + 2\varepsilon_k + \frac{3\sigma^2}{2(\mu+\kappa)}.$$

And thus, $\delta_k = 2\varepsilon_k + \frac{3\sigma^2}{2(\mu+\kappa)}$.

3.C Proofs of Main Results

3.C.1 Proof of Propositions 3.1 and 3.2

Proof. In order to treat both propositions jointly, we introduce the quantity

$$w_k = \begin{cases} x_k & \text{for variant } \mathcal{A} \\ x_k^* & \text{for variant } \mathcal{B} \end{cases},$$

and, for all $k \geq 1$,

$$v_k = w_k + \frac{1-\alpha_{k-1}}{\alpha_{k-1}}(w_k - x_{k-1}), \tag{3.22}$$

with $v_0 = x_0$, as well as $\gamma_k = \frac{\alpha_k - q}{1-q}$ for all $k \geq 0$.

Note that the following relations hold for all $k \geq 1$, keeping in mind that $\alpha_k^2 = (1-\alpha_k)\alpha_{k-1}^2 + q\alpha_k$:

$$1 - \gamma_k = \frac{1-\alpha_k}{1-q} = \frac{(\mu+\kappa)(1-\alpha_k)}{\kappa}$$
$$\gamma_k = \frac{\alpha_k - q}{1-q} = \frac{\alpha_k^2 - q\alpha_k}{\alpha_k - q\alpha_k} = \frac{\alpha_{k-1}^2(1-\alpha_k)}{\alpha_k - \alpha_k^2 + (1-\alpha_k)\alpha_{k-1}^2} = \frac{\alpha_{k-1}^2}{\alpha_{k-1}^2 + \alpha_k}.$$

Then, based on the previous relations, we have

$$y_k = w_k + \beta_k(w_k - x_{k-1}) + \frac{(\mu+\kappa)(1-\alpha_k)}{\kappa}(x_k - w_k)$$
$$= w_k + \frac{\alpha_{k-1}(1-\alpha_{k-1})}{\alpha_{k-1}^2 + \alpha_k}(w_k - x_{k-1}) + (1-\gamma_k)(x_k - w_k)$$
$$= w_k + \frac{\gamma_k(1-\alpha_{k-1})}{\alpha_{k-1}}(w_k - x_{k-1}) + (1-\gamma_k)(x_k - w_k)$$
$$= \gamma_k v_k + (1-\gamma_k)x_k,$$

which is similar to the relation used in [Lin et al., 2018] when $w_k = x_k$. Then, the proof differs from [Lin et al., 2018] since we introduce the surrogate function h_k. For all x in $\mathbb{R}^p$,

$$h_k(x) \geq h_k^* + \frac{\mu+\kappa}{2}\|x - x_k^*\|^2 \quad \text{(by strong convexity, see } \mathcal{H}_1) \tag{3.23}$$

$$= h_k^* + \frac{\mu+\kappa}{2}\|x - w_k\|^2 + \underbrace{\frac{\mu+\kappa}{2}\|w_k - x_k^*\|^2 + (\mu+\kappa)\langle x - w_k, w_k - x_k^*\rangle}_{-\Delta_k(x)}. \tag{3.24}$$

Introduce now the following quantity for the convergence analysis:

$$z_{k-1} = \alpha_{k-1}x^* + (1-\alpha_{k-1})x_{k-1},$$

and consider $x = z_{k-1}$ in (3.24) while taking expectations, noting that all random variables indexed by $k-1$ are deterministic given $\mathcal{F}_{k-1}$,

$$\mathbb{E}\left[F(x_k)\right] \leq \mathbb{E}\left[h_k^*\right] + \delta_k \qquad \text{(by } \mathcal{H}_3) \tag{3.25}$$

$$\leq \mathbb{E}\left[h_k(z_{k-1})\right] - \mathbb{E}\left[\frac{\mu+\kappa}{2}\|z_{k-1} - w_k\|^2\right] + \mathbb{E}\left[\Delta_k(z_{k-1})\right] + \delta_k \tag{3.26}$$

$$\leq \mathbb{E}\left[F(z_{k-1})\right] + \mathbb{E}\left[\frac{\kappa}{2}\|z_{k-1} - y_{k-1}\|^2\right] - \mathbb{E}\left[\frac{\mu+\kappa}{2}\|z_{k-1} - w_k\|^2\right] + \mathbb{E}\left[\Delta_k(z_{k-1})\right] + \delta_k, \tag{3.27}$$

where the last inequality is due to $(\mathcal{H}_2)$.

Let us now open a parenthesis and derive a few relations that will be useful to find a Lyapunov function. To use more compact notation, define $X_k = \mathbb{E}\left[\|x^* - x_k\|^2\right]$, $V_k = \mathbb{E}\left[\|x^* - v_k\|^2\right]$ and $F_k = \mathbb{E}\left[F(x_k) - F^*\right]$, and note that

$$\begin{aligned}
\mathbb{E}\left[F(z_{k-1})\right] &\leq \alpha_{k-1}F^* + (1-\alpha_{k-1})\mathbb{E}\left[F(x_{k-1})\right] - \frac{\mu\alpha_{k-1}(1-\alpha_{k-1})}{2}X_{k-1} \\
\mathbb{E}\left[\|z_{k-1} - w_k\|^2\right] &= \alpha_{k-1}^2 V_k \\
\mathbb{E}\left[\|z_{k-1} - y_{k-1}\|^2\right] &\leq \alpha_{k-1}(\alpha_{k-1} - \gamma_{k-1})X_{k-1} + \alpha_{k-1}\gamma_{k-1}V_{k-1}.
\end{aligned} \tag{3.28}$$

The first relation is due to the convexity of F; the second one can be obtained from the definition of v_k in (3.22) after simple calculations; the last one can be obtained as in the proof of Theorem 3 in [Lin et al., 2018].

We may now close the parenthesis, come back to (3.27) and we use the relations (3.28):

$$F_k + \frac{(\mu+\kappa)\alpha_{k-1}^2}{2}V_k \leq (1-\alpha_{k-1})F_{k-1} - \frac{\mu\alpha_{k-1}(1-\alpha_{k-1})}{2}X_{k-1} + \frac{\kappa}{2}\alpha_{k-1}(\alpha_{k-1} - \gamma_{k-1})X_{k-1} + \frac{\kappa}{2}\alpha_{k-1}\gamma_{k-1}V_{k-1} + \delta_k + \mathbb{E}\left[\Delta_k(z_{k-1})\right].$$

It is then easy to see that the terms involving X_{k-1} cancel each other since $\gamma_{k-1} = \alpha_{k-1} - (\mu/\kappa)\left(1-\alpha_{k-1}\right)$.

Lyapunov function. We may finally define the Lyapunov function

$$S_k = (1-\alpha_k)F_k + \frac{\kappa\alpha_k\gamma_k}{2}V_k. \tag{3.29}$$

and we obtain

$$\frac{S_k}{1-\alpha_k} \leq S_{k-1} + \delta_k + \mathbb{E}\left[\Delta_k(z_{k-1})\right], \tag{3.30}$$

For variant Algorithm 3.3, we have $\Delta_k(z_{k-1}) = 0$ since $w_k = x_k^*$, and we obtain the following relation by unrolling the recursion:

$$S_k \leq A_k\left(S_0 + \sum_{j=1}^{k}\frac{\delta_j}{A_{j-1}}\right) \quad \text{with} \quad A_j = \prod_{i=1}^{j}(1-\alpha_i). \tag{3.31}$$

Specialization to $\mu > 0$. When $\mu > 0$, we have $\alpha_0 = \sqrt{q}$ and

$$\begin{aligned} S_0 &= (1-\sqrt{q})\left(F(x_0) - F^*\right) + \frac{\kappa\sqrt{q}(\sqrt{q}-q)}{2(1-q)}\left\|x_0 - x^*\right\|^2 \\ &= (1-\sqrt{q})\left(F(x_0) - F^*\right) + \frac{(\mu+\kappa)\sqrt{q}(\sqrt{q}-q)}{2}\left\|x_0 - x^*\right\|^2 \\ &= (1-\sqrt{q})\left(F(x_0) - F^*\right) + \frac{\mu\left(1-\sqrt{q}\right)}{2}\left\|x_0 - x^*\right\|^2 \\ &\leq 2\left(1-\sqrt{q}\right)\left(F(x_0) - F^*\right), \end{aligned} \tag{3.32}$$

by using the strong convexity inequality $F(x_0) \geq F^* + (\mu/2)\left\|x_0 - x^*\right\|^2$. Then, noting that $\mathbb{E}\left[F(x_k) - F^*\right] \leq \frac{S_k}{1-\sqrt{q}}$ and $A_k = \left(1-\sqrt{q}\right)^k$ (Lemma 3.1), we immediately obtain the first part of (3.11) from (3.31).

Specialization to $\mu = 0$. When $\mu = 0$, we have $\alpha_0 = 1$ and $S_0 = \frac{\kappa}{2}\left\|x_0 - x^*\right\|^2$. Then, according to Lemma 3.1 and (3.31), for $k \geq 1$,

$$\mathbb{E}\left[F(x_k) - F^*\right] \leq \frac{S_k}{1-\alpha_k} \leq \frac{\kappa\left\|x_0 - x^*\right\|^2}{2}A_{k-1} + \sum_{j=1}^{k}\frac{\delta_j A_{k-1}}{A_{j-1}}, \tag{3.33}$$

and we obtain the second part of (3.11) noting that $A_{k-1} \leq \frac{4}{(k+1)^2}$ and that $A_{j-1} \geq \frac{2}{(j+1)^2}$. Then, Proposition 3.1 is proven.

Proof of Proposition 3.2. When $w_k = x_k$, we need to control the quantity $\Delta_k(z_{k-1})$. Consider any scalar θ_k in $(0,1)$. Then,

$$\begin{aligned}
\Delta_k(z_{k-1}) &= -\frac{\mu+\kappa}{2}\left\|x_k - x_k^*\right\|^2 - (\mu+\kappa)\left\langle z_{k-1} - x_k, x_k - x_k^*\right\rangle \\
&= -\frac{\mu+\kappa}{2}\left\|x_k - x_k^*\right\|^2 - (\mu+\kappa)\alpha_{k-1}\left\langle x^* - v_k, x_k - x_k^*\right\rangle \\
&\le -\frac{\mu+\kappa}{2}\left\|x_k - x_k^*\right\|^2 + (\mu+\kappa)\alpha_{k-1}\|x^* - v_k\|\|x_k - x_k^*\| \\
&\le \left(\frac{1}{\theta_k} - 1\right)\frac{\mu+\kappa}{2}\left\|x_k - x_k^*\right\|^2 + \frac{\theta_k(\mu+\kappa)\alpha_{k-1}^2}{2}\left\|x^* - vk\right\|^2 \quad \text{(Young's inequality)} \\
&\le \left(\frac{1}{\theta_k} - 1\right)\left(h_k(x_k) - h_k^*\right) + \frac{\theta_k(\mu+\kappa)\alpha_{k-1}^2}{2}\left\|x^* - vk\right\|^2 \quad \text{(since } \theta_k \le 1\text{)} \\
&\le \left(\frac{1}{\theta_k} - 1\right)\left(h_k(x_k) - h_k^*\right) + \frac{\theta_k(\mu+\kappa)(\alpha_k^2 - \alpha_k q)}{2(1-\alpha_k)}\left\|x^* - vk\right\|^2 \\
&= \left(\frac{1}{\theta_k} - 1\right)\left(h_k(x_k) - h_k^*\right) + \frac{\theta_k\kappa\alpha_k\gamma_k}{2(1-\alpha_k)}\left\|x^* - vk\right\|^2.
\end{aligned}$$

Then, we take expectations and, noticing that the quadratic term involving $\|x^* - v_k\|$ is smaller than $\theta_k S_k/(1-\alpha_k)$ in expectation (from the definition of S_k in (3.29)), we obtain

$$\mathbb{E}\left[\Delta_k(z_{k-1})\right] \le \left(\frac{1}{\theta_k} - 1\right)\varepsilon_k + \frac{\theta_k S_k}{1-\alpha_k},$$

and from (3.30),

$$S_k \le \frac{(1-\alpha_k)}{(1-\theta_k)}\left(S_{k-1} + \delta_k + \left(\frac{1}{\theta_k} - 1\right)\varepsilon_k\right).$$

By unrolling the recursion, we obtain

$$S_k \le \frac{A_k}{\Theta_k}\left(S_0 + \sum_{j=1}^{k}\frac{\Theta_{j-1}}{A_{j-1}}\left(\delta_j - \varepsilon_j + \frac{\varepsilon_j}{\theta_j}\right)\right) \quad \text{with} \quad A_j = \prod_{i=1}^{j}(1-\alpha_i) \quad \text{and} \quad \Theta_j = \prod_{i=1}^{j}(1-\theta_i). \tag{3.34}$$

Specialization to $\mu > 0$. When $\mu > 0$, we have $\alpha_k = \sqrt{q}$ for all $k \ge 0$. Then, we may choose $\theta_k = \frac{\sqrt{q}}{2}$; then, $1 - \sqrt{q} \le \left(1 - \frac{\sqrt{q}}{2}\right)^2$ and $\frac{A_k}{\Theta_k} \le \left(1 - \frac{\sqrt{q}}{2}\right)^k$ for all $k \ge 0$. By using the relation (3.32), we obtain

$$\begin{aligned}
S_k &\le 2\left(1 - \frac{\sqrt{q}}{2}\right)^k (1-\sqrt{q})\left(F(x_0) - F^*\right) + 2\sum_{j=1}^{k}\left(\frac{1-\sqrt{q}}{1-\frac{\sqrt{q}}{2}}\right)^{k-j+1}\left(\delta_j - \varepsilon_j + \frac{\varepsilon_j}{\sqrt{q}}\right) \\
&\le (1-\sqrt{q})\left(2\left(1 - \frac{\sqrt{q}}{2}\right)^k \left(F(x_0) - F^*\right) + 4\sum_{j=1}^{k}\left(\frac{1-\sqrt{q}}{1-\frac{\sqrt{q}}{2}}\right)^{k-j}\left(\delta_j - \varepsilon_j + \frac{\varepsilon_j}{\sqrt{q}}\right)\right) \\
&\le (1-\sqrt{q})\left(2\left(1 - \frac{\sqrt{q}}{2}\right)^k \left(F(x_0) - F^*\right) + 4\sum_{j=1}^{k}\left(1 - \frac{\sqrt{q}}{2}\right)^{k-j}\left(\delta_j - \varepsilon_j + \frac{\varepsilon_j}{\sqrt{q}}\right)\right),
\end{aligned}$$

where the second inequality uses $\left(1-\frac{\sqrt{q}}{2}\right)^{-1} \leq 2$. Since $\left(1-\sqrt{q}\right)\mathbb{E}\left[F(x_k)-F^*\right] \leq S_k$, we obtain the first part of Proposition 3.2.

Specialization to $\mu = 0$. When $\mu = 0$, we have $\alpha_0 = 1$ and $S_0 = \frac{\kappa}{2}\|x_0 - x^*\|^2$. We may then choose $\theta_k = \frac{\gamma}{(k+1)^{1+\gamma}}$ for any γ in $(0,1]$, leading to $e^{-(1+\gamma)} \leq \Theta_k \leq 1$ for all $k \geq 0$ according to Lemma 3.2. Besides, according to the proof of Lemma 3.1, $\frac{2}{(k+2)^2} \leq A_k \leq \frac{4}{(k+2)^2}$ for all $k \geq 1$.

Then, from (3.34),

$$\begin{aligned}\mathbb{E}\left[F(x_k)-F^*\right] &\leq \frac{A_{k-1}}{\Theta_k}\frac{\kappa\|x_0-x^*\|^2}{2} + \sum_{j=1}^{k}\frac{A_{k-1}\Theta_{j-1}}{\Theta_k A_{j-1}}\left(\delta_j - \varepsilon_j + \frac{\varepsilon_j}{\gamma}(1+j)^{1+\gamma}\right)\\ &\leq \frac{2e^{1+\gamma}}{(k+1)^2}\left(\kappa\|x_0-x^*\|^2 + \sum_{j=1}^{k}(j+1)^2(\delta_j-\varepsilon_j) + \frac{(j+1)^{3+\gamma}\varepsilon_j}{\gamma}\right),\end{aligned}$$

which yields the second part of Proposition 3.2. □

3.C.2 Proof of Proposition 3.3

Assume that for $k \geq 2$, we have the relation

$$\mathbb{E}\left[h_{k-1}(x_{k-1}) - h_{k-1}^*\right] \leq \varepsilon_{k-1}. \tag{3.35}$$

Then, we want to evaluate the quality of the initial point x_{k-1} to minimize h_k.

$$\begin{aligned}h_k(x_{k-1}) - h_k^* &= h_{k-1}(x_{k-1}) + \frac{\kappa}{2}\|x_{k-1}-y_{k-1}\|^2 - \frac{\kappa}{2}\|x_{k-1}-y_{k-2}\|^2 - h_k^*\\ &= h_{k-1}(x_{k-1}) - h_{k-1}^* + h_{k-1}^* - h_k^* + \frac{\kappa}{2}\|x_{k-1}-y_{k-1}\|^2 - \frac{\kappa}{2}\|x_{k-1}-y_{k-2}\|^2\\ &= h_{k-1}(x_{k-1}) - h_{k-1}^* + h_{k-1}^* - h_k^* -\\ &\qquad \kappa(x_{k-1}-y_{k-1})^\top(y_{k-1}-y_{k-2}) - \frac{\kappa}{2}\|y_{k-1}-y_{k-2}\|^2. \end{aligned}\tag{3.36}$$

Then, we may use the fact that h_k^* can be interpreted as the Moreau-Yosida smoothing of the objective F, defined as $G(y) = \min_{x\in\mathbb{R}^p} F(x) + \frac{\kappa}{2}\|x-y\|^2$, which gives us immediately a few useful results, as noted in [Lin et al., 2019]. Indeed, we know that G is κ-smooth with $\nabla G(y_{k-1}) = \kappa(y_{k-1} - x_k^*)$ for all $k \geq 1$ and

$$\begin{aligned}h_{k-1}^* = G(y_{k-2}) &\leq G(y_{k-1}) + \nabla G(y_{k-1})^\top(y_{k-2}-y_{k-1}) + \frac{\kappa}{2}\|y_{k-1}-y_{k-2}\|^2\\ &= h_k^* + \kappa(y_{k-1}-x_k^*)^\top(y_{k-2}-y_{k-1}) + \frac{\kappa}{2}\|y_{k-1}-y_{k-2}\|^2. \end{aligned}\tag{3.37}$$

Then, combining (3.36) and (3.37),

$$\begin{aligned}
h_k(x_{k-1}) - h_k^* &\leq h_{k-1}(x_{k-1}) - h_{k-1}^* + \kappa(x_{k-1} - x_k^*)^\top (y_{k-2} - y_{k-1}). \\
&\leq h_{k-1}(x_{k-1}) - h_{k-1}^* + \kappa(x_{k-1} - x_{k-1}^*)^\top (y_{k-2} - y_{k-1}) + \\
&\qquad \kappa(x_{k-1}^* - x_k^*)^\top (y_{k-2} - y_{k-1}) \\
&\leq h_{k-1}(x_{k-1}) - h_{k-1}^* + \kappa(x_{k-1} - x_{k-1}^*)^\top (y_{k-2} - y_{k-1}) + \kappa \left\| y_{k-1} - y_{k-2} \right\|^2 \\
&\leq h_{k-1}(x_{k-1}) - h_{k-1}^* + \frac{\kappa}{2} \left\| x_{k-1} - x_{k-1}^* \right\|^2 + \frac{3\kappa}{2} \left\| y_{k-1} - y_{k-2} \right\|^2 \\
&\leq \frac{3}{2}(h_{k-1}(x_{k-1}) - h_{k-1}^*) + \frac{3\kappa}{2} \left\| y_{k-1} - y_{k-2} \right\|^2,
\end{aligned}$$

where the third inequality uses the non-expansiveness of the proximal operator; the fourth inequality uses the inequality $a^\top b \leq \frac{\|a\|^2}{2} + \frac{\|b\|^2}{2}$ for vectors a, b, and the last inequality uses the strong convexity of h_{k-1}. Then, we may use the same upper-bound on $\|y_{k-1} - y_{k-2}\|$ as [Lin et al., 2018, Proposition 12], namely

$$\left\| y_{k-1} - y_{k-2} \right\|^2 \leq 36 \max \left\{ \left\| x_{k-1} - x^* \right\|^2, \left\| x_{k-2} - x^* \right\|^2, \left\| x_{k-3} - x^* \right\|^2 \right\},$$

where we define $x_{-1} = x_0$ if $k = 2$.

3.C.3 Proof of Proposition 3.4

The proof is similar to the derivation described in Section 3.B.3.

Inner-loop complexity. With the choice of δ_k, we have that $\xi_{k-1} = \mathcal{O}\left(\delta_{k-1}/\sqrt{q}\right)$. Besides, since we enforce $\mathbb{E}\left[H_k(x_k) - H_k^*\right] \leq \delta_k$ for all $k \geq 0$, the result of Proposition 3.3 can be applied and the discussion following the proposition still applies, such that the complexity for computing x_k is indeed $\tilde{O}(1/(\eta_k \tau))$.

Outer-loop complexity. Then, according to Proposition 3.1, it is easy to show that $\mathbb{E}\left[F(x_k) - F^*\right] \leq \mathcal{O}\left(\left(1 - \sqrt{q}/2\right)^k \left(F(x_0) - F^*\right)\right)/\sqrt{q}$ and thus it suffices to choose

$$K = \mathcal{O}\left(\frac{1}{\sqrt{q}} \log\left(\frac{F(x_0) - F^*}{\sqrt{q}\varepsilon}\right)\right)$$

iterations to guarantee $\mathbb{E}\left[F(x_K) - F^*\right] \leq \varepsilon$.

Global complexity. We use the exact same derivations as in Section 3.B.3 except that we use the fact that $\varepsilon = \mathcal{O}\left(\varepsilon_K/\sqrt{q}\right) = \mathcal{O}\left(\left(1 - \sqrt{q}/3\right)^K \left(F(x_0) - F^*\right)/\sqrt{q}\right)$ instead of $\varepsilon = \mathcal{O}\left(\varepsilon_K/q\right)$, which gives us the desired complexity.

3.D Methods $\mathcal{M}$ with Duality Gaps Based on Strongly-Convex Lower Bounds

In this section, we summarize a few results from Chapter 2 for convenience and introduce minor modifications to guarantee the condition (3.19). For solving a stochastic composite objectives such as (3.1), where F is μ-strongly convex, consider an algorithm $\mathcal{M}$ performing the following classical updates

$$z_t \leftarrow \mathrm{Prox}_{\eta\psi}\left[z_{k-1} - \eta g_t\right] \quad \text{with} \quad \mathbb{E}\left[g_t|\mathcal{F}_{k-1}\right] = \nabla f(z_{k-1}),$$

where $\eta \leq 1/L$, and the variance of g_t is upper-bounded by σ_t^2. Inspired by estimate sequences from [Nesterov, 2014], in Chapter 2, we have built recursively a μ-strongly convex quadratic function d_t of the form

$$d_t(z) = d_t^* + \frac{\mu}{2}\left\|z_t - z\right\|^2.$$

From the proof of Proposition 2.1 in Chapter 2, we then have

$$\mathbb{E}\left[d_t^*\right] \geq (1-\eta\mu)\mathbb{E}\left[d_{k-1}^*\right] + \eta\mu\mathbb{E}\left[F(z_t)\right] - \eta^2\mu\sigma_t^2,$$

which leads to

$$F^* - \mathbb{E}\left[d_t^*\right] + \eta\mu(\mathbb{E}\left[F(z_t)\right] - F^*) \leq (1-\eta\mu)\mathbb{E}\left[F^* - d_{k-1}^*\right] + \eta^2\mu\sigma_t^2,$$

which is a minor modification of Proposition 2.1 in Chapter 2 that is better suited to our purpose.

With constant variance. Assume now that $\sigma_t = \sigma$ for all $k \geq 1$. Following the iterate averaging procedure used in Theorem 2.1, which produces an iterate $\hat{z}_t$, we obtain

$$\mathbb{E}\left[F(\hat{z}_t) - d_t^*\right] \leq (1-\eta\mu)^t\left(F(z_0) - d_0^*\right) + \eta\sigma^2, \tag{3.38}$$

where d_0^* can be freely specified for the analysis: it is not used by the algorithm, but it influences d_t^* through the relation $\mathbb{E}\left[d_t(z)\right] \leq \Gamma_t d_0(z) + (1-\Gamma_t)\mathbb{E}\left[F(z)\right]$ with $\Gamma_t = (1-\mu\eta)^t$, see relation (2.10) in Chapter 2. In contrast, Theorem 2.1 would give here

$$\mathbb{E}\left[F(\hat{z}_t) - F^* + d_t(z^*) - d_t^*\right] \leq 2\,(1-\eta\mu)^t\left(F(z_0) - F^*\right) + \eta\sigma^2, \tag{3.39}$$

where z^* is a minimizer of F, which is sufficient to guarantee (3.3) given that $d_t(z^*) \geq d_t^*$.

Application to the minimization of H_k. Let us now consider applying the method to an auxiliary function H_k from (3.19) instead of F, with initialization x_{k-1}. After running T iterations, define h_k to be the corresponding function d_T defined above and $x_k = \hat{z}_T$. H_k is $(\mu+\kappa)$-strongly convex and thus h_k is also $(\mu+\kappa)$-strongly convex such that $(\mathcal{H}_1)$ is satisfied. Let us now check possible choices for d_0^* to ensure $(\mathcal{H}_2)$. For

$z = \alpha_{k-1}x^* + (1-\alpha_{k-1})x_{k-1}$, we have $\mathbb{E}\left[d_T(z_{k-1})\right] \leq \Gamma_T d_0(z_{k-1}) + (1-\Gamma_T)H_k(z_{k-1})$ such that we simply need to choose d_0^* such that $\mathbb{E}\left[d_0(z_{k-1})\right] \leq \mathbb{E}\left[H_k(z_{k-1})\right]$. Then, choose

$$d_0^* = H_k^* - F(x_{k-1}) + F^*, \tag{3.40}$$

and

$$\begin{aligned} d_0(z_{k-1}) &= d_0^* + \frac{\mu+\kappa}{2}\left\|x_{k-1} - z_{k-1}\right\|^2 = d_0^* + \frac{(\mu+\kappa)\alpha_{k-1}^2}{2}\left\|x_{k-1} - x^*\right\|^2 \\ &= d_0^* + \frac{\mu}{2}\left\|x_{k-1} - x^*\right\|^2 \leq d_0^* + F(x_{k-1}) - F^* \\ &= H_k^* \leq H_k(z), \end{aligned}$$

such that $(\mathcal{H}_2)$ is satisfied, and finally (3.38) becomes

$$\mathbb{E}\left[H_k(x_k) - h_k^*\right] \leq (1-\eta(\mu+\kappa))^T \left(H_k(x_{k-1}) - H_k^* + F(x_{k-1}) - F^*\right) + \eta\sigma^2,$$

which matches (3.19).

Variance-reduction methods. In Chapter 2, gradient estimators g_t with variance reduction are studied, leading to variants of SAGA [Defazio et al., 2014a], MISO [Mairal, 2015], and SVRG [Xiao and Zhang, 2014], which can deal with the stochastic finite-sum problem presented in Section 4.1. Then, the variance of σ_t^2 decreases over the iterations, see Proposition 2.2 Chapter 2, and their convergence bounds have the variance established in terms of $\tilde{\sigma}^2 \ll \sigma^2$ from (2.18).

Let us then consider again the guarantees of the method obtained when minimizing F with $\mu/L \leq 1/5n$. From Corollary 2.3, we have

$$\mathbb{E}\left[F(\hat{z}_t) - F^* + d_t(z^*) - d_t^*\right] \leq 8\left(1-\mu\eta\right)^t \left(F(x_0) - F^*\right) + 18\eta\tilde{\sigma}^2,$$

and (3.3) is satisfied. Consider now two cases at iteration T:

— if $\mathbb{E}\left[d_T(z^*)\right] \geq F^*$, then we have $\mathbb{E}\left[F(\hat{z}_T) - d_T^*\right] \leq 8\left(1-\mu\eta\right)^T \left(F(x_0) - F^*\right) + 18\eta\tilde{\sigma}^2$.

— otherwise, it is easy to modify Theorem 2 and Corollary 2.3 to obtain

$$\mathbb{E}\left[F(\hat{z}_T) - d_T^*\right] \leq \left(1-\mu\eta\right)^T \left(2(F(x_0) - F^*) + 6(F^* - d_0^*)\right) + 18\eta\tilde{\sigma}^2.$$

Application to the minimization of H_k. Consider now applying the method for minimizing H_k, with the same choice of d_0^* as (3.40), which ensures $(\mathcal{H}_2)$, and same definitions as above for x_k and h_k. Note that the conditions on μ and L above are satisfied when $\kappa = (L/5n) - \mu$ under the condition $L/5n \geq \mu$. Then, we have from the previous results, after replacing F by H_k making the right substitutions

$$\mathbb{E}\left[H_k(x_k) - h_k^*\right] \leq \left(1-(\mu+\kappa)\eta\right)^T \left(8(H_k(x_{k-1}) - H_k^*) + 6(F(x_{k-1}) - F^*)\right) + 18\eta\sigma^2,$$

and (3.19) is satisfied with $\tilde{\sigma}^2$ instead of σ^2. Therefore, we refer to $\tilde{\sigma}^2$ in Table 3.1 for variance-reduced methods.

Other schemes. Whereas we have presented approaches were d_t is quadratic, in Chapter 2 we also studied another class of algorithms where d_t is composite (see Section 2.2.2). The results we present in this chapter can be extended to such cases, but for simplicity, we have focused on quadratic surrogates.

Chapter 4

Sparse Recovery with Reduced-Variance Algorithms

One of the main contributions of Chapters 2 and 3 was development of various algorithms, which are robust to stochastic noise. In particular, several approaches were based on the variance reduction technique which utilizes finite-sum structure of optimization problems. Another example of a structure, which allows to mitigate the impact of noise that comes from an inexact oracle, is *sparsity*. From high-level perspective, an optimization problem is called sparse if the signal to be recovered has a minor fraction of significant components. Therefore, for sparse problems we may potentially reduce the noise effects coming from insignificant components and decrease the overall variance.

As was mentioned in Section 1.6, this field is well studied with various optimization methods of both SA and SAA types being developed. The latter include iterative thresholding techniques, particularly widely studied Lasso estimator (1.33) and Dantzig Selector. Most of them inherit the computational burden typical to SAA approaches. At the same time, some optimization algorithms of SA type with a lower per-iteration cost achieved a slow rate of asymptotical error $\mathcal{O}\left(\sigma\sqrt{s\log(d)/N}\right)$, which was then eventually improved to $\mathcal{O}\left(s\sigma^2\log(d)/(\mu N)\right)$ by a multi-stage procedure.

In this chapter, our objective is to achieve the convergence rate $\mathcal{O}\left(s\sigma^2\log(d)/(\mu N)\right)$ for a larger class of noise models. At the same time, we establish linear convergence of the initial error—the task which was not done in the papers on multi-stage procedures overviewed in Section 1.6.

To address these tasks, we develop a procedure called SMD-SR, which is based on the SMD algorithm. Roughly speaking, the SMD-SR procedure is a multi-stage hard thresholding stochastic approximation algorithm where the gradient step (1.34) is substituted with a full launch of the SMD algorithm. The main distinction of our procedure from the other multi-stage routines is the use of hard thresholding operator between stages, while we do not use projections on constraint sets and do not update the minimized objective, like was was done, for instance, in [Agarwal et al., 2012b, Steinhardt et al., 2014,

Sedghi et al., 2014].

Though we solve sparse stochastic optimization problems of the general form (1.8), our main focus in this chapter is on signal recovery in sparse linear regression (1.29) and low-rank matrix recovery (1.30).

This chapter is based on the following publication:

- A. Juditsky, A. Kulunchakov and H. Tsyntseus. Reduced variance algorithms of stochastic approximation for sparse recovery. *arXiv:2006.06365*, 2020

4.1 Introduction

We consider stochastic optimization problems of the following form

$$\min_{x\in X}\left\{f(x)=\mathbb{E}\left[\tilde{f}(x,\omega)\right]\right\} \tag{4.1}$$

where X is a given convex subset of the Euclidean space E, the function $\tilde{f}: X\times\Omega\to\mathbb{R}$ is a mapping, which is convex, finite and differentiable for all $x\in X$, and $\mathbb{E}$ stands for the expectation with respect to an unknown distribution of $\omega\in\Omega$. The optimal value of f is denoted as $f^*=f(x^*)$ where the optimum x^* is unique. In what follows, we assume that $E=\mathbb{R}^n$, unless stated otherwise. Note that, unlike previous chapters, we define n as the problem dimension.

There are several distinctions between (4.1) and the problem (1.8), which was solved in Chapters 2 and 3. First, (4.1) has a non-composite formulation. Second, we assume that the optimal solution x^* is *sparse*. The general notion of sparsity structure will be given in Section 4.2.1, and comprises "usual" sparsity, group sparsity, and low rank matrix structures as basic examples. For simplicity, in the introduction we stick to the "usual" s-sparsity of vectors, being the property of having at most s non-zero components. Finally, while the L-smoothness of $f(x)$ in Chapters 2 and 3 was defined in the Euclidean norm, we consider it with respect to a general norm $\|\cdot\|$. Specifically, we assume that the following Lipschitz property is satisfied for $\tilde{f}(\cdot,\omega)$

$$\left\|\nabla\tilde{f}(x,\omega)-\nabla\tilde{f}(x',\omega)\right\|_*\le\mathcal{L}(\omega)\|x-x'\|$$

where $\mathbb{E}[\mathcal{L}(\omega)]\le\nu<\infty$ for some positive ν and $\|\cdot\|_*$ is the conjugate norm $\|z\|_*=\max_x\left\{z^Tx:\ \|x\|\le1\right\}$. We also suppose that the gradient $\nabla f(x)$ of the expected function is Lipschitz-continuous on X with the constant $\mathcal{L}$ with respect to the same norm. The norm $\|\cdot\|$ will be usually chosen as $\|\cdot\|=\|\cdot\|_1$, so that $\|\cdot\|_*=\|\cdot\|_\infty$. Finally, we suppose that $f(x)$ satisfies a *quadratic growth condition* on X with respect to the Euclidean norm $\|\cdot\|_2$, being

$$f(x)-f^*\ge(\mu/2)\|x-x^*\|_2^2,\quad\forall\,x\in X. \tag{4.2}$$

Apparently, a uniformly strongly convex function satisfies (4.2).

Sparse linear regression Sparse optimization problems (4.1) have received a lot of attention in the literature as we have already seen in Section 1.6. In particular, they have been studied in relation with the sparse linear regression problem (1.29), which can be cast as minimization of either deterministic (1.32) or stochastic (1.31) objective function. In this chapter, we focus on SA approaches which minimize the latter, while addressing the reader to Section 1.6.1 for the overview of SAA techniques. which minimize (1.32). For convenience, we repeat here the statement of sparse linear regression problems

$$\min_{x\in X}\left\{f_{\mathrm{SA}}(x)=\frac{1}{2}\mathbb{E}\left\|\eta-\phi^\top x\right\|_2^2\right\}, \tag{4.3}$$

$$\text{where}\quad \eta=\langle\phi,x^*\rangle+\sigma\xi \tag{4.4}$$

with i.i.d. random variables $\phi\in\mathbb{R}^n$ and $\xi\in\mathbb{R}$. The parameter σ stands for the scale of noise, while the random variable ξ is assumed to have unit variance. The optimum vector x^* is supposed to be s-sparse, and the minimization is conducted over the set X of s-sparse vectors. For the general case of (4.1), the variance σ^2 would denote $\mathbb{E}\left\|\nabla\tilde{f}(x^*,\omega)-\nabla f(x^*)\right\|_*^2$ being the discrepancy measured only at the optimum point x^*.[1] Apparently, this definition is consistent with (4.4).

Related work In Section 1.6, we have overviewed the literature on solving (4.3). Let us briefly recall the key points. The case of sparse regression was considered in [Srebro et al., 2010, Shalev-Shwartz and Tewari, 2011], establishing a slow asymptotic convergence rate $\mathcal{O}\left(\sqrt{s\log(d)/N}\right)$, while not obtaining the linear convergence (1.35) of the initial error specific to hard thresholding techniques of SAA type in the strongly convex setting. However, this linear convergence is achieved by several multi-stage optimization procedures. For example, in [Agarwal et al., 2012b] authors develop a multi-stage procedure, called RADAR, which applies to (4.3). In the case of uniformly bounded regressors $\|\phi\|_\infty\le B$, RADAR enjoys the following convergence rate

$$\mathbb{E}\left[\|x_N-x^*\|_2^2\right]\le\mathcal{O}\left(\frac{sB^2\sigma^2\log n}{\mu^2N}\right). \tag{4.5}$$

Notice that this result appears only in Section 3.2 of [Agarwal et al., 2012b], while the other results of this paper appeal to the other definition of σ, being

$$\forall x\in X\quad \mathbb{E}\left\|\tilde{f}(x,\omega)-\nabla f(x)\right\|_\infty^2\le\hat{\sigma}^2\gg\sigma^2. \tag{4.6}$$

This is a more standard definition of the noise variance in the literature on stochastic approximation. However, the value of $\hat{\sigma}$ is generally much larger than of σ. Indeed, for example, in the setting of least-squares regression (4.3) with $\sigma=0$, $\|\cdot\|=\|\cdot\|_1$ and $\phi\sim\mathcal{N}(0,\mathbb{I})$, we have

$$\begin{aligned}\sigma&\ge\mathbb{E}\left\|\nabla\tilde{f}(x,\omega)-\nabla f(x)\right\|_\infty\sim\mathbb{E}\left\|x-x^*\right\|_1\left\|\phi^\top\phi-\mathbb{I}\right\|_\infty\\&\sim\left\|x-x^*\right\|_1\mathbb{E}\left\|\phi^\top\phi-\mathbb{I}\right\|_\infty,\end{aligned} \tag{4.7}$$

1. Further, we give a precise definition of σ^2 for the general case in Assumption 4.9.

where $\|x - x^*\|_1$ is typically a large factor. In this chapter, we focus on the large-scale setting where $n \gg N$. Therefore, we are looking for convergence bounds on the recovery error that are independent (or logarithmic at most) in the problem dimension n. For this reason, in the view of (4.7) it is particularly important to refer in (4.5) exactly to σ^2, not to $\hat{\sigma}^2$. We do not compare our results with other multi-stage procedures overviewed in Section 1.6, because they express their convergence bounds in terms of $\hat{\sigma}^2$. This argument rules out the comparison with algorithms of the standard "Euclidean" stochastic approximation as well.

Our approach Now, let us briefly describe the procedure—Stochastic Mirror Descent for Sparse Recovery (SMD-SR)—developed in this chapter for the general stochastic optimization problem (4.1) and, in particular, for sparse linear regression (4.3). For now, let us assume that the regressors ϕ_i are almost surely bounded $\|\phi_i\|_\infty = \mathcal{O}(1)$ and the covariance matrix $\Sigma = \mathbb{E}\left[\phi\phi^T\right]$ satisfies $\Sigma \succeq \mu I$. We also suppose that that we are given $R < \infty$, such that the initialization $x_0 \in \mathbb{R}^n$ satisfies $\mathbb{E}\|x_0 - x^*\|_1^2 \leq R^2$.

The multi-stage SMD-SR procedure is based on the stochastic mirror descent briefly introduced in Section 1.4.3. The procedure roughly follows the steps below

- The SMD-SR algorithm operates in stages divided into two groups—phases—named *preliminary* and *asymptotic*. Each stage is basically a launch of the SMD algorithm, after which we "sparsify" the resulted approximate solution, so that the next launch is initialized at it. The sparsification step is done by zeroing out all but s entries of largest amplitudes of the approximate solution.
- At each stage of the preliminary phase we perform a fixed number $m_0 = \widetilde{\mathcal{O}}(s \log n/\mu)$ of iterations of SMD so that the expected quadratic error $\mathbb{E}\|\widehat{y}_N - x^*\|_2^2$ decreases linearly as the total iteration count N grows. The coefficient of this linear decrease is proportional to $\frac{\mu}{s\nu \log n}$. When the expected quadratic error becomes $\mathcal{O}(\sigma^2 s/\mu)$, we pass to the asymptotic phase of the method.
- During the asymptotic phase, the number of iterations per stage m_k grows exponentially with the stage index k. The resulted expected quadratic error decreases as $\mathcal{O}\left(\frac{\sigma^2 s \log n}{\mu^2 N}\right)$ which corresponds to (4.5).

Now, let us explicitly present the list of main contributions of this chapter. Once again, the notation $\widetilde{\mathcal{O}}(1)$ is essentially $\mathcal{O}(1)$ that probably hides a logarithmic factor on n.

4.1.1 Contributions of Chapter 4

— The main contribution of this chapter is development of the SMD-SR procedure with the following convergence rate

$$\mathbb{E}\left[\|x_N - x^*\|_2^2\right] \leq \left(1 - \widetilde{\mathcal{O}}(1)\frac{\mu}{s\nu}\right)^N \frac{\|x_0 - x^*\|_1^2}{2s} + \widetilde{\mathcal{O}}(1)\frac{s\sigma^2}{\mu^2 N}. \tag{4.8}$$

In particular, we explicitly establish that the algorithm converges linearly during the preliminary phase eliminating quickly the initial error, when the variance part is small. This convergence is similar to the deterministic gradient descent algorithm,

when "full gradient observation" $\nabla f(x)$ is available. On the other hand, in the asymptotic regime, SMD-SR attains the rate $\widetilde{\mathcal{O}}\left(1/N\right)$ which is equivalent to the best known rates in this setting (4.5).

— Another important notion concerns the admissible values of s. While the asymptotic rate (4.5) of RADAR is optimal, it requires to perform at least $s^2\log[n]/\mu^2$ iterations of the dual averaging algorithm per stage. Consequently, it can be used only if the number of nonvanishing entries in the optimal solution x^* does not exceed $\mathcal{O}\left(\mu\sqrt{N/\log n}\right)$. However, the corresponding limit is $\mathcal{O}\left(N\mu/\log[n]\right)$ for Lasso [Raskutti et al., 2010] and iterative thresholding procedures [Barber and Ha, 2018, Foygel Barber and Liu, 2019]). In this chapter, we relax the condition on admissible values of sparsity from $s \leq \widetilde{\mathcal{O}}\left(\mu\sqrt{N/\log n}\right)$ of [Agarwal et al., 2012b] to become $s \leq \widetilde{\mathcal{O}}\left(N\mu/\log[n]\right)$.

— We establish the optimal asymptotic rate $\widetilde{\mathcal{O}}\left(s\sigma^2/\mu^2 N\right)$ for several different noise models, not only for regressors ϕ bounded in the ℓ_∞-norm. For example, these models include sub-Gaussian, Rademacher, multivariate Student distributions and scale mixtures. This amounts to derivation of the optimal asymptotic rate under the model assumptions which are close to the weakest known today [Foygel Barber and Liu, 2019, Raskutti et al., 2010].

— We show how one can straightforwardly enhance reliability of the corresponding solutions by using Median-of-Means like techniques [Nemirovsky and Yudin, 1983, Minsker, 2015]. Although the convergence bounds obtained for the expected risks do allow only for Chebyshev-type bounds for risks, their confidence can nonetheless be easily improved by applying an adapted version of median-of-means estimate.

The rest of the chapter is organized as follows. We define the key assumptions and introduce a general notion of sparsity in Section 4.2.1. The stochastic mirror descent is analyzed in Section 4.2.2. The precise scheme and the analysis of the SMD-SR procedure is presented in Section 4.3. Next, in Section 4.3.1 we show how sub-Gaussian confidence bounds for the error of approximate solutions can be obtained using an adopted analog of Median-of-Means approach. In Section 4.4 we discuss the properties of the method and conditions in which it leads to a "small error" solution when applied to sparse linear regression and low rank linear matrix recovery problems. Finally, Section 2.5 contains various experiments demonstrating the effectiveness of the proposed approach.

Notation In what follows, we use a generic notation C for an absolute constant; notation $a \lesssim b$ means that the ratio a/b is bounded by an absolute constant, the norm $\|\cdot\|$ is not the Euclidean norm by default anymore, and we explicitly distinguish ℓ_2-norm from other ℓ_p norms. For $Q \in \mathbb{R}^{p\times q}$ we denote $\|Q\|_\infty = \max_{ij}|[Q]_{ij}|$ and for symmetric positive-definite $Q \in \mathbb{R}^{n\times n}$ with $x \in \mathbb{R}^n$ we denote $\|x\|_Q = \sqrt{x^TQx}$. Besides, $\lfloor a \rfloor$ stands for the smallest integer greater or equal to a, and $\lceil a \rceil$ stands for the smallest integer strictly greater than a.

4.2 Prerequisites

4.2.1 Assumptions and sparsity

We start with formulating the key assumption about the structure of the gradient estimates, that will allow us to perform the aforementioned decrease in the variance from $\hat{\sigma}^2$ of (4.6) to σ^2 of (4.4).

Assumption 4.1. *Let the function $\tilde{f}(\cdot,\omega)$ be continuously differentiable*[2] *on X for almost all $\omega\in\Omega$, and have the following Lipschitz property*

$$\left\|\nabla\tilde{f}(x,\omega)-\nabla\tilde{f}(x',\omega)\right\|_*\leq\mathcal{L}(\omega)\left\|x-x'\right\|$$

with $\mathbb{E}\left[\mathcal{L}(\omega)\right]\leq\nu<\infty$ for some positive ν. Denote $\zeta(x,\omega)\triangleq\nabla\tilde{f}(x,\omega)-\nabla f(x)$ and

$$\varsigma^2(x)=\mathbb{E}\left[\left\|\nabla\tilde{f}(x,\omega)-\nabla f(x)\right\|_*^2\right].$$

Assume that there are constants $1\leq\varkappa,\varkappa'<\infty$ such that the following bound holds:

$$\varsigma^2(x)\leq\varkappa\nu\underbrace{\left[f(x)-f^*-\langle\nabla f(x^*),x-x^*\rangle\right]}_{=:V_f(x^*,x)}+\varkappa'\underbrace{\mathbb{E}\left[\left\|\zeta(x^*,\omega)\right\|_*^2\right]}_{=:\varsigma_*^2}. \tag{4.9}$$

Essentially, this assumption is necessary to decompose the discrepancy $\varsigma^2(x)$ to two parts. The first part $V_f(x^*,x)$ is then accumulated properly by our algorithm and decrease with linear speed, and ς_*^2 is the dimension-free variance that will appear in the final convergence bounds. Several examples of models, which satisfy this assumption, will be given further in Section 4.4. For now, let us note that the relation (4.9) is rather characteristic to the case of smooth stochastic observation. Indeed, let us consider the situation where the Lipschitz constant $\mathcal{L}(\omega)$ is *a.s. bounded*, i.e. $\mathcal{L}(\omega)\leq\nu$. Using a reasoning similar to the derivation of relation (2.47) in Section 2.D.1, we obtain in this case

$$\varsigma^2(x)\leq 16\nu\left[f(x)-f^*-\langle\nabla f(x^*),x-x^*\rangle\right]+2\varsigma_*^2.$$

The same strong assumption of uniform boundness $\mathcal{L}(\omega)\leq\nu$ was used by [Bietti and Mairal, 2017] and in Proposition 2.2 of our Chapter 2.

Sparsity structure. In what follows we assume to be given a *sparsity structure* [Juditsky et al., 2014] on E, that is a family $\mathcal{P}$ of projector mappings $P=P^2$ on E with associated nonnegative weights $\pi(P)$. For a nonnegative real s we set

$$\mathcal{P}_s=\left\{P\in\mathcal{P}:\pi(P)\leq s\right\}.$$

Given $s\geq 0$ we call $x\in E$ *s-sparse* if there exists $P\in\mathcal{P}_s$ such that $Px=x$. We will make the following assumption.

2. In what follows $\nabla\tilde{f}(\cdot,\omega)$ replaces notation $\nabla_x\tilde{f}(\cdot,\omega)$ for the gradient of $\tilde{f}$ w.r.t. the first argument.

Assumption 4.2. *Given $x \in X$, assume that we can efficiently compute a "sparse approximation" of x being an optimal solution $x_s := \mathrm{sparse}(x)$ to the optimization problem*

$$\min \|x - z\|_2 \quad \text{over } s\text{-sparse } z \in X \tag{4.10}$$

where $\|\cdot\|_2$ is the Euclidean norm $\|z\|_2 = \langle z, z\rangle^{1/2}$. Furthermore, we assume that for any s-sparse vector $z \in E$ the norm $\|\cdot\|$ satisfies $\|z\| \le \sqrt{s}\,\|z\|_2$.

In what follows we refer to x_s as *s-sparsification of x*. We are mainly interested in the following standard examples:

1. **"Vanilla" sparsity:** in this case $E = \mathbb{R}^n$ with the standard inner product, $\mathcal{P}$ is comprised of projectors on all coordinate subspaces of $\mathbb{R}^n$, $\pi(P) = \mathrm{rank}(P)$, and $\|\cdot\| = \|\cdot\|_1$.
   Assumption 4.2 clearly holds, for instance, when X is orthosymmetric, e.g., a ball of ℓ_p-norm on $\mathbb{R}^n$, $1 \le p \le \infty$.
2. **Group sparsity:** $E = \mathbb{R}^n$, and we partition the set $\{1, \dots, n\}$ of indices into K non-overlapping subsets $I_1, ..., I_K$, so that to every $x \in \mathbb{R}^n$ we associate blocks x^k with corresponding indices in I_k, $k = 1, \dots, K$. Now $\mathcal{P}$ is comprised of projectors $P = P_I$ onto subspaces $E_I = \{\left[x^1, ..., x^K\right] \in \mathbb{R}^n : x^k = 0\, \forall k \notin I\}$ associated with subsets I of the index set $\{1, ..., K\}$. We set $\pi(P_I) = \mathrm{card} I$, and define $\|x\| = \sum_{k=1}^K \|x_k\|_2$—*block ℓ_1/ℓ_2-norm.*
   Same as above, Assumption 4.2 holds in this case when X is "block-symmetric," for instance, is a ball of the block norm $\|\cdot\|$.
3. **Low rank sparsity structure:** in this example $E = \mathbb{R}^{p\times q}$ with, for the sake of definiteness, $p \ge q$, and the Frobenius inner product. Here $\mathcal{P}$ is the set of mappings $P(x) = P_\ell x P_r$ where P_ℓ and P_r are, respectively, $q \times q$ and $p \times p$ orthoprojectors, and $\|\cdot\|$ is the nuclear norm $\|x\| = \sum_{i=1}^q \sigma_i(x)$ where $\sigma_1(x) \ge \sigma_2(x) \ge ... \ge \sigma_q(x)$ are singular values of x.
   In this case, Assumption 4.2 holds due to the Eckart–Young approximation theorem, it suffices that X is a ball of a Schatten norm $\|x\|_r = \left(\sum_{i=1}^q \sigma_i^r(x)\right)^{1/r}$, $1 \le r \le \infty$.

In what follows we assume that the optimal solution x^* to the problem (4.1) is s-sparse. As the true value of s is always unknown, we denote an upper bound $\bar{s} \ge s$ for the signal sparsity. Our goal is to build approximate solutions $\widehat{x}_N$ to the problem (4.1) utilizing N queries to the stochastic oracle. We quantify the accuracy of an approximate solution $\widehat{x}$ using the following risk measures:

— **Recovery risks:** first, the maximal over $x^* \in X$ expected squared error

$$\mathrm{Risk}_{|\cdot|}(\widehat{x}|X) = \sup_{x^*\in X} \mathbb{E}\left[|\widehat{x} - x^*|^2\right]$$

where $|\cdot|$ stands for $\|\cdot\|_2$- or $\|\cdot\|$-norm, and, second, the *ε-risks* of recovery, being essentially the smallest maximal over $x^* \in X$ radius of $(1-\varepsilon)$-confidence ball of norm $|\cdot|$ centered at $\widehat{x}$:

$$\mathrm{Risk}_{|\cdot|,\varepsilon}(\widehat{x}|X) = \inf\left\{r : \sup_{x^*\in X} \mathrm{Prob}|\widehat{x} - x^*| \ge r \le \varepsilon\right\}$$

— **Prediction risks:** first, the maximal over $x^* \in X$ expected sub-optimality

$$\mathrm{Risk}_f(\widehat{x}|X) = \sup_{x^* \in X} \mathbb{E}\left[f(\widehat{x})\right] - f^*,$$

of $\widehat{x}$, and, second, the smallest maximal over $x^* \in X$ $(1-\varepsilon)$-confidence interval

$$\mathrm{Risk}_{f,\varepsilon}(\widehat{x}|X) = \inf\left(r : \sup_{x^* \in X} \mathrm{Prob}\left(f(\widehat{x}) - f^* \geq r\right) \leq \varepsilon\right).$$

4.2.2 Stochastic Mirror Descent algorithm

Let $\vartheta : E \to \mathbb{R}$ be a distance-generating function from Section 1.4.3, such that the problem (1.24) is easy (for example, admits a closed form solution or may be solved by a simple linear search). From now on, w.l.o.g. we assume that $\vartheta(x) \geq \vartheta(0) = 0$. We say that Θ is the constant of quadratic growth of $\vartheta(\cdot)$ if

$$\forall x \in E \quad \vartheta(x) \leq \Theta\|x\|^2.$$

We also utilize associated Bregman divergence

$$V_{x_0}(x,z) = \vartheta(z - x_0) - \vartheta(x - x_0) - \left\langle \nabla\vartheta(x - x_0), z - x\right\rangle, \quad \forall\, z, x, x_0 \in X,$$

which is a slightly modified version of (1.22). The proximal operator (1.24) is naturally associated with $V_{x_0}(x,z)$, and we note that it is different from the proximal operator (1.9) used in Chapters 2 and 3. Here we denote this proximal operator for $x, x_0 \in X$, $u \in E$, and $\beta > 0$ as follows

$$\begin{aligned}\mathrm{Prox}_\beta(u, x; x_0) &\triangleq \operatorname*{argmin}_{z \in X}\left\{\langle u, z\rangle + \beta V_{x_0}(x,z)\right\} \\ &= \operatorname*{argmin}_{z \in X}\left\{\langle u - \beta\nabla\vartheta(x - x_0), z\rangle + \beta\vartheta(z - x_0)\right\}. \end{aligned} \tag{4.11}$$

Then, for $i = 1, 2, \ldots$, we consider the following stochastic mirror descent recursion

$$x_i = \mathrm{Prox}_{\beta_{i-1}}\left(\nabla\tilde{f}(x_{i-1}, \omega_i), x_{i-1}; x_0\right), \quad x_0 \in X, \tag{4.12}$$

Here $\beta_i > 0$ is a step size parameter which is defined later, and $\omega_1, \omega_2, \ldots$ are independent identically distributed (i.i.d.) realizations of the random variable ω from (4.1).

The approximate solution to the problem (4.1) after N iterations is defined as the following weighted average

$$\widehat{x}_N = \left[\sum_{i=1}^N \beta_{i-1}^{-1}\right]^{-1} \sum_{i=1}^N \beta_{i-1}^{-1} x_i. \tag{4.13}$$

The next result describes some useful properties of the recursion (4.12).

Proposition 4.1. *Suppose that the SMD algorithm (4.12) is applied to the problem (4.1). We assume that Assumption 4.3 holds and that initial condition $x_0 \in X$ is independent of $(\omega)_{i \geq 1}$ and such that $\mathbb{E}\left[\|x_0 - x^*\|^2\right] \leq R^2$. Choose the constant step size*

$$\beta_i \equiv \beta \geq 2\varkappa\nu, \;\; i = 1, 2, ..., m.$$

Then the approximate solution $\widehat{x}_m = \frac{1}{m}\sum_{i=1}^m x_i$ after m steps of the algorithm satisfies

$$\mathbb{E}\left[f(\widehat{x}_m)\right] - f^* \leq \frac{2R^2}{m}\left(\Theta\beta + \frac{\varkappa\nu^2}{2\beta}\right) + \frac{2\varkappa'\varsigma_*^2}{\beta}. \tag{4.14}$$

4.3 Multistage SMD algorithm

We are using the stochastic mirror descent algorithm as the building block in the iterative approach of [Juditsky and Nemirovski, 2011a, Juditsky and Nesterov, 2014] to improve its accuracy bounds. The proposed Stochastic Mirror Descent algorithm for Sparse Recovery (SMD-SR) works in stages, which are split into two groups—phases—corresponding to two essentially different regimes of the method. We refer to the first phase as preliminary, and refer to the second as asymptotic. Each stage consists of a launch of the SMD algorithm initialized at some proper point and subsequent post-processing of the resulted approximate solution. By the end of each stage, we have $\mathrm{Risk}_{\|\cdot\|}(\widehat{y}_N|X)$ being halved in expectation. Then, we "sparsify" the obtained approximate solution by zeroing out all but s entries of largest amplitudes. This sparse point is used then to initialize the SMD algorithm at the next stage. At the preliminary phase we perform a fixed number m_0 of iterations of the SMD so that the expected quadratic error $\mathbb{E}\|\widehat{y}_N - x^*\|_2^2$ decreases linearly. The asymptotic phase is dedicated to decrease the variance part, and we increase exponentially the number of iterations per stage m_k. The resulted asymptotic expected quadratic error decreases as $\widetilde{\mathcal{O}}\left(\frac{\sigma^2 s \log n}{\mu^2 N}\right)$ which corresponds to (4.5).

We are in shape to present the precise scheme of the SMD-SR procedure in Algorithm 4.1, where we assume to be given $R < \infty$ and $x_0 \in X$ such that $\|x^* - x_0\| \le R$, along with the problem parameters $\varkappa, \varkappa', \nu, \varsigma_*^2, \mu$ and an upper bound $\bar{s}$ for signal sparsity. The number of steps of the SMD to perform locally $(m_k)_{k=0}^{K'}$ and the step sizes $(\beta_k)_{k=0}^{K'}$ are chosen such that $\mathrm{Risk}_{\|\cdot\|}(\widehat{y}_N|X)$ is halved in expectation based on the convergence guarantees of Proposition 4.1. Properties of the proposed procedure are summarized in the following statement.

Theorem 4.1. *In the situation of this section, suppose that $N \ge m_0$ so at least one preliminary stage of Algorithm 4.1 is completed. Then there is an absolute $c > 0$ such that approximate solutions $\widehat{x}_N$ and $\widehat{y}_N$ produced by the algorithm satisfy*

$$\mathrm{Risk}_f(\widehat{x}_N|X) \le \frac{\mu R^2}{\bar{s}} \exp\left(-\frac{cN\mu}{\Theta\varkappa\bar{s}\nu}\right) + C\frac{\varsigma_*^2 \bar{s}\varkappa'\Theta}{\mu N}, \tag{4.15}$$

$$\begin{aligned}\mathrm{Risk}_{\|\cdot\|}(\widehat{y}_N|X) &\le 2s\mathrm{Risk}_{\|\cdot\|_2}(\widehat{y}_N|X) \le 8s\mathrm{Risk}_{\|\cdot\|}(\widehat{x}_N|X)\\ &\lesssim R^2 \exp\left(-\frac{cN\mu}{\Theta\varkappa\bar{s}\nu}\right) + \frac{\Theta\varkappa'\varsigma_*^2\bar{s}^2}{\mu^2 N}.\end{aligned} \tag{4.16}$$

4.3.1 Enhancing reliability of SMD-SR solutions

In this section, our objective is to build approximate solutions to problem (4.1) utilizing Algorithm 4.1 which obey "sub-Gaussian type" bounds on their ε-risks. Note that bounds (4.15) and (4.16) of Theorem 4.1 do allow only for Chebyshev-type bounds for risks of $\widehat{y}_N$ and $\widehat{x}_N$. Nevertheless, their confidence can be easily improved by applying, for instance, an adapted version of "median-of-means" estimate [Nemirovsky and Yudin, 1983, Minsker, 2015].

Algorithm 4.1 Stochastic Mirror Descent algorithm for Sparse Recovery [SMD-SR]

Input: x_0 in $\mathbb{R}^p$ (initial point); N (budget of iterations);

Initialization:

Set $y_0 = x_0$, $R_0 = R \geq \|x_0 - x^*\|$, constant step size

$$\beta_0 = 2\varkappa\nu, \tag{4.17}$$

the number of steps of the SMD algorithm to perform locally

$$m_0 = \left\lfloor 16\mu^{-1}\bar{s}(8\Theta\varkappa + 1)\nu \right\rfloor; \tag{4.18}$$

and the number of preliminary stages performed

$$\overline{K} = \left\lfloor \log_2\left(\frac{R_0^2\mu\nu\varkappa}{32\varsigma_*^2\bar{s}\varkappa'}\right)\right\rfloor \quad \text{and} \quad K = \min\left\{\left\lfloor\frac{N}{m_0}\right\rfloor, \overline{K}\right\}. \tag{4.19}$$

for $k = 1, \ldots, K$ **do stages of the preliminary phase**

— Launch the SMD algorithm, initialized with y_{k-1}, for m_0 iterations with constant step size parameter β_0. Obtain an approximate solution $\widehat{x}_{m_0}(y_{k-1}, \beta_0)$.

— Define y_k as s-sparsification of $\widehat{x}_{m_0}(y_{k-1}, \beta_0)$, being $y_k = \mathrm{sparse}(\widehat{x}_{m_0}(y_{k-1}, \beta_0))$.

end for

Output of the phase:

define $\widehat{y}^{(1)} = y_K$ and $\widehat{x}^{(1)} = \widehat{x}_{m_0}(y_{K-1}, \beta)$.

Initialization of the asymptotic phase:

Set the remaining budget $M = N - m_0\overline{K}$.

Set the number of steps of the SMD to perform per stage

$$m_k = \left\lfloor 512\frac{\bar{s}\Theta\nu\varkappa}{\mu}2^k \right\rfloor.$$

if $m_1 > M$ **then** we do not have a budget for the asymptotic phase.

We output $\widehat{y}_N = \widehat{y}^{(1)}$ and $\widehat{x}_N = \widehat{x}^{(1)}$.

end if

Set $y'_0 = \widehat{y}^{(1)}$ and $\beta_k = 2^k\nu\varkappa$ and the number of asymptotic stages to perform

$$K' = \max\left\{k : \sum_{i=1}^{k} m_i \leq M\right\}.$$

for $k = 1, \ldots, K'$ **do stages of the asymptotic phase**

— Launch the SMD initialized with y'_{k-1}, for m_k iterations and constant step size parameter β_k. Obtain an approximate solution $\widehat{x}_{m_k}(y'_{k-1}, \beta_k)$.

— Define y'_k as s-sparsification of $\widehat{x}_{m_k}(y'_{k-1}, \beta_k)$.

end for

Output: $\widehat{y}_N = y'_{K'}$ and $\widehat{x}_N = \widehat{x}_{m_K}(y'_{K'-1}, \beta_{K'})$.

Reliable recovery utilizing geometric median of SMD-SR solutions. Suppose that available sample of length N can be split into L independent samples of length $M = N/L$ (for the sake of simplicity let us assume that N is a multiple of L). We run Algorithm 4.1 on each subsample thus obtaining L independent recoveries $\widehat{x}_M^{(1)}, ..., \widehat{x}_M^{(L)}$ and compute "enhanced solutions" using an aggregation procedure of geometric median-type. Note that we are in the situation where Theorem 4.1 applies, meaning that approximate solutions $\widehat{x}_M^{(1)}, ..., \widehat{x}_M^{(L)}$ satisfy

$$\forall \ell \quad \mathbb{E}\left[f(\widehat{x}_M^{(\ell)})\right] - f^* \le \tau_M^2 := \frac{\mu R^2}{\bar{s}} \exp\left(-\frac{cM\mu}{\Theta \varkappa \bar{s} \nu}\right) + C\frac{\varsigma_*^2 \bar{s} \varkappa' \Theta}{\mu M}, \tag{4.20}$$

and so

$$\forall \ell \quad \mathbb{E}\left[\left\|\widehat{x}_M^{(\ell)} - x^*\right\|_2^2\right] \le \theta_M^2 := \frac{2}{\mu}\tau_M^2 \lesssim \frac{R^2}{\bar{s}} \exp\left(-\frac{cM\mu}{\Theta \varkappa \bar{s} \nu}\right) + \frac{\Theta \varkappa' \varsigma_*^2 \bar{s}}{\mu^2 M}. \tag{4.21}$$

We are to select among $\widehat{x}_M^{(\ell)}$ the solution which attains similar bounds "reliably."

1. The first reliable solution $\widehat{x}_{N,1-\varepsilon}$ of x^* is a "pure" geometric median of $\widehat{x}_M^{(1)}, ..., \widehat{x}_M^{(L)}$: we put

$$\widehat{x}_{N,1-\varepsilon} \in \operatorname*{Argmin}_x \sum_{\ell=1}^{L} \left\|x - \widehat{x}_M^{(\ell)}\right\|_2, \tag{4.22}$$

   and then define $\widehat{y}_{N,1-\varepsilon} = \mathrm{sparse}(\widehat{x}_{N,1-\varepsilon})$.
   Computing reliable solutions $\widehat{x}_{N,1-\varepsilon}$ and $\widehat{y}_{N,1-\varepsilon}$ as optimal solutions to (4.22) amounts to solving a nontrivial optimization problem. A simpler reliable estimation can be computed by replacing the geometric median $\widehat{x}_{N,1-\varepsilon}$ by its "empirical counterparts" (note that, number L of solutions to be aggregated is not large—it is typically order of $\log[1/\varepsilon]$).
2. We can replace $\widehat{x}_{N,1-\varepsilon}$ with

$$\widehat{x}'_{N,1-\varepsilon} \in \operatorname*{Argmin}_{x \in \left\{\widehat{x}_M^{(1)}, ..., \widehat{x}_M^{(L)}\right\}} \sum_{\ell=1}^{L} \left\|x - \widehat{x}_M^{(\ell)}\right\|_2$$

   and compute its sparse approximation $\widehat{y}'_{N,1-\varepsilon} = \mathrm{sparse}(\widehat{x}'_{N,1-\varepsilon})$.
3. Another reliable solution (with slightly better guarantees) was proposed in [Hsu and Sabato, 2014]. Let $i \in \{1, ..., L\}$, we set

$$r_{ij} = \left\|\widehat{x}_M^{(i)} - \widehat{x}_M^{(j)}\right\|_2$$

   and denote $r_{(1)}^i \le r_{(2)}^i \le ... \le r_{(L-1)}^i$ corresponding order statistics (i.e., $r_{i\cdot}$'s sorted in the increasing order). We define reliable solution $\widehat{x}''_{N,1-\varepsilon} = \widehat{x}_M^{(\widehat{i})}$ where

$$\widehat{i} \in \operatorname*{Argmin}_{i \in \{1,...,L\}} r^i_{\lceil L/2 \rceil} \tag{4.23}$$

   and put $\widehat{y}''_{N,1-\varepsilon} = \mathrm{sparse}(\widehat{x}''_{N,1-\varepsilon})$.

Theorem 4.2. *Let $\varepsilon \in (0, \frac{1}{4}]$, and let $\overline{x}_N$ (resp. $\overline{y}_N$) be one of reliable solutions $\widehat{x}_{1-\varepsilon,N}, \widehat{x}'_{1-\varepsilon,N}$ and $\widehat{x}''_{1-\varepsilon,N}$ (resp., $\widehat{y}_{1-\varepsilon,N}, \widehat{y}'_{1-\varepsilon,N}$ and $\widehat{y}''_{1-\varepsilon,N}$) described above using $L = \lfloor \alpha \log[1/\varepsilon] \rfloor$ [3] independent approximate solutions $\widehat{x}_M^{(1)}, ..., \widehat{x}_M^{(L)}$ by Algorithm 4.1. When $N \geq Lm_0$ we have*

$$\begin{aligned}\mathrm{Risk}_{\|\cdot\|,\varepsilon}(\overline{y}_N|X) &\leq \sqrt{2s}\mathrm{Risk}_{\|\cdot\|_2,\varepsilon}(\overline{y}_N|X) \leq \sqrt{8s}\mathrm{Risk}_{\|\cdot\|_2,\varepsilon}(\overline{x}_N|X)\\ &\lesssim R\exp\left(-\frac{cN\mu}{\Theta\varkappa\bar{s}\nu\log[1/\varepsilon]}\right) + \frac{\varsigma_*\bar{s}}{\mu}\sqrt{\frac{\Theta\varkappa'\log[1/\varepsilon]}{N}}. \qquad (4.24)\end{aligned}$$

Remark. Notice that the term $\log[1/\varepsilon]$ enters the bound (4.24) as a multiplier which is typical for accuracy estimates of solutions which relies upon median to enhance confidence. A better dependence on reliability tolerance parameter with the corresponding term entering as in $\Theta + \log[1/\varepsilon]$ may be obtained for algorithm utilizing "trimmed" stochastic gradients [Juditsky et al., 2019].

Algorithm 4.2 Reliable aggregation

Input: approximate solutions of the first step $\widehat{x}_M^{(1)}, ..., \widehat{x}_M^{(L)}$; observation samples lengths N and K; algorithm parameters $\varepsilon \in (0, \frac{1}{2}]$, $L' \in \mathbb{Z}_+$ and $m = K/L'$ (for the sake of simplicity we assume, as usual, that $K = mL'$).

Compute $\widehat{x}''_{N,1-\varepsilon} = \widehat{x}_M^{(\widehat{i})}$ the reliable solution as defined in (4.23) and denote $\widehat{I} = \{i_1, ..., i_{\lceil L/2 \rceil}\}$, the set of indices of $\lceil L/2 \rceil$ closest to $\widehat{x}''_{N,1-\varepsilon}$ points among $\widehat{x}_M^{(1)}, ..., \widehat{x}_M^{(L)}$.

Comparison procedure:

Split the (second) sample ω^K into L' independent subsamples $\left(\omega^\ell\right)_{l=1}^{L'}$ of size m.

For all $i \in \widehat{I}$, compute the index

$$\widehat{v}_i = \max_{j \in \widehat{I}, j \neq i} \left\{ \operatorname*{median}_\ell [\widehat{v}_{ji}^\ell] - \rho_{ij} \right\}$$

where

$$\widehat{v}_{ji}^\ell = \frac{1}{m}\sum_{k=1}^m \left\langle \nabla \tilde{f}(\widehat{x}_M^{(j)} + t_k(\widehat{x}_M^{(i)} - \widehat{x}_M^{(j)}), \omega_k^\ell), \widehat{x}_M^{(i)} - \widehat{x}_M^{(j)} \right\rangle, \quad \ell = 1, ..., L',$$

are estimates of $v_{ji} = f(\widehat{x}_M^{(i)}) - f(\widehat{x}_M^{(j)})$, $t_k = \frac{2k-1}{2m}$, $k = 1, ..., m$, and coefficients $\rho_{ij} > 0$ to be defined depend on $r_{ij} = \left\|\widehat{x}_M^{(i)} - \widehat{x}_M^{(j)}\right\|_2$.

Output: We say that $x_M^{(i)}$ *is admissible* if $\widehat{v}_i \leq 0$. When the set of admissible $\widehat{x}_M^{(i)}$'s is nonempty we define the procedure output $\overline{x}_{N+K,1-\varepsilon}$ as one of admissible $\widehat{x}_M^{(i)}$'s, and define $\overline{x}_{N+K,1-\varepsilon} = \widehat{x}_M^{(1)}$ otherwise.

3. The exact value of the numeric constant α is specific for each construction, and can be retrieved from the proof of the theorem.

Reliable solution aggregation. Let us assume that two independent observation samples of lengths N and K are available. In the present approach, we use the first sample to compute, same as in the construction presented above, L independent approximate SMD-SR solutions $\widehat{x}_M^{(\ell)}$, $\ell = 1, ..., L$, $M = N/L$. Then we "aggregate" $\widehat{x}_M^{(1)}, ..., \widehat{x}_M^{(L)}$—select the best of them in terms of the objective value $f(\widehat{x}_M^{(\ell)})$ by computing reliable estimations of differences $f(\widehat{x}_M^{(i)}) - f(\widehat{x}_M^{(j)})$ using observations of the second subsample.

The proposed procedure for reliable selection of the "best" solution $\widehat{x}_M^{(\ell)}$ is presented in Algorithm 4.2. Now, consider the following (cf. Assumption 4.3)

Assumption 4.3. *There are $1 \le \chi, \chi' < \infty$ such that for any $x \in X$ and $z \in E$ the following bound holds:*

$$\mathbb{E}\left[\langle \zeta(x,\omega), z\rangle^2\right] \le \|z\|_2^2 \left[\chi \mathcal{L}_2 \left(f(x) - f^*\right) + \chi' \varsigma_*^2\right] \tag{4.25}$$

where $\mathcal{L}_2$ is the Lipschitz constant of the gradient ∇f of f with respect to the Euclidean norm,

$$\|\nabla f(x') - \nabla f(x'')\|_2 \le \mathcal{L}_2 \|x' - x''\|_2, \quad \forall x', x'' \in X.$$

Theorem 4.3. *Let Assumption 4.3 hold, and let τ_M and θ_M be as in (4.20) and (4.21) respectively. Further, in the situation of this section, let $\varepsilon \in (0, \frac{1}{2}]$, $L = \lfloor \alpha \log[1/\varepsilon] \rfloor$ for large enough α, and let $\overline{x}_{N+K,1-\varepsilon}$ be an approximate solution by Algorithm 4.2 in which we set $L' \ge \left\lfloor 7 \log[2/\varepsilon] \right\rfloor$ and*

$$\rho_{ij} = 2r_{ij}\sqrt{\frac{\mathcal{L}_2 \chi}{m}}(\gamma(r_{ij}) + \tau_M) + 2r_{ij}\varsigma_* \sqrt{\frac{\chi'}{m}}$$

where

$$\gamma(r) = \left(\left[4r\sqrt{\frac{\chi \mathcal{L}_2}{m}} + \tau_M\right]^2 + 4r\zeta_* \sqrt{\frac{\chi'}{m}}\right)^{1/2}. \tag{4.26}$$

Then

$$\mathrm{Risk}_{f,\varepsilon}(\overline{x}_{N+K,1-\varepsilon} | X) \le \bar{\gamma}^2 := \gamma^2(8\theta_M),$$

In particular, when $K = mL' \ge c \max\left\{\frac{\chi \mathcal{L}_2 \log[1/\varepsilon]}{\mu}, \frac{N\chi'}{\Theta \varkappa' \bar{s}}\right\}$ for an appropriate absolute $c > 0$, one has

$$\mathrm{Risk}_{f,\varepsilon}(\overline{x}_{N+K,1-\varepsilon} | X) \lesssim \frac{\mu R^2}{\bar{s}} \exp\left(-\frac{cN\mu}{\Theta \varkappa \bar{s} \nu \log[1/\varepsilon]}\right) + \frac{\varsigma_*^2 \bar{s} \Theta \varkappa' \log[1/\varepsilon]}{\mu N}.$$

4.4 Applications

4.4.1 Sparse linear regression by stochastic approximation

Let us consider the problem of recovery of a sparse signal $x^* \in \mathbb{R}^n$, $n \ge 3$, from independent and identically distributed observations

$$\eta_i = \phi_i^T x^* + \sigma \xi_i, \quad i = 1, 2, ..., N$$

with ϕ_i and ξ_i mutually independent and such that $\mathbb{E}\left[\phi_i\phi_i^T\right] = \Sigma$, $\mu_\Sigma I \preceq \Sigma$, and $\|\Sigma\|_\infty \le \upsilon$, with known $\mu_\Sigma > 0$ and υ; we also assume that $\mathbb{E}[\xi_i] = 0$ and $\mathbb{E}[\xi_i^2] \le 1$.

We suppose that x^* is *s-sparse*. Furthermore, we assume to be given a convex and closed subset X of $\mathbb{R}^n$ (e.g., a large enough ball of ℓ_1- or ℓ_2-norm centered at the origin) such that $x^* \in X$, along with $R < \infty$ and $x_0 \in X$ such that $\|x^* - x_0\|_1 \le R$.

We are about to apply the SMD-SR approach from Algorithm 4.1. To this end, consider the following stochastic optimization problem

$$\min_{x\in X}\left\{f(x) = \frac{1}{2}\mathbb{E}\left[(\eta - \phi^T x)^2\right]\right\}, \tag{4.27}$$

so that $\tilde f(x,\omega) = (\eta - \phi^T x)^2$ with $\omega = [\phi, \eta]$. Note that x^* is the unique optimal solution to the above problem. Indeed, we have

$$f(x) = \frac{1}{2}\mathbb{E}\left[[\phi^T(x^* - x) + \sigma\xi]^2\right] = \frac{1}{2}\underbrace{(x - x^*)^T\Sigma(x - x^*)}_{=:\|x-x^*\|_\Sigma^2} + \frac{1}{2}\sigma^2.$$

Observe that $\nabla f(x) = \Sigma(x - x^*) = \mathbb{E}\left[\nabla\tilde f(x,\omega)\right]$ where

$$\nabla\tilde f(x,\omega) = \phi\phi^T(x - x^*) - \sigma\xi\phi, \tag{4.28}$$

and

$$\zeta(x,\omega) = \nabla\tilde f(x,\omega) - \nabla f(x) = \left[\phi\phi^T - \Sigma\right](x - x^*) - \sigma\xi\phi.$$

Further, the objective f is strongly convex with respect to the ℓ_2-norm, so that the condition (4.2) holds with the parameter $\mu = \mu_\Sigma$. We set $\|\cdot\| = \|\cdot\|_1$ with $\|\cdot\|_* = \|\cdot\|_\infty$, and we use the "$\ell_1$-proximal setup" of the SMD-SR algorithm with a quadratically growing for $n > 2$ distance-generating function, see, for instance, Theorem 2.1 in [Nesterov and Nemirovski, 2013]

$$\vartheta(x) = \frac{1}{2}e\log(n)\, n^{(p-1)(2-p)/p}\,\|x\|_p^2, \quad p = 1 + \frac{1}{\log n},$$

where the corresponding Θ satisfies $\Theta \le \frac{1}{2}e^2\log n$.

Note that

$$\varsigma^2(x) = \mathbb{E}\left[\left\|\nabla\tilde f(x,\omega) - \nabla f(x)\right\|_\infty^2\right] = \mathbb{E}\left[\left\|\left[\phi\phi^T - \Sigma\right](x - x^*) - \sigma\xi\phi\right\|_\infty^2\right].$$

Therefore, in our present situation Assumption 4.3 reads

$$\mathbb{E}\left[\left\|\left[\phi\phi^T - \Sigma\right](x - x^*) - \sigma\xi\phi\right\|_\infty^2\right] \le \frac{1}{2}\varkappa\nu\,\|x - x^*\|_\Sigma^2 + \varsigma_*^2. \tag{4.29}$$

The following statement is a straightforward corollary of Theorems 4.1 and 4.2; it describes the properties of approximate solutions obtained by Algorithm 4.1 when applied to the observation model in this section. We assume that the problem parameters—values of $\varkappa$, ν, μ_Σ, σ^2 and an upper bound $\bar s$ on sparsity of x^*—are known.

Proposition 4.2. *Suppose that (4.29) holds.*
(i) Let the sample size N satisfy

$$N \geq m_0 = \left\rfloor \frac{16\nu\bar{s}}{\mu_\Sigma}(4e^2\varkappa\log[n]+1)\right\lfloor$$

so at least one preliminary stage of Algorithm 4.1 is completed. Then, approximate solutions $\widehat{x}_N$ and $\widehat{y}_N$ produced by the algorithm satisfy

$$\begin{aligned}\mathrm{Risk}_{\|\cdot\|}(\widehat{y}_N|X) &\leq 8s\mathrm{Risk}_{\|\cdot\|}(\widehat{x}_N|X) \lesssim R^2\exp\left(-\frac{cN\mu_\Sigma}{\varkappa\bar{s}\nu\log n}\right)+\frac{\nu\sigma^2\bar{s}^2\log n}{\mu_\Sigma^2 N}, \qquad (4.30)\\ \mathrm{Risk}_f(\widehat{x}_N|X) &\lesssim \frac{\mu_\Sigma R^2}{\bar{s}}\exp\left(-\frac{cN\mu_\Sigma}{\varkappa\bar{s}\nu\log n}\right)+\frac{\nu\sigma^2\bar{s}\varkappa'\log n}{\mu_\Sigma N}.\end{aligned}$$

(ii) Furthermore, when observation size satisfies $N \geq \alpha m_0\log[1/\varepsilon]$ with large enough absolute $\alpha > 0$, $1-\varepsilon$ reliable solutions $\widehat{y}_{N,1-\varepsilon}$ and $\widehat{x}_{N,1-\varepsilon}$ as defined in Section 4.3.1 satisfy

$$\begin{aligned}\mathrm{Risk}_{\|\cdot\|,\varepsilon}(\widehat{y}_{N,1-\varepsilon}|X) &\leq \sqrt{2s}\mathrm{Risk}_{\|\cdot\|_2,\varepsilon}(\widehat{y}_{N,1-\varepsilon}|X) \leq 2\sqrt{2s}\mathrm{Risk}_{\|\cdot\|_2,\varepsilon}(\widehat{x}_{N,1-\varepsilon}|X)\\ &\lesssim R\exp\left(-\frac{cN\mu_\Sigma}{\varkappa\bar{s}\nu\log[1/\varepsilon]\log n}\right)+\frac{\sigma\bar{s}}{\mu_\Sigma}\sqrt{\frac{\nu\log[1/\varepsilon]\log n}{N}}, \qquad (4.31)\end{aligned}$$

with $\widehat{x}'_{N,1-\varepsilon}$, $\widehat{x}''_{N,1-\varepsilon}$ and $\widehat{y}'_{N,1-\varepsilon}$, $\widehat{y}''_{N,1-\varepsilon}$ verifying similar bounds.

Note that Assumption 4.3 in the case of sparse linear regression holds when for some $1 \leq \chi < \infty$

$$\forall x \in X,\, z \in \mathbb{R}^n \;\; \mathbb{E}\left[\left(z^T\phi\right)^2\left(\phi^T(x-x^*)\right)^2\right] \leq \frac{1}{2}\chi\left\|z\right\|_2^2\sigma_1(\Sigma)\left\|x-x^*\right\|_\Sigma^2 \qquad (4.32)$$

where $\sigma_1(\Sigma)$ is the principal eigenvalue (the spectral norm) of Σ. Indeed, in this case we have

$$\begin{aligned}\mathbb{E}\left[\left(z^T\zeta(x,\omega)\right)^2\right] &= \mathbb{E}\left[\left(z^T\left\{\left[\phi\phi^T-\Sigma\right](x-x^*)-\sigma\phi\xi\right\}\right)^2\right]\\ &= \mathbb{E}\left[\left(z^T\left[\phi\phi^T-\Sigma\right](x-x^*)\right)^2\right]+\sigma^2\mathbb{E}\left[\xi^2\left(z^T\phi\right)^2\right]\\ &\leq \mathbb{E}\left[\left(z^T\phi\right)^2\left(\phi^T(x-x^*)\right)^2\right]+\sigma^2\left\|z\right\|_2^2\sigma_1(\Sigma)\\ &\leq \underbrace{\frac{1}{2}\left\|x-x^*\right\|_\Sigma^2\chi}_{=f(x)-f^*}\left\|z\right\|_2^2\sigma_1(\Sigma)+\sigma_1(\Sigma)\sigma^2\left\|z\right\|_2^2\end{aligned}$$

implying Assumption 4.3 with $\chi' = \sigma_1(\Sigma)/\nu$.

The following result is a corollary of Theorem 4.3.

Proposition 4.3. *Suppose that (4.29) and (4.32) hold true, and let*

$$N \geq c\max\left\{\frac{\varkappa\nu\bar{s}}{\mu_\Sigma}\log[1/\varepsilon]\log n,\, \frac{\chi\sigma_1(\Sigma)}{\mu_\Sigma}\log[1/\varepsilon]\right\}$$

with large enough $c > 0$. Then aggregated solution $\overline{x}_{2N,1-\varepsilon}$ (with $K = N$) by Algorithm 4.2 satisfies

$$\mathrm{Risk}_{f,\varepsilon}(\overline{x}_{2N,1-\varepsilon}|X) \lesssim \frac{\mu_\Sigma R^2}{\bar{s}} \exp\left(-\frac{cN\mu_\Sigma}{\varkappa\bar{s}\nu\log[1/\varepsilon]\log n}\right) + \frac{\sigma^2\nu\bar{s}\log[1/\varepsilon]\log n}{\mu_\Sigma N}. \qquad (4.33)$$

Note that when $\sigma_1(\Sigma) = \mathcal{O}(\nu\log n)$ and $\varkappa$ and χ are both $\mathcal{O}(1)$ bounds (4.31) and (4.33) hold for $N \geq c\frac{\nu\bar{s}}{\mu_\Sigma}\log[1/\varepsilon]\log n$.

Remark. Results of Propositions 4.2 and (4.3) merit some comments. If compared to now standard accuracy bounds for sparse recovery by ℓ_1-minimization [Candes, 2006, 2008, Bickel et al., 2009, Van De Geer and Bühlmann, 2009, Raskutti et al., 2010, Juditsky and Nemirovski, 2011b, Candes and Plan, 2011a, Rudelson and Zhou, 2012], to the best of our knowledge, (4.29) and (4.32) provide the most relaxed conditions under which the bounds such as (4.30)–4.33 can be established. One may notice a degradation of bounds (4.31) and (4.33) with respect to comparable results [Juditsky and Nemirovski, 2011b, Raskutti et al., 2010, Dalalyan and Thompson, 2019] as far as dependence in factors which are logarithmic in n and ε^{-1} is concerned—bound (4.24) depends on the product $\log[n]\log[1/\varepsilon]$ of these terms instead of the sum $\log[n] + \log[\varepsilon^{-1}]$ in the "classical" results. [4] This seems to be "an artifact" of the reliability enhancement approach using median of estimators we have adopted in this work, cf. the comment after Theorem 4.2. Nevertheless, it is rather surprising to see that conditions on the regressor model in Proposition 4.2, apart from positive definiteness of regressor covariance matrix, essentially resume to (cf. 4.29)

$$\mathbb{E}\left[\left\|\phi\phi^T z\right\|_\infty^2\right] \lesssim \nu\|z\|_\Sigma^2 \quad \forall z \in \mathbb{R}^n.$$

Below we consider some examples of situations where bounds (4.29) and (4.32) hold with constants which are "almost dimension-independent," i.e. are, at most, *logarithmic in problem dimension.* When this is the case, and when observation count N satisfies $N \geq \alpha m_0\log[1/\varepsilon]\log[R/\sigma]$ for large enough absolute α, so that the preliminary phase of the algorithm is completed, the bounds of Propositions 4.2 and 4.3 coincide (up to already mentioned logarithmic in n and $1/\varepsilon$ factors) with the best accuracy bound available for sparse recovery in the situation in question. [5]

1. *Sub-Gaussian regressors:* suppose now that $\phi_i \sim \mathcal{SG}(0, S)$, i.e., regressors ϕ_i are sub-Gaussian with zero mean and matrix parameter S, meaning that

$$\mathbb{E}\left[e^{u^T\phi}\right] \leq e^{\frac{u^TSu}{2}} \text{ for all } u \in \mathbb{R}^n.$$

Let us assume that sub-Gaussianity matrix S is "similar" to the covariance matrix Σ of ϕ, i.e. $S \preceq \mu\Sigma$ with some $\mu < \infty$. Note that $\mathbb{E}\left[(\phi^Tz)^4\right] \leq 16(z^TSz)^2 \leq 16\mu^2\|z\|_\Sigma^4$, and thus

$$\mathbb{E}\left[(z^T\phi\phi^Tx)^2\right] \leq \mathbb{E}\left[(z^T\phi)^4\right]^{1/2}\mathbb{E}\left[(x^T\phi)^4\right]^{1/2} \leq 16z^TSz\,x^TSx \leq 16\mu^2\sigma_1(\Sigma)\|z\|_2^2\|x\|_\Sigma^2,$$

4. Note that a similar deterioration was noticed in [Candes and Plan, 2011a].

5. In the case of "isotropic sub-Gaussian" regressors, see [Lecué et al., 2018], the bounds of Proposition 4.2 are comparable to bounds of [Lecué et al., 2020, Theorem 5] for Lasso recovery under relaxed moment assumptions on the noise ξ.

what is (4.32) with $\chi = 16\mu^2$. Let us put $\bar{\upsilon} = \max_i [S]_{ii}$. One easily verifies that in this case

$$\nu = \mathbb{E}\left[\|\phi\|_\infty^2\right] \le 2\bar{\upsilon}(\log[2n]+1) \le 2\mu\upsilon(\log[2n]+1),$$

and

$$\mathbb{E}\left[\|\phi\|_\infty^4\right] \le 4\bar{\upsilon}^2(\log^2[2n] + 2\log[2n] + 2) \le 4\mu^2\upsilon^2(\log^2[2n] + 2\log[2n] + 2).$$

As a result, we have

$$\begin{aligned}
\zeta^2(x) &= \mathbb{E}\left[\left\|[\phi\phi^T - \Sigma](x - x^*) - \sigma\xi\phi\right\|_\infty^2\right] \\
&\le \left[(\mathbb{E}\left[\|\phi\|_\infty^4\right])^{1/4}(\mathbb{E}\left[(\phi^T(x-x^*))^4\right])^{1/4} + \sigma(\mathbb{E}\left[\|\phi\|_\infty^2\right])^{1/2} + \sqrt{\upsilon}\,\|x - x^*\|_\Sigma\right]^2 \\
&\le \left[\sqrt{8\bar{\upsilon}(\log[2n]+2)}\,\|x-x^*\|_S + \sigma\sqrt{2\bar{\upsilon}(\log[2n]+1)} + \sqrt{\upsilon}\,\|x-x^*\|_\Sigma\right]^2 \\
&\le 2\left(\mu\sqrt{8(\log[2n]+2)}+1\right)^2 \upsilon\,\|x-x^*\|_\Sigma^2 + 4\mu\upsilon(\log[2n]+1)\sigma^2.
\end{aligned}$$

whence, (4.3) holds with $\varkappa\nu \lesssim \mu^2\upsilon\log n$, $\varkappa' \lesssim 1$, and $\varsigma_*^2 \lesssim \mu\upsilon\sigma^2\log n$.

2. *Bounded regressors:* we assume that $\|\phi_i\|_\infty \le \mu$ a.s.. One has

$$\begin{aligned}
\varsigma^2(x) &= \mathbb{E}\left[\left\|\left[\phi\phi^T - \Sigma\right](x-x^*) - \sigma\xi\phi\right\|_\infty^2\right] \\
&\le \mathbb{E}\left[\|\phi\|_\infty\left(|\phi^T(x-x^*)| + \sigma|\xi|\right) + \left\|\mathbb{E}\{\phi\phi^T(x-x^*)\right\|_\infty^2\right] \\
&\le \mu^2\mathbb{E}\left[\left[|\phi^T(x-x^*)| + \sigma|\xi| + \sqrt{\upsilon}\,\|x-x^*\|_\Sigma\right]^2\right] \\
&\le 2(\mu+\sqrt{\upsilon})^2\,\|x-x^*\|_\Sigma + 2\mu^2\sigma^2
\end{aligned}$$

implying (4.29) with $\varkappa\nu \le 4(\mu+\sqrt{\upsilon})^2$ and $\varsigma_*^2 \le \mu^2\sigma^2$. In particular, this condition is straightforwardly satisfied when ϕ_j are sampled from an orthogonal system with uniformly bounded elements, e.g., $\phi_j = \sqrt{n}\psi_{\kappa_j}$ where $\{\psi_j, j = 1, ..., n\}$ is a trigonometric or Hadamard basis of $\mathbb{R}^n$, and κ_j are independent and uniformly distributed over $\{1, ..., n\}$. On the other hand, in this case, for $z = x = \psi_1$ we have

$$\mathbb{E}\left[(z^T\phi\phi^T x)^2\right] = \mathbb{E}\left[(\psi_1\phi\phi^T\psi_1)^2\right] = n = n\,\|\psi_1\|_2^4 = n\,\|x\|_2^2\,\|z\|_2^2,$$

implying that (4.32) can only hold with $\chi = \mathcal{O}(n)$ in this case.

Besides this, when ϕ is a linear image of a Rademacher vector, i.e. $\phi = A\eta$ where $A \in \mathbb{R}^{m\times n}$ and η has independent components $[\eta]_i \in \{\pm 1\}$ with $\mathrm{Prob}\{[\eta]_i = 1\} = \mathrm{Prob}\{[\eta]_i = -1\} = 1/2$, one has $\Sigma = AA^T$, and

$$\begin{aligned}
\mathbb{E}\left[(z^T\phi\phi^T(x-x^*))^2\right] &\le \mathbb{E}\left[(z^T\phi)^4\right]^{1/2}\mathbb{E}\left[((x-x^*)^T\phi)^4\right]^{1/2} \\
&\le 2z^T\Sigma z\,(x-x^*)^T\Sigma(x-x^*) \le 2\sigma_1(\Sigma)\,\|z\|_2^2\,\|x-x^*\|_\Sigma^2
\end{aligned}$$

implying (4.32) with $\chi = 4$. On the other hand, $\mathbb{E}\left[(\phi^T x)^4\right] \le 2\left\|A^T x\right\|_2^4$ (cf. the case of sub-Gaussian regressors above), and, denoting $\mu = \max_j \|\mathrm{Row}_j(A)\|_2$, we get $\mathrm{Prob}\left(\|\phi\|_\infty^4 \ge t\mu\right) \le 2ne^{-t^2/2}$, with

$$\mathbb{E}\left[\|\phi\|_\infty^2\right] \le 2\mu^2[\log[2n]+1] \text{ and } \mathbb{E}\left[\|\phi\|_\infty^4\right] \le 4\mu^4[\log^2[2n]+2\log[2n]+2].$$

Thus,

$$\begin{aligned}
\zeta^2(x) &= \mathbb{E}\left[\left\|[\phi\phi^T-\Sigma](x-x^*)-\sigma\xi\phi\right\|_\infty^2\right] \\
&\le \left[(\mathbb{E}\left[\|\phi\|_\infty^4\right])^{1/4}(\mathbb{E}\left[(\phi^T(x-x^*))^4\right])^{1/4}+\sigma(\mathbb{E}\left[\|\phi\|_\infty^2\right])^{1/2}+\sqrt{\upsilon}\,\|x-x^*\|_\Sigma\right]^2 \\
&\le \left[\sqrt{4(\log[2n]+2)}\mu\,\|x-x^*\|_\Sigma+\sigma\sqrt{2(\log[2n]+1)\mu}+\sqrt{\upsilon}\,\|x-x^*\|_\Sigma\right]^2 \\
&\le 2\left(\mu\sqrt{4(\log[2n]+2)}+\sqrt{\upsilon}\right)^2\|x-x^*\|_\Sigma^2+4\mu^2(\log[2n]+1)\sigma^2
\end{aligned}$$

what is (4.9) with $\varkappa\nu \lesssim \mu^2\log n+\upsilon$, $\varkappa' \lesssim 1$, and $\varsigma_*^2 \lesssim \mu^2\sigma^2\log n$.

3. *Scale mixtures:* Let us now assume that

$$\phi \sim \sqrt{Z}\eta, \tag{4.34}$$

where Z is a scalar a.s. positive random variable, and $\eta \in \mathbb{R}^n$ is independent of Z with covariance matrix $\mathbb{E}\left[\eta\eta^T\right] = \Sigma_0$. Because

$$\mathbb{E}\left[\|\phi\|_\infty^2\right] = \mathbb{E}\left[Z\right]\mathbb{E}\left[\|\eta\|_\infty^2\right], \quad \mathbb{E}\left[\left\|\phi\phi^T z\right\|_\infty^2\right] = \mathbb{E}\left[Z^2\right]\mathbb{E}\left[\left\|\eta\eta^T z\right\|_\infty^2\right]$$

and

$$[\Sigma :=]\ \mathbb{E}\left[\phi\phi^T\right] = \mathbb{E}\left[Z\right]\mathbb{E}\left[\eta\eta^T\right],$$

we conclude that if random vector η satisfies (4.29) with Σ_0 substituted for Σ and $\mathbb{E}[Z^2]$ is finite then a similar bound also holds for ϕ. It is obvious that if η satisfies (4.32) then

$$\begin{aligned}
\mathbb{E}\left[(z^T\phi\phi^T x)^2\right] = \mathbb{E}\left[Z^2\right]\mathbb{E}\left[(z^T\eta\eta^T x)^2\right] &\le \frac{\mathbb{E}[Z^2]}{\mathbb{E}[Z]^2}\chi\,\|z\|_\Sigma^2\,\|x\|_\Sigma^2 \\
&\le \chi\frac{\mathbb{E}[Z^2]}{\mathbb{E}[Z]^2}\sigma_1(\Sigma)\,\|z\|_2^2\,\|x\|_\Sigma^2,
\end{aligned}$$

and (4.32) holds for ϕ with χ replaced with $\chi\frac{\mathbb{E}[Z^2]}{\mathbb{E}[Z]^2}$.

Let us consider the situation where $\eta \sim \mathcal{N}(0,\Sigma_0)$ with positive definite Σ_0. In this case ϕ is referred to as Gaussian scale mixture with a standard example provided by *n-variate t-distributions* $t_n(q,\Sigma_0)$ (multivariate Student distributions with q degrees of freedom, see [Kotz and Nadarajah, 2004] and references therein). Here, by definition,

$t_n(q,\Sigma_0)$ is the distribution of the random vector $\phi=\sqrt{Z}\eta$ with $Z=q/\zeta$, where ζ is the independent of η random variable following χ^2-distribution with q degrees of freedom. One can easily see that all one-dimensional projections $e^T\phi$, $\|e\|_2=1$, of ϕ are random variables with univariate t_q-distribution. When $\phi_i\sim t_n(q,\Sigma_0)$ with $q>4$, we have for $\zeta\sim\chi^2_q$

$$\mathbb{E}\left[\frac{q}{\zeta}\right]=\frac{q}{q-2},\quad \mathbb{E}\left[\frac{q^2}{\zeta^2}\right]=\frac{3q^2}{(q-2)(q-4)},$$

so that $\Sigma=\frac{q}{q-2}\Sigma_0$, and

$$\begin{aligned}\varsigma^2(x)&=\mathbb{E}\left[\left\|[\phi\phi^T-\Sigma](x-x^*)-\sigma\xi\phi\right\|_\infty^2\right]\\&\lesssim\frac{q-2}{q-4}\upsilon\log[n]\|x-x^*\|_\Sigma+\sigma^2\upsilon\log n\end{aligned}$$

implying (4.9) with $\varkappa,\varkappa'\lesssim 1$ and $\varsigma_*^2\lesssim\sigma^2\upsilon\log n$. Moreover, in this case

$$\mathbb{E}\left[(z^T\phi\phi^Tx)^2\right]=\mathbb{E}\left[Z^2\right]\mathbb{E}\left[z^T\eta\eta^Tx)^2\right]\le 3\frac{\mathbb{E}[Z^2]}{\mathbb{E}[Z]^2}\|z\|_\Sigma^2\|x\|_\Sigma^2\le 9\frac{q-2}{q-4}\sigma_1(\Sigma)\|z\|_2^2\|x\|_\Sigma^2.$$

Another example of Gaussian scale mixture (4.34) is the *n-variate Laplace distribution* $\mathcal{L}_n(\lambda,\Sigma_0)$ [Eltoft et al., 2006] in which Z has exponential distribution with parameter λ. In this case all one-dimensional projections $e^T\phi$, $\|e\|_2=1$, of ϕ are Laplace random variables. If $\phi_i\sim\mathcal{L}_n(\lambda,\Sigma_0)$ one has

$$\varsigma^2(x)\lesssim\upsilon\log[n]\|x-x^*\|_\Sigma+\sigma^2\upsilon\log n$$

and

$$\mathbb{E}\left[(z^T\phi\phi^Tx)^2\right]\lesssim\sigma_1(\Sigma)\|z\|_2^2\|x\|_\Sigma^2.$$

4.4.2 Stochastic Mirror Descent for low-rank matrix recovery

In this section we consider the problem of recovery of matrix $x_*\in\mathbb{R}^{p\times q}$, from independent and identically distributed observations

$$\eta_i=\langle\phi_i,x^*\rangle+\sigma\xi_i,\quad i=1,2,...,N,\tag{4.35}$$

with $\phi_i\in\mathbb{R}^{p\times q}$ which are random independent over i with covariance operator Σ (defined according to $\Sigma(x)=\mathbb{E}[\phi\langle\phi,x\rangle]$). We assume that $\xi_i\in\mathbf{R}$ are mutually independent and independent of ϕ_i with $\mathbb{E}[\xi_i]=0$ and $\mathbb{E}[\xi_i^2]\le 1$.

In this application, E is the space of $p\times q$ matrices equipped with the Frobenius scalar product

$$\langle a,b\rangle=\mathrm{Tr}\left(a^Tb\right)$$

with the corresponding norm $\|a\|_2=\langle a,a\rangle^{1/2}$. For the sake of definiteness, we assume that $p\ge q\ge 2$. Our choice for the norm $\|\cdot\|$ is the nuclear norm $\|x\|=\|\sigma(x)\|_1$

where $\sigma(x)$ is the singular spectrum of x, so that the conjugate norm is the spectral norm $\|y\|_* = \|\sigma(y)\|_\infty$. We suppose that

$$\mu_\Sigma \|x\|_2^2 \le \langle x, \Sigma(x)\rangle \le \upsilon \|x\|_2^2 \ \forall x \in \mathbb{R}^{p\times q},$$

with known $\mu_\Sigma > 0$ and υ, we write $\mu_\Sigma I \preceq \Sigma \preceq \upsilon I$; for $x \in \mathbb{R}^{p\times q}$ we denote $\|x\|_\Sigma = \sqrt{\langle x, \Sigma(x)\rangle}$. Finally, we assume that matrix x^* is of rank $s \le \bar{s} \le q$, and moreover, that we are given a convex and closed subset X of $\mathbb{R}^{p\times q}$ such that $x^* \in X$, along with $R < \infty$ and $x_0 \in X$ satisfying $\|x^* - x_0\| \le R$.

Consider the following stochastic optimization problem

$$\min_{x\in X} \left\{ f(x) = \frac{1}{2}\mathbb{E}\left[(\eta - \phi^T x)^2\right] \right\}, \tag{4.36}$$

so that $\tilde{f}(x,\omega) = (\eta - \phi^T x)^2$ with $\omega = [\phi, \eta]$. We are to apply SMD algorithm to solve (4.36) with the proximal setup associated with the nuclear norm with quadratically growing for $q \ge 2$ distance-generating function

$$\vartheta(x) = 2e\log(2q)\left[\sum_{j=1}^q \sigma_j^{1+r}(x)\right]^{\frac{2}{1+r}}, \quad r = (12\log[2q])^{-1},$$

(here $\sigma_j(x)$ are singular values of x) with the corresponding parameter $\Theta \le C\log[2q]$ (cf. [Nesterov and Nemirovski, 2013, Theorem 2.3]). Note that, in the premise of this section,

$$f(x) = \frac{1}{2}\mathbb{E}\left[(\sigma\xi + \langle \phi, x^* - x\rangle)^2\right] = \frac{1}{2}(\|x - x^*\|_\Sigma^2 + \sigma^2),$$

with

$$\nabla f(x) = \Sigma(x - x^*) = \mathbb{E}[\underbrace{\phi(\langle \phi, x - x^*\rangle - \sigma\xi)}_{=\nabla\tilde{f}(x,\omega)}]$$

and

$$\zeta(x,\omega) = \nabla\tilde{f}(x,\omega) - \nabla f(x) = [\phi\,\langle \phi, x - x^*\rangle - \Sigma(x - x^*)] - \sigma\phi\xi.$$

Let us now consider the case regressors $\phi_i \in \mathbb{R}^{p\times q}$ drawn independently from a *sub-Gaussian ensemble*, $\phi_i \sim \mathcal{SG}(0, S)$ with sub-Gaussian operator S. The latter means that

$$\mathbb{E}\left[e^{\langle x,\phi\rangle}\right] \le e^{\frac{1}{2}\langle x, S(x)\rangle} \ \ \forall x \in \mathbb{R}^{p\times q}$$

with linear positive definite $S(\cdot)$. To show the bound of Theorems 4.1–4.3 in this case we need to verify that relationships (4.9) and (4.25) of Assumptions 4.3 and 4.3 are satisfied. To this end, let us assume that S is "similar" to the covariance operator Σ of ϕ, namely, $S \preceq \mu\Sigma$ with some $\mu < \infty$. This setting covers, for instance, the situation where the entries

in the regressors matrix $\phi \in \mathbb{R}^{p\times q}$ are standard Gaussian or Rademacher i.i.d. random variables (in these models, $S = \Sigma$ is the identity, and $f(x) - f(x^*) = \frac{1}{2}\|x - x^*\|_2^2$).

Note that, more generally, when $S \preceq \mu\Sigma$ we have $S \preceq \mu\upsilon I$ with

$$\mathbb{E}\left[\|\phi\|_*^4\right] \le C^2\mu^2\upsilon^2(p+q)^2,$$

cf. Lemma 4.3 of the appendix, and

$$\mathbb{E}\left[\langle\phi, x - x^*\rangle^4\right] \le 16\langle x - x^*, S(x - x^*)\rangle^2 \le 16\mu^2\|x - x^*\|_\Sigma^2$$

for sub-Gaussian random variable $\langle\phi, x - x^*\rangle \sim \mathcal{SG}(0, \langle x - x^*, S(x - x^*)\rangle)$. Therefore,

$$\begin{aligned}\mathbb{E}\left[\|\phi\langle\phi, x - x^*\rangle_* - \Sigma(x - x^*)\|^2\right] &\le 2\mathbb{E}\left[\|\phi\langle\phi, x - x^*\rangle\|_*^2\right] + 2\upsilon\|x - x^*\|_\Sigma^2\\ &\le 2\mathbb{E}\left[\|\phi\|_*^4\right]^{1/2}\mathbb{E}\left[\langle\phi, x - x^*\rangle^4\right]^{1/2} + 2\upsilon\|x - x^*\|_\Sigma^2\\ &\le 8C\mu^2(p+q)\upsilon\|x - x^*\|_\Sigma^2 + 2\upsilon\|x - x^*\|_\Sigma^2.\end{aligned}$$

Taking into account that $\nu = \mathbb{E}\left[\|\phi\|_*^2\right] \le C\mu\upsilon(p+q)$ in this case, we have

$$\begin{aligned}\varsigma^2(x) = \mathbb{E}\left[\|\zeta(x,\omega)\|_*^2\right] &\le 2\mathbb{E}\left[\|\phi\langle\phi, x - x^*\rangle - \Sigma(x - x^*)\|_*^2\right] + 2\sigma^2\mathbb{E}\left[\|\phi\|_*^2\right]\\ &\le 8(4C\mu^2(p+q)+1)\upsilon[f(x) - f^*] + 2\underbrace{C\mu\upsilon(p+q)\sigma^2}_{=\varsigma_*^2}\end{aligned}$$

implying (4.9) with $\varkappa \lesssim \mu$ and $\varkappa' \lesssim 1$.

Similarly, we estimate $\forall x \in X$, $z \in \mathbb{R}^{p\times q}$

$$\mathbb{E}\left[\langle\phi, z\rangle^2\langle\phi, x\rangle^2\right] \le \mathbb{E}\left[\langle z, \phi\rangle^4\right]^{1/2}\mathbb{E}\left[\langle\phi, x\rangle^4\right]^{1/2} \le 16\langle z, S(z)\rangle\langle x, S(x)\rangle \le 16\mu^2\upsilon\|z\|_2^2\|x\|_\Sigma^2,$$

so that

$$\begin{aligned}\mathbb{E}\left[\langle z, \zeta(x,\omega)\rangle^2\right] &= \mathbb{E}\left[\langle z, \phi\langle\phi, x - x^*\rangle - \Sigma(x - x^*) - \sigma\phi\xi\rangle^2\right]\\ &= \mathbb{E}\left[(\langle z, \phi\rangle\langle\phi, x - x^*\rangle - \langle z, \Sigma(x - x^*)\rangle)^2\right] + \sigma^2\mathbb{E}\left[\xi^2\langle z, \phi\rangle^2\right]\\ &\le \mathbb{E}\left[\langle z, \phi\rangle^2\langle\phi, x - x^*\rangle^2\right] + \sigma^2\upsilon\|z\|_2^2\\ &\le 16\mu^2\upsilon(f(x) - f^*)\|z\|_2^2 + \sigma^2\upsilon\|z\|_2^2\end{aligned}$$

implying the bound (4.25) with $\chi \lesssim \mu(p+q)^{-1}$ and $\chi' \lesssim \mu^{-1}(p+q)^{-1}$. When substituting the above bounds for problem parameters into statements of Theorems 4.1–4.3 we obtain the following statement summarizing the properties of the approximate solutions by the SMD-SR algorithm utilizing observations (4.35).

Proposition 4.4. *In the situation of this section,*

(i) let the sample size N satisfy

$$N \ge \alpha\left[\frac{\mu^2\upsilon(p+q)\bar{s}\log q}{\mu_\Sigma}\right]$$

for an appropriate absolute α, implying that at least one preliminary stage of Algorithm 4.1 is completed. Then there is an absolute $c > 0$ such that approximate solutions $\widehat{x}_N$ and $\widehat{y}_N$ produced by the algorithm satisfy

$$\mathrm{Risk}_{\|\cdot\|}(\widehat{y}_N|X) \le 8s\mathrm{Risk}_{\|\cdot\|_2}(\widehat{x}_N|X) \lesssim R^2\exp\left(-\frac{cN\mu_\Sigma}{\mu^2\upsilon(p+q)\bar{s}\log q}\right) + \frac{\sigma^2\mu\upsilon(p+q)\bar{s}^2\log q}{\mu_\Sigma^2 N},$$
$$\mathrm{Risk}_f(\widehat{x}_N|X) \lesssim \frac{\mu_\Sigma R^2}{\bar{s}}\exp\left(-\frac{cN\mu_\Sigma}{\mu^2\upsilon(p+q)\bar{s}\log q}\right) + \frac{\sigma^2\mu\upsilon(p+q)\bar{s}\log q}{\mu_\Sigma N}.$$

(ii) Furthermore, when observation size satisfies

$$N \ge \alpha'\left[\frac{\mu^2\upsilon(p+q)\bar{s}\log[1/\varepsilon]\log q}{\mu_\Sigma}\right]$$

with large enough α', $1-\varepsilon$ reliable solutions $\widehat{y}_{N,1-\varepsilon}$ and $\widehat{x}_{N,1-\varepsilon}$ defined in Section 4.3.1 satisfy for some $c' > 0$

$$\begin{aligned}\mathrm{Risk}_{\|\cdot\|,\varepsilon}(\widehat{y}_{N,1-\varepsilon}|X) &\le \sqrt{2s}\mathrm{Risk}_{\|\cdot\|_2,\varepsilon}(\widehat{y}_{N,1-\varepsilon}|X) \le 2\sqrt{2s}\mathrm{Risk}_{\|\cdot\|_2,\varepsilon}(\widehat{x}_{N,1-\varepsilon}|X)\\ &\lesssim R\exp\left(-\frac{c'N\mu_\Sigma}{\mu^2\upsilon(p+q)\bar{s}\log[1/\varepsilon]\log q}\right) + \frac{\sigma\bar{s}}{\mu_\Sigma}\sqrt{\frac{\mu\upsilon(p+q)\log[1/\varepsilon]\log q}{N}},\end{aligned} \tag{4.37}$$

with solutions $\widehat{x}'_{N,1-\varepsilon}$, $\widehat{x}''_{N,1-\varepsilon}$ and $\widehat{y}'_{N,1-\varepsilon}$, $\widehat{y}''_{N,1-\varepsilon}$ verifying analogous bounds. Finally, the following bound holds for the aggregated solution $\overline{x}_{2N,1-\varepsilon}$ (with $K = N$) by Algorithm 4.2:

$$\mathrm{Risk}_{f,\varepsilon}(\overline{x}_{2N,1-\varepsilon}|X) \lesssim \frac{\mu_\Sigma R^2}{\bar{s}}\exp\left(-\frac{c'N\mu_\Sigma}{\mu^2\upsilon(p+q)\bar{s}\log[1/\varepsilon]\log q}\right) + \frac{\sigma^2\mu\upsilon(p+q)\bar{s}\log[1/\varepsilon]\log q}{\mu_\Sigma N}.$$

Remark. Let us now compare the bounds of the proposition to available accuracy estimates for low rank matrix recovery. Notice first, that when assuming that $\mu \lesssim 1$ the bounds of the proposition hold if (the upper bound on unknown) signal rank $\bar{s}$ satisfies

$$\bar{s} \lesssim \frac{N\mu_\Sigma}{(p+q)\upsilon\log[1/\varepsilon]\log q}.$$

The above condition is essentially the same, up to logarithmic in $1/\varepsilon$ factor, as the best condition on rank of the signal to be recovered under which the recovery is exact in the case of exact—noiseless—observation [Candes and Plan, 2011b, Recht et al., 2010]. The risk bounds of Proposition 4.4 can be compared to the corresponding accuracy bounds for recovery $\widehat{x}_{N,\mathrm{Lasso}}$ by Lasso with nuclear norm penalization, as in [Koltchinskii et al., 2011, Negahban and Wainwright, 2011]. For instance, when regressors ϕ_i have i.i.d. $\mathcal{N}(0,1)$ entries they state (cf. [Negahban and Wainwright, 2011, Corollary 5]) that the $\|\cdot\|_2, \varepsilon$-risk of the recovery satisfies the bound

$$\mathrm{Risk}_{\|\cdot\|_2,\varepsilon}(\widehat{x}_{N,\mathrm{Lasso}}|X) \lesssim \frac{\sigma^2 r(p+q)}{N}$$

for $\varepsilon \ge \exp -(p+q)$. Observe that the above bound coincides, up to logarithmic in q and $1/\varepsilon$ factors with the second—asymptotic—term in the bound (4.37). This result

is all the more surprising if we recall that its validity is not limited to sub-Gaussian regressors—what we need in fact is the bound (cf. the remark after Proposition 4.3)

$$\mathbb{E}\left[\|\phi\,\langle\phi, z\rangle\|_*^2\right] \lesssim (p+q)\,\|x - x^*\|_\Sigma^2\,. \tag{4.38}$$

For instance, one straightforwardly verifies that the latter bound holds, for instance, in the case where regressor ϕ is a scale mixtures of matrices satisfying (4.38) (e.g., scale mixture of sub-Gaussian matrices).

4.5 Experiments

Here we present a small simulation study illustrating the performance of the SMD algorithm.

Problem statement and model description We analyze the application of the SMD-SR algorithm to sparse linear regression model of Section 4.4.1,

$$\eta_i = \phi_i^T x^* + \sigma\xi_i, \quad i = 1, 2, ..., N.$$

The signal $x^* \in \mathbb{R}^n$ is assumed to be s-sparse; we consider large-scale setting with $(N, n, s) = (100000, 100000, 50)$ of the case $N \leq n$. The random observations $(\phi_i, \xi_i)_{i=1}^N$ are assumed to be mutually independent. In our first experiments, the noise terms $(\xi_i)_{i=1}^N$ are standard Gaussian, and regressors $(\phi_i)_{i=1}^N$ are normally distributed with zero mean and covariance matrix Σ. Matrix Σ is diagonal with entries $\mu = \Sigma_{1,1} \leq \Sigma_{2,2} \leq \cdots \leq \Sigma_{n,n} = \nu$. The parameters (μ, ν) are specific for each experiment, while the rest $(\Sigma_{i,i})_{i=2}^{n-1}$ are evenly spaced in the segment $[\mu, \nu]$. The components of the optimal solution x^* are evenly spaced in $[1, n]$ with the non-zero entries being sampled from the standard Gaussian distribution. The number s of non-zero components is assumed to be known.

We consider the following stochastic optimization problem

$$\min_{x\in\mathbb{R}^n}\left\{f(x) = \frac{1}{2}\mathbb{E}\left[\left(\eta - \phi^T x\right)^2\right]\right\},$$

with inexact gradients $\nabla\tilde{f}(x, \omega) = \phi\phi^T(x - x^*) - \sigma\xi\phi$ being accessible at each iteration.

We compare the SMD-SR procedure with the SMD algorithm having the same proximal operator and with the coordinate descent algorithm (CDA) solving the Lasso problem

$$\min_{x\in\mathbb{R}^n}\left\{\frac{1}{2N}\sum_{i=1}^N\left(\eta_i - \phi_i^T x\right)^2 + \lambda\,\|x\|_1\right\}, \quad \text{with } \lambda = 2\sqrt{2}\sigma\sqrt{\frac{\log(n)}{N}}, \tag{4.39}$$

implemented in a Python package *sklearn*. The choice of λ in (4.39) is theoretically optimal according to [Bickel et al., 2009, Koltchinskii et al., 2011]. The code of the SMD algorithm is the same as the one used by the SMD-SR procedure.

Algorithm 4.3 Asymptotic phase of Algorithm 4.1, when using mini-batches of exponentially increasing sizes.

1: **Initialization:**
Set the constant step size $\beta_0 = 2\varkappa\nu$,
and the number of steps of the SMD to perform per stage

$$m_0 = \left\rfloor 16\mu^{-1}\bar{s}(8\Theta\varkappa + 1)\nu \right\lfloor,$$

and the coefficient for the mini-batch sizes $l_0 = \log(n)$.
2: **for** $k = 1, \ldots, K'$ **do stages of the asymptotic phase**
3: — Launch the SMD, initialized with y'_{k-1}, for m_0 iterations with the constant step size parameter β_0, using mini-batches with size $l_k = l_0 2^k$. Obtain an approximate solution $\widehat{x}_{m_0}(y'_{k-1}, \beta_0)$.
4: — Define y'_k as s-sparsification of $\widehat{x}_{m_0}(y'_{k-1}, \beta_0)$.
5: **end for**
6: **Output:** $\widehat{y}_N = y'_{K'}$ and $\widehat{x}_N = \widehat{x}_{m_0}(y'_{K'-1}, \beta_0)$.

Mini-batches in the asymptotic phase In our simulations, we use mini-batch implementation of the asymptotic phase of the SMD-SR algorithm. It allows to reduce the computational complexity of the method while preserving the convergence bounds of Theorem 4.1. Mini-batch technique essentially amounts to averaging stochastic gradients when staying at one point. Then the updates are made using the averaged oracle responses. We use mini-batches of exponentially increasing sizes $l_k = 2^k l_0$, where k is the index of the asymptotic phase of the algorithm. This allows to run the algorithm with the constant step size parameter $\beta_k = \beta_0$ and constant number of steps per asymptotic stage $m_k = m_0$. This might significantly reduce the computational complexity of the algorithm in the case with expensive computation of proximal mapping (4.11) and/or in a distributed setting. The updated scheme of the asymptotic phase with mini-batches is presented in Algorithm 4.3.

The theoretical guarantees for approximate solutions by the Algorithm 4.3 with mini-batches are essentially the same as those of the standard implementation given in Algorithm 4.1. We will assume that for mini-batch with size l the following holds

$$\mathbb{E}\left[\left\|\frac{1}{l}\sum_{i=1}^{l}\zeta(x,\omega_i)\right\|_*^2\right] \leq C\varsigma_*^2/l$$

for every $x \in X$ with the factor $C = C(n)$ depending on problem dimension.[6] Let us then set up $l_0 = \rfloor C \lfloor$.

Workarounds for SMD-SR The choice of algorithm parameters given in Section 4.4.1 is derived from the theoretical worst-case perspective and is typically very conservative in practice. We give a brief overview of the workarounds used in our simulations.

6. Note that the factor $C(n) = 1$ in the case of Gaussian regressors and is always bounded with $\mathcal{O}(\log(n))$. For instance, $C(n) \sim \log n$ in the case of Rademacher and bounded regressors.

— First of all, the number of steps m_0 to be performed by the SMD algorithm on each preliminary stage is taken as $m_0 = \rfloor(1/2)s\nu(\log(n)+1)\lfloor$, which coincides with (4.18) in the case of $\mu = 1.0$.
— In our simulations, we apply the CUSUM test for monitoring a change detection [Ploberger and Krämer, 1992, Lee et al., 2003] to define the switching point between the preliminary and the asymptotic phase of the algorithm. In any case, we perform at least 4 preliminary stages: if we pass to the asymptotic phase slightly lately, the SMD-SR procedure is fast to regain the pace even if there is an unstable behavior at the end of the preliminary phase.
— We use mini-batches instead of the exponentially increasing m_k and β_k. This allows to significantly accelerate the method at the asymptotic regime alleviating the computational burden of expensive prox-evaluations.
— Our step size strategy is to use variable step sizes $\beta_i = \beta_0 \|\phi_i\|_\infty^2$ with a constant factor $\beta_0 = 1.0$ both for SMD-SR and SMD. The optimal choice of β_0 was made in accordance with the condition $\beta_0 \geq \nu$, neglecting the constants derived in the theoretical analysis. In order to compute the current approximate solution, the estimates of the SMD algorithm are then weighted according to the corresponding step sizes β_i.

Setups We conduct comparisons of the SMD-SR procedure with the SMD algorithm and with the coordinate descent algorithm for the Lasso version of sparse linear regression problem. First, we consider four setups, each corresponding to a specific pair (μ, σ), where μ belongs to $\{0.1, 1.0\}$ and $\sigma \in \{0.001, 0.1\}$. We run 15 simulations for each combination of parameters. On the figures below we present the median curve along with the tube of 25% and 75% quantiles around it. For the SMD algorithm we plot both the averaged and the non-averaged solutions. Plots presented in Figure 4.1 illustrate the improvement by the SMD-SR procedure over the plain SMD algorithm in the considered settings. The acceleration of the initial error convergence is clearly seen on the plots for $\sigma = 0.001$.

We present a comparison with the CDA for Lasso in Figure 4.2 in the case of large noise, $\sigma = 10.0$. Due to memory limitations of the CDA, we consider the setup with $(N, n, s) = (10000, 50000, 50)$. The CDA is restarted for different sizes of the observation sample, each time the number of iterations of the algorithm is limited to $N = 10000$. Same as in the previous comparison, we run 15 simulations; the plots represent the median curve along with the tube of 25% and 75% quantiles.

While the coordinate descent algorithm outperforms the SMD-SR procedure for smaller values of (N, n, σ), the proposed algorithm appears to be competitive in the large noise setting. This improvement is especially notable in terms of the time of performance.

Similar results were observed when utilizing other distributions of ϕ_i and ξ_i. For instance, the results of simulations with regressors randomly drawn from the Hadamard basis in $\mathbb{R}^n$ with $n = 65536 = 2^{16}$ are presented in Figures 4.3 and 4.4.

Finally, we consider the case of heavy-tail noises—we run simulations with t_4-distribution (multivariate Student distribution with 4 degrees of freedom) of the noise and regressors. Regressors components were scaled in accordance with the diagonal entries of the covariance matrix Σ. We present the results of the simulations in Figures 4.5 and 4.6.

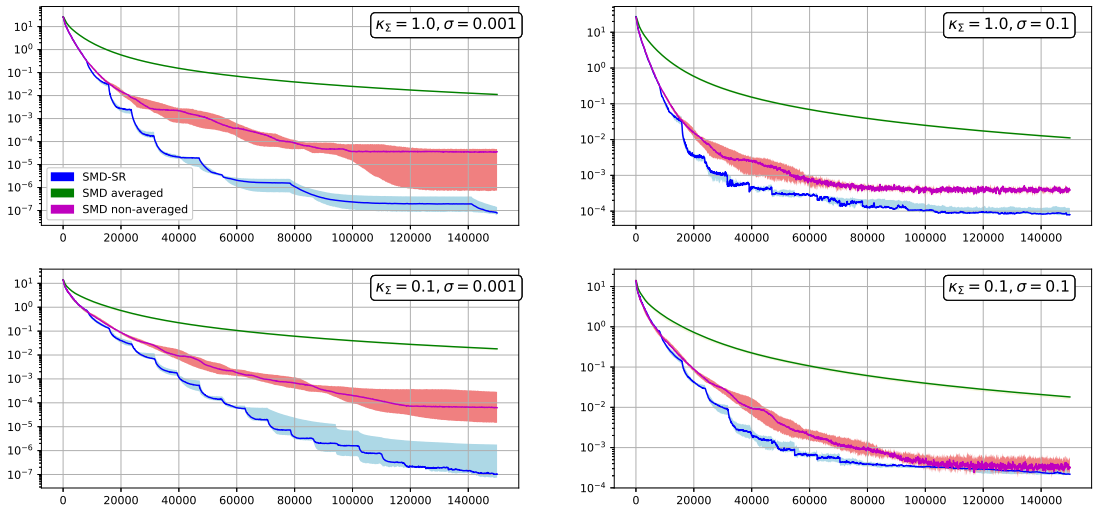


Figure 4.1 – Comparison of SMD-SR and SMD for $(N, n, s) = (150000, 100000, 50)$ in the Gaussian setting. For all the plots, the prediction error is given on a logarithmic scale, and the x-axis denotes the number of steps (oracle calls). The colored tubes represent the 25% and 75% quantiles.

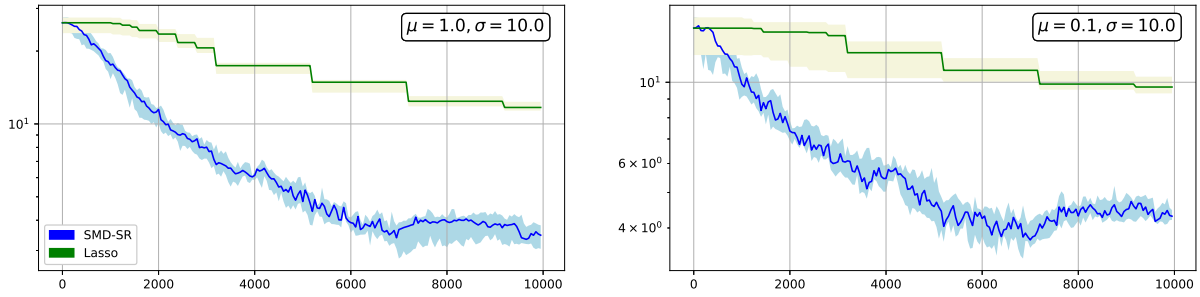


Figure 4.2 – Comparison of SMD-SR and CDA for $(N, n, s) = (10000, 50000, 50)$ on the Lasso problem in the Gaussian setting. For all the plots, the prediction error is given on a logarithmic scale, and the x-axis denotes the number of steps (oracle calls). The colored tubes represent the 25% and 75% quantiles.

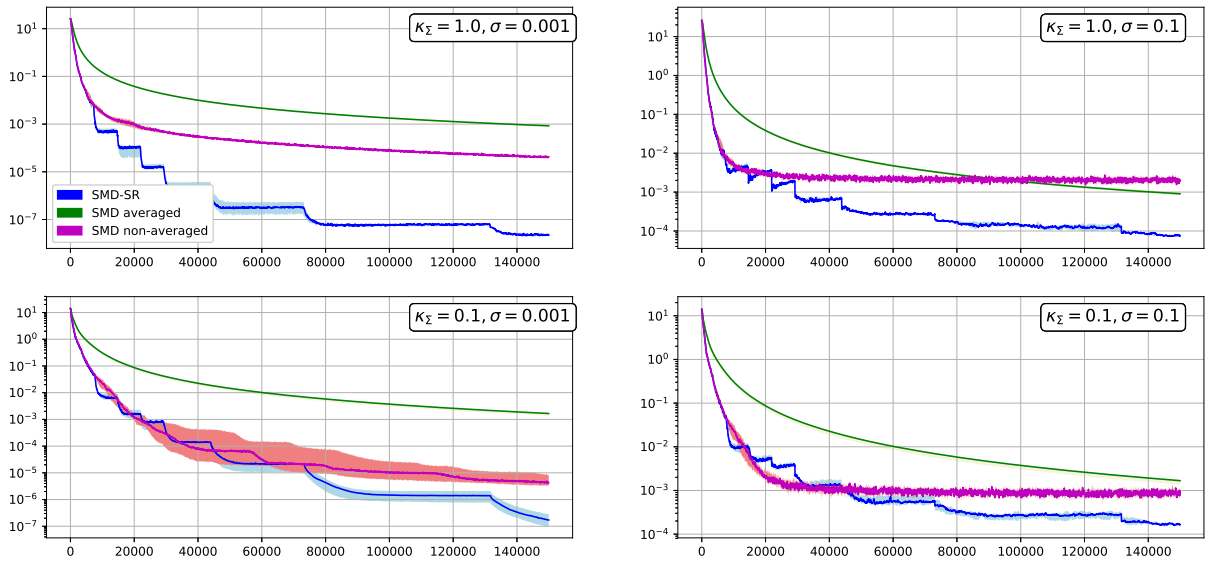


Figure 4.3 – Comparison of SMD-SR and SMD for $(N, n, s) = (150000, 65536, 50)$, when regressors ϕ_i constitute Hadamard basis and ξ is a standard Gaussian noise. For all the plots, the prediction error is given on a logarithmic scale, and the x-axis denotes the number of steps (oracle calls). The colored tubes represent the 25% and 75% quantiles.

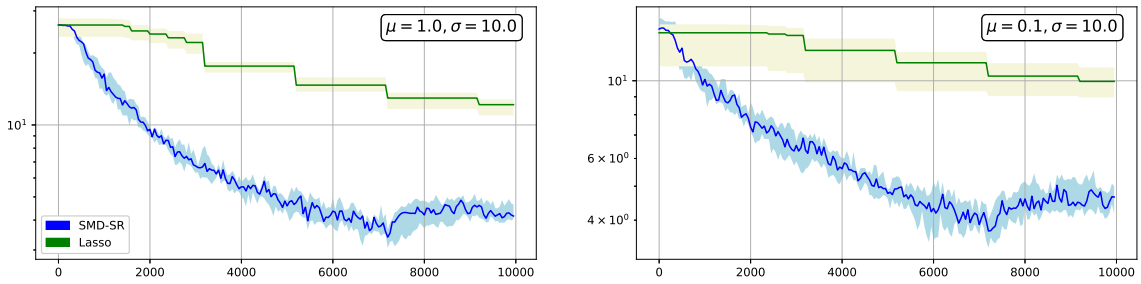


Figure 4.4 – Comparison of SMD-SR and CDA on the Lasso problem with $(N, n, s) = (10000, 65536, 50)$, when regressors ϕ_i constitute Hadamard basis and ξ is standard Gaussian noise. For all the plots, the prediction error is given on a logarithmic scale, and the x-axis denotes the number of steps (oracle calls). The colored tubes represent the 25% and 75% quantiles.

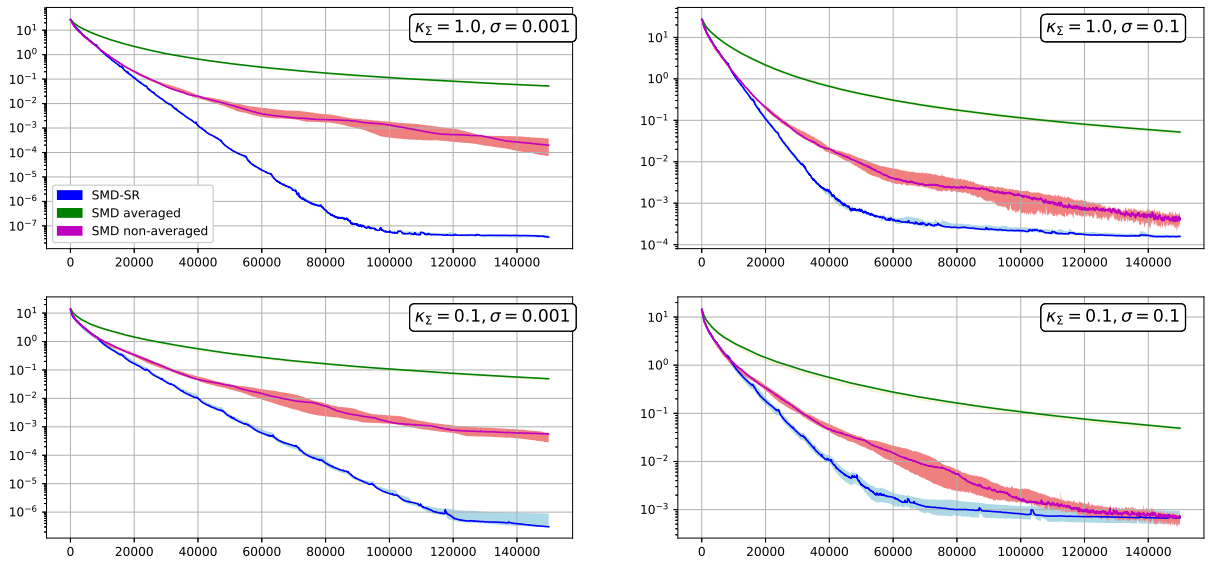


Figure 4.5 – Comparison of SMD-SR and SMD for $(N, n, s) = (150000, 100000, 50)$ in the setting of t_4-distributed noise. For all the plots, the prediction error is given on a logarithmic scale, and the x-axis denotes the number of steps (oracle calls).

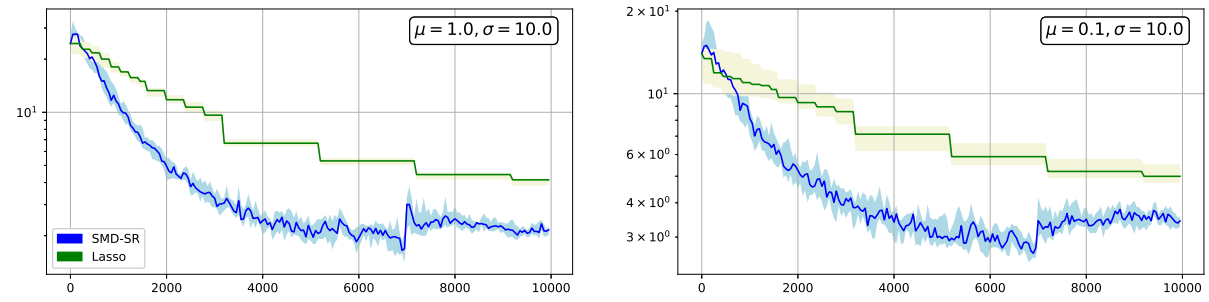


Figure 4.6 – Comparison of SMD-SR and CDA on the Lasso problem with $(N, n, s) = (10000, 50000, 50)$ in the setting of t_4-distributed noise. For all the plots, the prediction error is given on a logarithmic scale, and the x-axis denotes the number of steps (oracle calls). The colored tubes represent the 25% and 75% quantiles.

Appendix

4.A Proof of Proposition 4.1

We start with a technical result on the SMD algorithm which we formulate in a more general setting of composite minimization. Specifically, assume that we aim at solving the problem

$$\min_{x\in X}\Big\{F(x)=\mathbb{E}_x\left[\tilde{f}(x,\omega)\right]+h(x):=f(x)+h(x)\Big\}, \tag{4.40}$$

where the set X and the function $\tilde{f}$ are as in Section 4.2.1 and h is convex and continuous. We consider a more general *composite proximal mapping:* for $\zeta\in E$, $x,x_0\in X$, and $\beta>0$ we define

$$\begin{aligned}\mathrm{Prox}_\beta(\zeta,x;x_0) :&= \operatorname*{argmin}_{z\in X}\{\langle\zeta,z\rangle+h(z)+\beta V_{x_0}(x,z)\}\\ &= \operatorname*{argmin}_{z\in X}\{\langle\zeta-\beta\vartheta'(x-x_0),z\rangle+h(z)+\beta\vartheta(z-x_0)\}\end{aligned} \tag{4.41}$$

and consider for $i=1,2,\ldots$ the stochastic mirror descent recursion (4.12). Same as before, the approximate solution after N iterations of the algorithm is defined as the weighted average of $(x_i)_{i=1}^N$ according to (4.13). Obviously, one can recover the setting of Section 4.2.2 by putting $h(x)\equiv 0$. Without loss of generality, we can assume that $x_0=0$ and denote $V(x,z)=V_0(x,z)$. Besides, we denote

$$\zeta_i=\nabla\tilde{f}(x_{i-1},\omega_i)-\nabla f(x_{i-1})$$

and

$$\varepsilon(x^N,z)=\sum_{i=1}^N\beta_{i-1}^{-1}\left(\langle\nabla f(x_{i-1}),x_i-z\rangle+h(x_i)-h(z)\right)+\frac{1}{2}V(x_{i-1},x_i), \tag{4.42}$$

with $x^N=(x_0,\ldots,x_N)$. In the sequel, we use the following result.

Proposition 4.5. *In the situation of this section, let $\beta_i \geq 2\mathcal{L}$ for all $i = 0, 1, ...,$ and let $\widehat{x}_N$ be defined in (4.13), where x_i are the iterations (4.12). Then for any $z \in X$ we have*

$$\left[\sum_{i=1}^{N} \beta_{i-1}^{-1}\right] [F(\widehat{x}_N) - F(z)] \leq \sum_{i=1}^{N} \beta_{i-1}^{-1} [F(x_i) - F(z)] \leq \varepsilon(x^N, z)$$

$$\leq V(x_0, z) - V(x_N, z) + \sum_{i=1}^{N} \left[\frac{\langle \zeta_i, z - x_{i-1}\rangle}{\beta_{i-1}} + \frac{\|\zeta_i\|_*^2}{\beta_{i-1}^2}\right] \quad (4.43)$$

$$\leq 2V(x_0, z) + \sum_{i=1}^{N} \left[\frac{\langle \zeta_i, z_{i-1} - x_{i-1}\rangle}{\beta_{i-1}} + \frac{3}{2}\frac{\|\zeta_i\|_*^2}{\beta_{i-1}^2}\right], \quad (4.44)$$

where z_i is a random vector with values in X depending only on $x_0, \zeta_1, \ldots, \zeta_i$.

Proof of Proposition 4.5. **1^o.** Let $x_0, \ldots, x_N$ be some points of X; let

$$\varepsilon_{i+1}(z) := \langle \nabla f(x_i), x_{i+1} - z\rangle + \langle h'(x_{i+1}), x_{i+1} - z\rangle + \mathcal{L}V(x_i, x_{i+1}).$$

Note that $V(x, z) \geq \frac{1}{2}\|x - z\|^2$ due to the strong convexity of $V(x, \cdot)$. Thus, by convexity of f and h and the Lipschitz continuity of ∇f we get for any $z \in X$

$$\begin{aligned} F(x_{i+1}) - F(z) &= [f(x_{i+1}) - f(z)] + [h(x_{i+1}) - h(z)] \\ &= [f(x_{i+1}) - f(x_i)] + [f(x_i) - f(z)] + [h(x_{i+1}) - h(z)] \\ &\leq [\langle \nabla f(x_i), x_{i+1} - x_i\rangle + \mathcal{L}V(x_i, x_{i+1})] + \langle \nabla f(x_i), x_i - z\rangle + h(x_{i+1}) - h(z) \\ &\leq \langle \nabla f(x_i), x_{i+1} - z\rangle + \langle h'(x_{i+1}), x_{i+1} - z\rangle + \mathcal{L}V(x_i, x_{i+1}) = \varepsilon_{i+1}(z); \end{aligned}$$

i.e., the following inequality holds for any $z \in X$:

$$F(x_{i+1}) - F(z) \leq \varepsilon_{i+1}(z). \quad (4.45)$$

2^o. Let us first prove inequality (4.43). In view of (4.41), the optimality condition for x_{i+1} in (4.12.a) has the form

$$\left\langle \nabla \tilde{f}(x_i, \omega_{i+1}) + h'(x_{i+1}) + \beta_i[\vartheta'(x_{i+1}) - \vartheta'(x_i)], z - x_{i+1}\right\rangle \geq 0, \ \ \forall\, z \in X,$$

or, equivalently,

$$\begin{aligned} \left\langle \nabla \tilde{f}(x_i, \omega_{i+1}) + h'(x_{i+1}), x_{i+1} - z\right\rangle &\leq \beta_i \langle \vartheta'(x_{i+1}) - \vartheta'(x_i), z - x_{i+1}\rangle \\ &= \beta_i \langle V'(x_i, x_{i+1}), z - x_{i+1}\rangle \\ &= \beta_i \left(V(x_i, z) - V(x_{i+1}, z) - V(x_i, x_{i+1})\right), \ \ \forall\, z \in X \end{aligned}$$ [7]

what results in

$$\begin{aligned} \langle \nabla f(x_i), x_{i+1} - z\rangle + \langle h'(x_{i+1}), x_{i+1} - z\rangle \leq \beta_i \left(V(x_i, z) - V(x_{i+1}, z) - V(x_i, x_{i+1})\right) \\ - \langle \zeta_{i+1}, x_{i+1} - z\rangle. \end{aligned} \quad (4.46)$$

7. The last equality follows from the following remarkable identity see, for instance, [Chen and Teboulle, 1993]: for any u, u' and $w \in X$

$$\langle V'(u, u'), w - u'\rangle = V(u, w) - V(u', w) - V(u, u').$$

It follows from (4.45) and condition $\beta_i \geq 2\mathcal{L}$ that

$$F(x_{i+1}) - F(z) \leq \varepsilon_{i+1}(z) \leq \langle \nabla f(x_i), x_{i+1} - z\rangle + \langle h'(x_{i+1}), x_{i+1} - z\rangle + \frac{\beta_i}{2} V(x_i, x_{i+1}).$$

Together with (4.46), this inequality implies

$$\varepsilon_{i+1}(z) \leq \beta_i \left[V(x_i, z) - V(x_{i+1}, z) - \frac{1}{2} V(x_i, x_{i+1})\right] - \langle \zeta_{i+1}, x_{i+1} - z\rangle .$$

On the other hand, due to the strong convexity of $V(x, \cdot)$ we have

$$\begin{aligned}\langle \zeta_{i+1}, z - x_{i+1}\rangle - \frac{\beta_i}{2} V(x_i, x_{i+1}) &= \langle \zeta_{i+1}, z - x_i\rangle + \langle \zeta_{i+1}, x_i - x_{i+1}\rangle - \frac{\beta_i}{2} V(x_i, x_{i+1}) \\ &\leq \langle \zeta_{i+1}, z - x_i\rangle + \frac{\|\zeta_{i+1}\|_*^2}{\beta_i}.\end{aligned}$$

Combining these inequalities, we obtain

$$F(x_{i+1}) - F(z) \leq \varepsilon_{i+1}(z) \leq \beta_i [V(x_i, z) - V(x_{i+1}, z)] - \langle \zeta_{i+1}, x_i - z\rangle + \frac{\|\zeta_{i+1}\|_*^2}{\beta_i} \quad (4.47)$$

for all $z \in X$. Dividing (4.47) by β_i and taking the sum over i from 0 to $N-1$ we obtain (4.43).

3^o. We now prove the bound (4.44). Applying Lemma 6.1 of [Nemirovski et al., 2009] with $z_0 = x_0$ we get for all $z \in X$

$$\sum_{i=1}^N \beta_{i-1}^{-1} \langle \zeta_i, z - z_{i-1}\rangle \leq V(x_0, z) + \frac{1}{2} \sum_{i=1}^N \beta_{i-1}^{-2} \|\zeta_i\|_*^2 , \quad (4.48)$$

where $z_i = \operatorname{argmin}_{z \in X}\{-\beta_{i-1}^{-1} \langle \zeta_i, z\rangle + V(z_{i-1}, z)\}$ depends only on $z_0, \zeta_1, \ldots, \zeta_i$. Further,

$$\begin{aligned}\sum_{i=1}^N \beta_{i-1}^{-1} \langle \zeta_i, z - x_{i-1}\rangle &= \sum_{i=1}^N \beta_{i-1}^{-1} [\langle \zeta_i, z_{i-1} - x_{i-1}\rangle + \langle \zeta_i, z - z_{i-1}\rangle] \leq \\ &\leq V(x_0, z) + \sum_{i=1}^N \beta_{i-1}^{-1} \langle \zeta_i, z_{i-1} - x_{i-1}\rangle + \frac{1}{2} \beta_{i-1}^{-2} \|\zeta_i\|_*^2 .\end{aligned}$$

Combining this inequality with (4.43) we arrive at (4.44). □

Proof of Proposition 4.1. Note that, by definition, $\nu \geq \mathcal{L}$ and $\varkappa \geq 1$, thus, Proposition 4.5 can be applied to the corresponding SMD recursion. When applying recursively the bound (4.43) of the proposition with $z = x^*$ and $h(x) \equiv 0$ we conclude that $\mathbb{E}[V_{x_0}(x_i, x^*)]$ is finite along with $\mathbb{E}\left[\|x_i - x^*\|^2\right]$, and so $\mathbb{E}[\langle \zeta_{i+1}, x_i - x^*\rangle] = 0$. Thus, after taking expectation we obtain

$$\begin{aligned}\sum_{i=1}^m [\mathbb{E}[f(x_i)] - f^*] &\leq \beta \mathbb{E}[V_{x_0}(x_0, x^*) - V_{x_0}(x_m, x^*)] + \beta^{-1} \sum_{i=1}^m \mathbb{E}\left[\|\zeta_i\|_*^2\right] \\ &\leq \mathbb{E}[V_{x_0}(x_0, x^*) - V_{x_0}(x_m, x^*)] \\ &\quad + \beta^{-1} \sum_{i=1}^m \left(\varkappa\nu[\mathbb{E}[f(x_{i-1}) - \langle \nabla f(x^*), x_{i-1} - x^*\rangle] - f^*] + \varkappa' \varsigma_*^2\right),\end{aligned}$$

what, thanks to convexity of F, leads to

$$\left[1-\frac{\varkappa\nu}{\beta}\right]\sum_{i=1}^{m}\mathbb{E}\left[f(x_i)-f^*\right]+\beta\mathbb{E}\left[V_{x_0}(x_m,x^*)\right]$$
$$\leq \beta\mathbb{E}\left[V_{x_0}(x_0,x^*)\right]+\frac{\varkappa\nu}{\beta}\left[\mathbb{E}\left[f(x_0)-\langle\nabla f(x^*),x_0-x^*\rangle\right]-f^*\right]+\frac{m\varkappa'\varsigma_*^2}{\beta}.$$

Because, due to the convexity of f, we have $f(\widehat{x}_m)\leq\frac{1}{m}\sum_{i=1}^m f(x_i)$ and

$$\mathbb{E}\left[f(x_0)-\langle\nabla f(x^*),x_0-x^*\rangle\right]-f^*\leq\frac{1}{2}\nu\mathbb{E}\left[\|x_0-x^*\|^2\right]\leq\frac{1}{2}\nu R^2$$

we conclude that when $\beta\geq 2\varkappa\nu$

$$\mathbb{E}\left[f(\widehat{x}_m)\right]-f^*\leq\frac{2}{m}\left(\beta\Theta R^2+\frac{\varkappa\nu}{\beta}\mathbb{E}\left[V_f(x^*,x_0)\right]\right)+\frac{2\varkappa'\varsigma_*^2}{\beta}$$
$$\leq\frac{2R^2}{m}\left(\Theta\beta+\frac{\varkappa\nu^2}{2\beta}\right)+\frac{2\varkappa'\varsigma_*^2}{\beta}$$

and we obtain (4.14). □

4.B Proof of Theorem 4.1

We start with the following straightforward result:

Lemma 4.1. *Let $x^*\in X\subset E$ be s-sparse, $x\in X$, and let $x_s=\mathrm{sparse}(x)$—an optimal solution to (4.10). We have*

$$\|x_s-x^*\|\leq\sqrt{2s}\,\|x_s-x^*\|_2\leq 2\sqrt{2s}\,\|x-x^*\|_2\,. \tag{4.49}$$

Proof. Indeed, we have

$$\|x_s-x^*\|_2\leq\|x_s-x\|_2+\|x-x^*\|_2\leq 2\,\|x-x^*\|_2$$

(recall that x^* is s-sparse). Because x_s-x^* is $2s$-sparse we have by Assumption 4.2

$$\|x_s-x^*\|\leq\sqrt{2s}\,\|x_s-x^*\|_2\leq 2\sqrt{2s}\,\|x-x^*\|_2\,.$$

□

Proof of the theorem relies upon the following characterization of the properties of approximate solutions y_k, x_k, x'_k and y'_k.

Proposition 4.6. *Under the premise of Theorem 4.1,*

(i) after k preliminary stages of the algorithm one has

$$\mathbb{E}\left[\|y_k - x^*\|^2\right] \le 2s\mathbb{E}\left[\|y_k - x^*\|_2^2\right] \le 2^{-k}R^2 + 32\frac{\varsigma_*^2\bar{s}\varkappa'}{\mu\nu\varkappa}, \tag{4.50}$$

$$\mathbb{E}\left[f(\widehat{x}_{m_0}(y_{k-1},\beta))\right] - f^* \le 2^{-k-4}\frac{\mu R_0^2}{\bar{s}} + \frac{2\varkappa'\varsigma_*^2}{\varkappa\nu}. \tag{4.51}$$

In particular, upon completion of $K = \overline{K}$ preliminary stages approximate solutions $\widehat{x}^{(1)}$ and $\widehat{y}^{(1)}$ satisfy

$$\mathbb{E}\left[\left\|\widehat{y}^{(1)} - x^*\right\|^2\right] \le 2s\mathbb{E}\left[\left\|\widehat{y}^{(1)} - x^*\right\|_2^2\right] \le 64\frac{\varsigma_*^2\bar{s}\varkappa'}{\mu\nu\varkappa}, \tag{4.52}$$

$$\mathbb{E}\left[f(\widehat{x}^{(1)})\right] - f^* \le \frac{4\varkappa'\varsigma_*^2}{\varkappa\nu}. \tag{4.53}$$

(ii) Suppose that at least one asymptotic stage is complete. Let $r_k^2 = 2^{-k}r_0^2$ where $r_0^2{=}64\frac{\varsigma_*^2\bar{s}\varkappa'}{\mu\nu\varkappa}$. Then after k stages of the asymptotic phase one has

$$\mathbb{E}\left[\|y_k' - x^*\|^2\right] \le 2s\mathbb{E}\left[\|y_k' - x^*\|_2^2\right] \le r_k^2 = 2^{-k}r_0^2, \tag{4.54}$$

$$\mathbb{E}\left[f(\widehat{x}_{m_k}(y_{k-1}',\beta))\right] - f^* \le \frac{4\varsigma_*^2\varkappa'}{\beta_k} \le 2^{-k+2}\frac{\varsigma_*^2\varkappa'}{\varkappa\nu}. \tag{4.55}$$

Proof. **1^o.** We first show that under the premise of the proposition the following relationship holds for $1 \le k \le K$:

$$\mathbb{E}\left[\|y_k - x^*\|^2\right] \le R_k^2 := \frac{1}{2}R_{k-1}^2 + \frac{16\varsigma_*\bar{s}\varkappa'}{\mu\nu\varkappa}, \;\; R_0 = R. \tag{4.56}$$

Obviously, (4.56) implies (4.50) for all $1 \le k \le K$. Observe that (4.56) clearly holds for $k = 1$. Let us now perform the recursive step $k-1 \to k$. Indeed, bound (4.14) of Proposition 4.1 implies that after m_0 iterations of the SMD with the step size parameter satisfying (4.17) and initial condition x_0 such that $\mathbb{E}\left[\|x_0 - x^*\|^2\right] \le R_{k-1}$ one has

$$\begin{aligned}\mathbb{E}\left[f(\widehat{x}_{m_0})\right] - f^* &\le \frac{2}{m_0}\left[2\Theta\varkappa\nu + \frac{\nu}{4}\right]R_{k-1}^2 + \frac{\varkappa'\varsigma_*^2}{\varkappa\nu} \\ &\le \frac{[8\Theta\varkappa+1]\nu}{2m_0}R_{k-1}^2 + \frac{\varkappa'\varsigma_*^2}{\varkappa\nu}.\end{aligned} \tag{4.57}$$

Note that when $m_0 \ge 16\mu^{-1}\bar{s}(8\Theta\varkappa+1)\nu$ we have

$$\frac{8\bar{s}}{\mu}\frac{[8\Theta\varkappa+1]\nu}{m_0} \le \frac{1}{2}.$$

Therefore, when utilizing the bound (4.49) of Lemma 4.1 we get

$$\begin{aligned}\mathbb{E}\left[\|y_k - x^*\|^2\right] \le 2\bar{s}\mathbb{E}\left[\|y_k - x^*\|^2\right] &\le 8\bar{s}\mathbb{E}\left[\|\widehat{x}_{m_0} - x^*\|_2^2\right] \le \frac{16\bar{s}}{\mu}[\mathbb{E}\left[f(\widehat{x}_{m_0})\right] - f^*] \\ &\le \frac{16\bar{s}}{\mu}\left(\frac{[8\Theta\varkappa+1]\nu}{2m_0}R_{k-1}^2 + \frac{\varkappa'\varsigma_*^2}{\varkappa\nu}\right) \le R_k^2 := \frac{1}{2}R_{k-1}^2 + \frac{16\varsigma_*^2\bar{s}\varkappa'}{\mu\varkappa\nu}\end{aligned}$$

what is (4.56). Finally, when using (4.57) along with (4.50) we obtain

$$\mathbb{E}\left[f(\widehat{x}_{m_0}(y_{k-1},\beta))\right]-f^*\le\frac{\mu R_{k-1}^2}{32\bar{s}}+\frac{\varkappa'\varsigma_*^2}{\varkappa\nu}\le 2^{-k-4}\frac{\mu R_0^2}{\bar{s}}+\frac{2\varkappa'\varsigma_*^2}{\varkappa\nu}$$

what implies (4.51). Now, (4.52) and (4.53) follow straightforwardly by applying (4.50) and (4.51) with $K=\overline{K}$.

2^o. Let us prove (4.54). Recall that at the beginning of the first stage of the second phase we have $\mathbb{E}\left[\|\bar{y}_0-x^*\|\right]\le r_0^2$. Now, let us do the recursive step, i.e., assume that (4.54) holds for some $0\le k<K'$, and let us show that it holds for $k+1$. Because $\Theta\ge 1$ and $\varkappa\ge 1$ we have $\beta_k^2\ge\frac{\varkappa\nu^2}{2\Theta}$, $k=1,...$, and, by (4.14),

$$\begin{aligned}\mathbb{E}\left[f(\widehat{x}_{m_k}(y'_{k-1},\beta_k))\right]-f^*&\le\frac{2r_{k-1}^2}{m_k}\left(\Theta\beta_k+\frac{\varkappa\nu^2}{2\beta_k}\right)+\frac{2\varkappa'\varsigma_*^2}{\beta_k}\le\frac{4\Theta\beta_k r_{k-1}^2}{m_k}+\frac{2\varkappa'\varsigma_*^2}{\beta_k}\\&\le 2^{-k}\frac{r_0^2\mu}{64\bar{s}}+2^{1-k}\frac{\varkappa'\varsigma_*^2}{\varkappa\nu}\le 2^{-k}\frac{r_0^2\mu}{16\bar{s}}\le 2^{-k+2}\frac{\varkappa'\varsigma_*^2}{\varkappa\nu}.\end{aligned}\tag{4.58}$$

Observe that

$$\mathbb{E}\left[\left\|x_{m_k}(y'_{k-1},\beta_k)-x^*\right\|_2^2\right]\le\frac{2}{\mu}[\mathbb{E}\left[f(\widehat{x}_{m_k}(y'_{k-1},\beta_k))\right]-f^*]\le 2^{-k}\frac{r_0^2}{8\bar{s}},$$

so that by Lemma 4.1

$$\mathbb{E}\left[\|y'_k-x^*\|^2\right]\le 8s\mathbb{E}\left[\left\|x_{m_k}(y'_{k-1},\beta_k)-x^*\right\|_2^2\right]\le 2^{-k}r_0^2=r_k^2,$$

and (4.54) follows. Now (4.55) is an immediate consequence of (4.54) and (4.58). □

Proof of the theorem. 1^o. Let us start with the situation where no asymptotic stage takes place. Because we have assumed that N is large enough so that at least one preliminary stage took place this can only happen when either $m_0K\ge\frac{N}{2}$ or $m_1\ge\frac{N}{2}$. Due to $m_0>1$, by (4.52) we have in the first case:

$$\mathbb{E}\left[\|y_K-x^*\|^2\right]\le R_K^2:=2^{-K}R_0^2+\frac{32\varsigma_*^2\bar{s}\varkappa'}{\mu\nu\varkappa}\le 2^{-K+1}R_0^2\le R_0^2\exp\left(-\frac{cN\mu}{\Theta\varkappa\bar{s}\nu}\right)$$

for some absolute $c>0$. Furthermore, due to (4.51) we also have in this case

$$\mathbb{E}\left[f(\widehat{x}_{m_0}(y_{K-1},\beta))\right]-f^*\le 2^{-K-4}\frac{\mu R_0^2}{\bar{s}}+\frac{2\varkappa'\varsigma_*^2}{\varkappa\nu}\le 2^{-K-3}\frac{\mu R_0^2}{\bar{s}}\le\frac{\mu R_0^2}{\bar{s}}\exp\left(-\frac{cN\mu}{\Theta\varkappa\bar{s}\nu}\right).$$

Next, $m_1\ge\frac{N}{2}$ implies that

$$\frac{\bar{s}}{\mu}\ge\frac{cN}{\Theta\nu\varkappa}\tag{4.59}$$

for some absolute constant c, so that approximate solution y_K at the end of the preliminary phase satisfies (cf. 4.52)

$$\mathbb{E}\left[\|\widehat{y}-x^*\|^2\right]\le C\frac{\varsigma_*^2\bar{s}\varkappa'}{\mu\nu\varkappa}\le C\frac{\Theta\varkappa'\varsigma_*^2\bar{s}^2}{\mu^2N}.$$

Same as above, using (4.52) and (4.59) we conclude that in this case

$$\mathbb{E}\left[f(\widehat{x})\right] - f^* \le C\frac{\varkappa'\varsigma_*^2}{\varkappa\nu} \le C\frac{\Theta\varkappa'\varsigma_*^2\bar{s}}{\mu N}.$$

2^o. Now, let us suppose that at least one stage of the asymptotic phase was completed. Applying the bound (4.54) of Proposition 4.6 we have $\mathbb{E}\left[\|y'_K - x^*\|^2\right] \le r_0^2$. When $M < N/2$, same as above, we have

$$\mathbb{E}\left[\|\widehat{y}_N - x^*\|^2\right] \le r_0^2 \le R_0^2 \exp\left(-\frac{cN\mu}{\Theta\varkappa\bar{s}\nu}\right)$$

and

$$\mathbb{E}\left[f(\widehat{x}_{m_{K'}}(y_{K'-1},\beta))\right] - f^* \le \mathbb{E}\left[f(\widehat{x})\right] - f^* \le \frac{\mu R_0^2}{\bar{s}} \exp\left(-\frac{cN\mu}{\Theta\varkappa\bar{s}\nu}\right). \tag{4.60}$$

When $M \ge N/2$, since $m_k \le C\bar{m}_k$ where $\bar{m}_k = 512\frac{\bar{s}\Theta\nu\varkappa}{\mu}2^k$ we have

$$\frac{N}{2} \le C\sum_{k=1}^{K'} \bar{m}_k \le C2^{K'+1}\bar{m}_1 \le C2^{K'}\frac{\bar{s}\Theta\nu\varkappa}{\mu}.$$

We conclude that $2^{-K'} \le C\frac{\bar{s}\Theta\nu\varkappa}{\mu N}$ so that

$$\mathbb{E}\left[\|\widehat{y}_N - x^*\|^2\right] = \mathbb{E}\left[\|\widehat{y}_{K'} - x^*\|^2\right] \le 2^{-K'}r_0^2 \le C\frac{\Theta\varkappa'\varsigma_*^2\bar{s}^2}{\mu^2 N}.$$

Finally, by (4.55),

$$\mathbb{E}\left[f(\widehat{x}_{m_{K'}}(y_{K'-1},\beta))\right] - f^* \le 2^{-K'+2}\frac{\varsigma_*^2\varkappa'}{\varkappa\nu} \le C\frac{\varsigma_*^2\bar{s}\varkappa'\Theta}{\mu N};$$

together with (4.60) this implies (4.15). □

4.C Proof of Theorem 4.2

1^o. By the Chebyshev inequality,

$$\forall \ell \quad \mathrm{Prob}\left\{\left\|\widehat{x}_M^{(\ell)} - x^*\right\|_2 \ge 2\theta_M\right\} \le \tfrac{1}{4}; \tag{4.61}$$

applying [Minsker, 2015, Theorem 3.1] we conclude that

$$\mathrm{Prob}\left\{\|\widehat{x}_{N,1-\varepsilon} - x^*\|_2 \ge 2C_\alpha\theta_M\right\} \le e^{-L\psi(\alpha,\frac{1}{4})}$$

where

$$\psi(\alpha,\beta) = (1-\alpha)\log\frac{1-\alpha}{1-\beta} + \alpha\log\frac{\alpha}{\beta} \tag{4.62}$$

and $C_\alpha = \frac{1-\alpha}{\sqrt{1-2\alpha}}$. When choosing $\alpha = \frac{\sqrt{3}}{2+\sqrt{3}}$ which corresponds to $C_\alpha = 2$ we obtain $\psi(\alpha, \frac{1}{4}) = 0.1070... > 0.1$ so that

$$\mathrm{Prob}\left\{\|\widehat{x}_{N,1-\varepsilon} - x^*\|_2 \geq 4\theta_M\right\} \leq \varepsilon$$

if $L \geq 10\log[1/\varepsilon]$. When combining this result with that of Lemma 4.1 we arrive at the theorem statement for solutions $\widehat{x}_{N,1-\varepsilon}$ and $\widehat{y}_{N,1-\varepsilon}$.

2^o. The corresponding result for $\widehat{x}'_{N,1-\varepsilon}$ and its "sparsification" $\widehat{y}'_{N,1-\varepsilon}$ is due to the following simple statement.

Proposition 4.7. *Let $0 < \alpha < \frac{1}{2}$, $|\cdot|$ be a norm on E, $z \in E$, and let z_ℓ, $\ell = 1, ..., L$ be independent and satisfy*

$$\mathrm{Prob}\{|z_\ell - z| \geq \delta\} \leq \beta$$

for some $\delta > 0$ and $\beta < \alpha$. Then for $\widehat{z}$,

$$\widehat{z} \in \operatorname*{Argmin}_{u \in \{z_1, ..., z_L\}} \sum_{\ell=1}^{L} |u - z_\ell|, \tag{4.63}$$

it holds

$$\mathrm{Prob}\{|\widehat{z} - z| \geq C'_\alpha \delta\} \leq e^{-L\psi(\alpha,\beta)}$$

with $C'_\alpha = \frac{2+\alpha}{1-2\alpha}$.

Proof. Without loss of generality we may put $\delta = 1$ and $z = 0$. Proof of the proposition follows that of [Minsker, 2015, Theorem 3.1] with Lemma 2.1 of [Minsker, 2015] replaced with the following result. □

Lemma 4.2. *Let $z_1, ..., z_L \in E$, and let $\widehat{z}$ be an optimal solution to (4.63). Let $0 < \alpha < \frac{1}{2}$, and let $|\widehat{z}| \geq C'_\alpha$. Then there exists a subset I of $\{1, ..., L\}$ of cardinality $\mathrm{card} I > \alpha L$ such that for all $\ell \in I$ $|z_\ell| > 1$.*

Proof. Let us assume that $|z_\ell| \leq 1$, $\ell = 1, ..., \bar{L}$ for $\bar{L} \geq (1-\alpha)L$. Then

$$\begin{aligned}
\sum_{\ell=1}^{L} |z_\ell - \widehat{z}| &= \sum_{\ell \leq \bar{L}} |z_\ell - \widehat{z}| + \sum_{\ell > \bar{L}} |z_\ell - \widehat{z}| \geq \bar{L}(C_\alpha - 1) + \sum_{\ell > \bar{L}} [|z_\ell| - C_\alpha] \\
&\geq \sum_{\ell \leq \bar{L}} |z_\ell| + \bar{L}(C_\alpha - 2) + \sum_{\ell > \bar{L}} |z_\ell| - (L - \bar{L})C_\alpha \\
&\geq \sum_{\ell=1}^{L} |z_\ell| + \bar{L}(C_\alpha - 2) - (L - \bar{L})C_\alpha \\
&\geq \sum_{\ell=1}^{L} |z_\ell - z_1| + \bar{L}(2C_\alpha - 2) - LC_\alpha + L - 1 > \sum_{\ell=2}^{L} |z_\ell - z_1|
\end{aligned}$$

for $\bar{L} > \frac{LC_\alpha + L - 1}{2(C_\alpha - 1)}$. We conclude that $1 - \alpha \leq \frac{C_\alpha + 1}{2(C_\alpha - 1)}$, same as $C_\alpha \leq \frac{2+\alpha}{1-2\alpha}$. □

For instance, when choosing $\alpha = 1/6$ with $C_\alpha = 13/4$, and β such that $C_\alpha/\sqrt{\beta} = 10$ we obtain $\psi(\alpha,\beta) = 0.0171...$ so that for $L = \lfloor 58.46 \log[1/\varepsilon] \rfloor$ we have $L\psi(\alpha,\beta) \geq \log[1/\varepsilon]$. Because

$$\mathrm{Prob}\left\{\left\|\widehat{x}_M^{(\ell)} - x^*\right\|_2 \geq \frac{\theta_M}{\sqrt{\beta}}\right\} \leq \beta, \quad \ell = 1, ..., L,$$

by Lemma 4.2 we conclude that

$$\mathrm{Prob}\left\{\left\|\widehat{x}'_{1-\varepsilon,N} - x^*\right\|_2 \geq 10\theta_M\right\} \leq \varepsilon,$$

implying statement of the theorem for $\widehat{x}'_{1-\varepsilon,N}$ and $\widehat{y}'_{1-\varepsilon,N}$.

3^o. The proof of the claim for solutions $\widehat{x}''_{1-\varepsilon,N}$ and $\widehat{y}''_{1-\varepsilon,N}$ follows the lines of that of [Hsu and Sabato, 2014, Theorem 4]. We reproduce it here (with improved parameters of the procedure) to meet the needs of the proof of Theorem 4.3.

Let us denote $I(\tau_M)$ the subset of $\{1, ..., L\} \cup \emptyset$ such that $f(\widehat{x}_M^{(i)}) - f^* \leq 2\tau_M$ and thus $\left\|\widehat{x}_M^{(i)} - x^*\right\|_2 \leq 2\theta_M$ for $i \in I(\tau_M)$. Assuming the latter set is nonempty we have for all $i, j \in I(\tau_M)$ $\left\|\widehat{x}_M^{(i)} - \widehat{x}_M^{(j)}\right\|_2 \leq 4\theta_M$. On the other hand, using (4.61) and independence of $\widehat{x}_M^{(i)}$ we conclude that (cf. e.g., [Lerasle and Oliveira, 2011, Lemma 23])

$$\mathrm{Prob}\left\{|I| \geq \rceil L/2 \lceil\right\} \geq \mathrm{Prob}\left\{B(L, \tfrac{1}{4}) \geq \rceil L/2 \lceil\right\} \geq 1 - \exp\left(-L\psi\left(\frac{\lceil L/2 \rceil}{L}, \frac{1}{4}\right)\right)$$

where $\lceil a \rceil$ is the largest integer strictly less than a, $B(N,p)$ is a (N,p)-binomial random variable, and $\psi(\cdot,\cdot)$ is as in (4.62). When $\varepsilon \leq \frac{1}{4}$ and $L = \lfloor 12.05 \log[1/\varepsilon] \rfloor \geq 16$ we have

$$\mathrm{Prob}\{|I| \geq \rceil L/2 \lceil\} \geq 1 - e^{-L\psi(\frac{7}{16}, \frac{1}{4})} \geq 1 - e^{-0.083L} \geq 1 - \varepsilon.$$

Therefore, if we denote $\overline{\Omega}_\varepsilon$ a subset of Ω^N such that $|I(\tau_M)| > L/2$ for $\omega^N \in \overline{\Omega}_\varepsilon$ we have $P\{\overline{\Omega}_\varepsilon\} \geq 1 - \varepsilon$. Let now $\omega^N \in \overline{\Omega}_\varepsilon$ be fixed. Observe that the optimal value $\widehat{r} = r_{\widehat{\imath}_{\rceil L/2 \lceil}}$ of (4.23) satisfies $\widehat{r} \leq 4\theta_M$, and that among $\rceil L/2 \lceil$ closest to $\widehat{x}''_{N,1-\varepsilon}$ points there is at least one, let it be $\widehat{x}_M^{(\bar{i})}$ satisfying $f(\widehat{x}_M^{(\bar{i})}) - f^* \leq 2\tau_M$ and $\left\|\widehat{x}_M^{(\bar{i})} - x^*\right\|_2 \leq 2\theta_M$. We conclude that whenever $\omega^N \in \overline{\Omega}$ one has

$$\left\|\widehat{x}''_{N,1-\varepsilon} - x^*\right\|_2 \leq \left\|\widehat{x}''_{N,1-\varepsilon} - \widehat{x}_M^{(\bar{i})}\right\|_2 + \left\|\widehat{x}_M^{(\bar{i})} - x^*\right\|_2 \leq 4\theta_M + 2\theta_M \leq 6\theta_M,$$

implying that

$$\mathrm{Prob}\left(\left\|\widehat{x}''_{N,1-\varepsilon} - x^*\right\|_2 \geq 6\theta_M\right) \leq \varepsilon$$

whenever $L = \lfloor 12.05 \log[1/\varepsilon] \rfloor$. □

4.D Proof of Theorem 4.3

1^o. Let $\omega^N \in \overline{\Omega}_{\varepsilon/2}$ defined as in 3^o of the proof of Theorem 4.2; we choose $L \geq \lfloor 12.05 \log[2/\varepsilon] \rfloor$ so that $\mathrm{Prob}\{\overline{\Omega}_{\varepsilon/2}\} \leq \varepsilon/2$. We denote $\widehat{r}$ the optimal value of (4.23); recall that $\widehat{r} \leq 4\theta_M$. Then for any $i, j \in \widehat{I}$ we have

$$\left\|\widehat{x}_M^{(i)} - \widehat{x}_M^{(j)}\right\|_2 \leq 2\widehat{r} \leq 8\theta_M, \tag{4.64}$$

and for some $\bar{i} \in \widehat{I}$ we have

$$f(\widehat{x}_M^{(\bar{i})}) - f^* \leq 2\tau_M^2 \tag{4.65}$$

where τ_M and θ_M are defined in (4.20) and (4.21) respectively. W.l.o.g. we can assume that $\widehat{x}_M^{(\bar{i})}$ is the minimizer of $f(x)$ over $\widehat{x}_M^{(i)}$, $i \in \widehat{I}$.

Let us consider the aggregation procedure. From now on all probabilities are assumed to be computed with respect to the distribution P^K of the (second) sample ω^K, conditional to realization ω^N of the first sample (independent of ω^K). To alleviate notation we drop the corresponding "conditional indices."

The proof of the theorem relies on the following statement which may be of independent interest.

Proposition 4.8. *Let $U : [0,1] \times \Omega \to \mathbb{R}$ be continuously differentiable and such that $u(t) = \mathbb{E}[U(t,\omega)]$ is finite for all $t \in [0,1]$, convex and differentiable with Lipschitz-continuous gradient:*

$$|u'(t') - u'(t)|_* \leq \mathcal{M}|t - t'|, \qquad \forall\, t, t' \in [0,1].$$

In the situation in question, let $\varepsilon \in (0, \frac{1}{4}]$, $J \geq \left\rfloor 7\log[2/\varepsilon] \right\lfloor$, and $t_i = \frac{2i-1}{2m}$, $i = 1, ..., m$. Consider the estimate

$$\widehat{v} = \operatorname*{median}_j[\widehat{v}^j], \;\; \widehat{v}^j = \frac{1}{m}\sum_{i=1}^m U'(t_i, \omega_i^j) \;\; j = 1, ..., J$$

of the difference $v = u(1) - u(0)$ using $M = mJ$ independent realizations ω_i^j, $i = 1, ..., m$, $j = 1, ..., L$. Then

$$\mathrm{Prob}\{|\widehat{v} - v| \geq \rho\} \leq \varepsilon \tag{4.66}$$

where

$$\rho = \frac{1}{4m}\left[\sqrt{2\mathcal{M}(u(1) - u_*)} + \sqrt{2\mathcal{M}(u(0) - u_*)}\right] + \frac{2}{m}\sqrt{\sum_{i=1}^m \mathbb{E}\left[[\zeta^1(t_i)]^2\right]},$$

(here and below, $\zeta^j(t_i) = U'(t_i, \omega_i^j) - u'(t_i)$ and $u_* = \min_{0 \leq t \leq 1} u(t)$).

In particular, if for $\mu \geq \mathcal{M}$

$$\mathbb{E}\left[[\zeta^1(t)]^2\right] \leq \mu(u(t) - u_*) + \varsigma^2 \tag{4.67}$$

then

$$\mathrm{Prob}\{|\widehat{v} - v| \geq \bar{\rho}\} \leq \varepsilon \tag{4.68}$$

where

$$\bar{\rho} = 2\sqrt{\frac{\mu}{m}}\left[\sqrt{u(1) - u_*} + \sqrt{u(0) - u_*}\right] + \frac{2\varsigma}{\sqrt{m}}.$$

We postpone the proof of the proposition to the end of this section. We now finish the proof of the theorem.

2^o. Denote $\widehat{v}_{ji} = \mathrm{median}_\ell[\widehat{v}^\ell_{ji}]$. For $j \in \widehat{I}$, $j \neq \bar{i}$ let $x(t) = \widehat{x}^{(j)}_M + t\left(\widehat{x}^{(\bar{i})}_M - \widehat{x}^{(j)}_M\right)$. Note that $U(t,\omega) = \tilde{f}(x(t),\omega)$ and $u(t) = f(x(t))$ satisfy the premise of Proposition 4.8 with $\mathcal{M} = r^2_{j\bar{i}}\mathcal{L}_2$ where $r_{j\bar{i}} = \left\|\widehat{x}^{(\bar{i})}_M - \widehat{x}^{(j)}_M\right\|_2$, $\mu = \chi\mathcal{L}_2 r^2_{j\bar{i}}$, and $\varsigma^2 = \chi'\varsigma^2_* r^2_{j\bar{i}}$. When applying the proposition with $\varepsilon = \varepsilon/L$, $J = L'$, and $K = mL'$ we conclude that

$$\forall j \in \widehat{I},\, j \neq \bar{i} \quad \mathrm{Prob}\left\{|\widehat{v}_{j\bar{i}} - v_{j\bar{i}}| \geq \varrho_{j\bar{i}}\right\} \leq \frac{\varepsilon}{L},$$

implying that

$$\mathrm{Prob}\left\{\max_{j\in\widehat{I},j\neq\bar{i}} |\widehat{v}_{j\bar{i}} - v_{j\bar{i}}| \geq \varrho_{j\bar{i}}\right\} \leq \frac{\varepsilon}{2} \tag{4.69}$$

where

$$\varrho_{ij} = 2r_{j\bar{i}}\sqrt{\frac{\mathcal{L}_2\chi}{m}}\left[\sqrt{f(\widehat{x}^{(i)}_M) - f^*} + \sqrt{f(\widehat{x}^{(j)}_M) - f^*}\right] + 2r_{j\bar{i}}\varsigma_*\sqrt{\frac{\chi'}{m}}.$$

Let now $\Omega'_{\varepsilon/2} \subset \Omega^K$ such that for all

$$\max_{\bar{i}\neq j\in\widehat{I}} |\widehat{v}_{j\bar{i}} - v_{j\bar{i}}| \leq \varrho_{j\bar{i}}, \quad \forall \omega^K \in \Omega'_{\varepsilon/2};$$

by (4.69) $\mathrm{Prob}\{\Omega'_{\varepsilon/2}\} \geq 1 - \varepsilon/2$.

3^o. Let us fix $\omega^K \in \Omega'_{\varepsilon/2}$; our current objective is to show that in this case the set of admissible $\widehat{x}^{(i)}_M$'s is nonempty—it contains $\widehat{x}^{(\bar{i})}_M$—and, moreover, all admissible $\widehat{x}^{(j)}_M$'s satisfy the bound $f(\widehat{x}^{(j)}_M) \leq \gamma^2(r_{\bar{i}j})$ with $\gamma(r)$ defined as in (4.26).

Let $\alpha, \beta, \tau > 0$, and let $v(\gamma) = \gamma^2 - \tau^2 - 2[\alpha(\gamma+\tau) + \beta]$; then $v(\gamma) > 0$ for $\gamma \geq \sqrt{(2\alpha+\tau)^2 + 4\beta}$. Indeed, $v(\cdot)$ being nondecreasing for $\gamma \geq \alpha$, it suffices to verify the inequality for $\gamma = \sqrt{(2\alpha+\tau)^2+4\beta}$. Because

$$2\alpha + \tau + \beta/\alpha > \sqrt{(2\alpha+\tau)^2 + 4\beta}$$

we have

$$4\alpha^2 + 4\alpha\tau + 2\beta > 2\alpha\left(\sqrt{(2\alpha+\tau)^2+4\beta} + \tau\right),$$

and

$$v(\gamma) = [(2\alpha+\tau)^2 + 4\beta] - \tau^2 - 2\alpha\left(\sqrt{(2\alpha+\tau)^2+4\beta} + \tau\right) - 2\beta > 0.$$

Applying the above observation to $\alpha = 2r_{j\bar{i}}\sqrt{\frac{\mathcal{L}_2\chi}{m}}$, $\beta = 2r_{j\bar{i}}\varsigma_*\sqrt{\frac{\chi'}{m}}$, and $\tau = \tau_M$ we conclude that whenever $f(\widehat{x}^{(j)}_M) - f^* \geq \gamma^2(r_{j\bar{i}})$

$$v_{j\bar{i}} = f(\widehat{x}^{(\bar{i})}_M) - f(\widehat{x}^{(j)}_M) \leq \tau^2_M - f(\widehat{x}^{(j)}_M) < -2\varrho_{j\bar{i}}. \tag{4.70}$$

Therefore, for $f(\widehat{x}_M^{(j)}) \geq \gamma^2(r_{j\bar{i}})$

$$\operatorname*{median}_{\ell}[\widehat{v}_{j\bar{i}}^{\ell}] - \rho_{\bar{i}j} = [\operatorname*{median}_{\ell}[\widehat{v}_{j\bar{i}}^{\ell}] - v_{j\bar{i}}] + v_{j\bar{i}} - \rho_{\bar{i}j} < \varrho_{j\bar{i}} - 2\varrho_{j\bar{i}} - \rho_{\bar{i}j} < 0 \ \ \forall\, \omega^K \in \Omega'_{\varepsilon/2}.$$

Furthermore, for $f(\widehat{x}_M^{(j)}) - f^* < \gamma^2(r_{j\bar{i}})$ we have

$$\operatorname*{median}_{\ell}[\widehat{v}_{j\bar{i}}^{\ell}] - \rho_{\bar{i}j} \leq \varrho_{\bar{i}j} - \rho_{\bar{i}j} < 0 \quad \forall\, \omega^K \in \Omega'_{\varepsilon/2},$$

and we conclude that $\widehat{x}_M^{(\bar{i})}$ is admissible.

On the other hand, whenever $f(\widehat{x}_M^{(j)}) - f^* \geq \gamma^2(r_{j\bar{i}})$ we have $v_{\bar{i}j} > 2\varrho_{\bar{i}j}$ (cf. 4.70), and

$$\operatorname*{median}_{\ell}[\widehat{v}_{\bar{i}j}^{\ell}] - \rho_{j\bar{i}} = [\operatorname*{median}_{\ell}[\widehat{v}_{j\bar{i}}^{\ell}] - v_{\bar{i}j}] + v_{\bar{i}j} - \rho_{\bar{i}j} > -\varrho_{\bar{i}j} + 2\varrho_{\bar{i}j} - \rho_{\bar{i}j} \geq 0 \ \ \forall\, \omega^K \in \Omega'_{\varepsilon/2}.$$

We conclude that $\widehat{x}_M^{(j)}$ is not admissible if $f(\widehat{x}_M^{(j)}) \geq \gamma^2(r_{j\bar{i}})$ and $\omega^K \in \Omega'_{\varepsilon/2}$.

4^o. Now we are done. So, assume that $[\omega^N, \omega^K] \in \overline{\Omega}_{\varepsilon/2} \times \Omega'_{\varepsilon/2}$ (what is the case with probability $\geq 1 - \varepsilon$). We have $r_{ij} \leq 8\theta_M$ for $i, j \in \widehat{I}$ by (4.64), and $f(x_M^{(\bar{i})}) \leq \tau_M^2$ for some admissible $\bar{i} \in \widehat{I}$ by (4.65). In this situation, all $\widehat{x}_M^{(j)}$ such that $f(\widehat{x}_M^{(j)}) - f^* \geq \gamma^2(r_{j\bar{i}})$, $j \in \widehat{I}$, are not admissible, implying that the sub-optimality of the selected solution $\overline{x}_{N+K,1-\varepsilon}$ is bounded with $\gamma^2(8\theta_M)$, thus

$$\mathrm{Risk}_{f,\varepsilon}(\overline{x}_{N+K,1-\varepsilon}|X) \leq \bar{\gamma}^2 = \gamma^2(8\theta_M).$$

The "in particular" part of the statement of the theorem can be verified by direct substitution of the corresponding values of m, θ_M, and τ_M into the expression for $\bar{\gamma}^2$. □

Proof of Proposition 4.8. Let us denote

$$\bar{v} = \mathbb{E}\left[\widehat{v}^j\right] = \frac{1}{m}\sum_{i=1}^{m} u'(t_i);$$

we have

$$|\widehat{v} - v| \leq |\widehat{v} - \bar{v}| + |\bar{v} - v|. \tag{4.71}$$

1^o. Note that

$$\widehat{v}^j - \bar{v} = \frac{1}{m}\sum_{i=1}^{m} U'(t_i, \omega_i^j) - u'(t_i) = \frac{1}{m}\sum_{i=1}^{m} \zeta^j(t_i),$$

and

$$\mathbb{E}\left[(\widehat{v}^j - \bar{v})^2\right] \leq \frac{1}{m^2}\sum_{i=1}^{m} \mathbb{E}\left[[\zeta^j(t_i)]^2\right] =: \upsilon^2.$$

By the Chebyshev inequality, $\mathrm{Prob}\{|\widehat{v}^j - \bar{v}| \geq 2\upsilon\} \leq \frac{1}{4}$, and

$$\begin{aligned}\mathrm{Prob}\{\operatorname*{median}_{j}[\widehat{v}^j] - \bar{v} \geq 2\upsilon\} &\leq \mathrm{Prob}\Big\{\sum_j 1\{\widehat{v}^j - \bar{v} \geq 2\upsilon\} \geq J/2\Big\} \\ &\leq \mathrm{Prob}\{B(J, \tfrac{1}{4}) \geq J/2\} \leq e^{-J\psi(\frac{1}{2}, \frac{1}{4})} \leq e^{-0.1438J}\end{aligned}$$

where $\psi(\cdot,\cdot)$ is defined in (4.62). Because the same bound holds for $\mathrm{Prob}\{\mathrm{median}_j[\widehat{v}^j]-\bar{v}\le -2\upsilon\}$ we conclude that

$$\mathrm{Prob}\{|\widehat{v}-\bar{v}|\ge 2\upsilon\}=\mathrm{Prob}\{|\operatorname*{median}_j[\widehat{v}^j]-\bar{v}|\ge 2\upsilon\}\le 2e^{-J/7}\le\varepsilon \tag{4.72}$$

for $J\ge 7\log(2/\varepsilon)$. Furthermore, if (4.67) holds we have

$$\mathbb{E}\left[(\widehat{v}^j-\bar{v})^2\right]\le\frac{1}{m^2}\sum_{i=1}^m\left[\mu(u(t_i)-f^*)+\varsigma^2\right]\le\frac{1}{2m}\left[(u(1)-u*)+(u(0)-u_*)\right]+\frac{\varsigma^2}{m}=:\bar{v}^2$$

implying (4.72) with υ replaced with $\bar{v}$:

$$\mathrm{Prob}\{|\widehat{v}-\bar{v}|\ge 2\bar{v}\}\le 2e^{J/7}\le\varepsilon \tag{4.73}$$

2^o. Next, we bound the difference $\bar{v}-v$.

Let $s_i=i/m$, $i=0,...,m$, and $r_i=u'(s_i)-u'(s_{i-1})$. Let us show that

$$v-\bar{v}\le\frac{1}{4m}\left[\sqrt{2\mathcal{M}(u(1)-u_*)}+\sqrt{2\mathcal{M}(u(0)-u_*)}\right].$$

Note that

$$\delta_i=\int_{s_{i-1}}^{s_i}[u'(s)-u'(t_i)]ds\le\tfrac{1}{4}r_i(s_i-s_{i-1})=(4m)^{-1}r_i,$$

so that

$$v-\bar{v}\le\sum_{i=1}^m\delta_i\le(4m)^{-1}[u'(1)-u'(0)].$$

Let now $t_*\in[0,1]$ be a minimizer of u on $[0,1]$. Due to the smoothness and convexity of u we have

$$|u'(0)-u'(t_*)|^2\le 2\mathcal{M}[u(0)-u_*+t_*u'(t_*)]\le 2\mathcal{M}[u(0)-u_*]$$

and

$$|u'(1)-u'(t_*)|^2\le 2\mathcal{M}[u(1)-u_*-(1-t_*)u'(t_*)]\le 2\mathcal{M}[u(1)-u_*].$$

We conclude that

$$u'(1)-u'(0)\le u'(1)-u'(t_*)+u'(t_*)-u'(0)\le\sqrt{2\mathcal{M}[u(0)-u_*]}+\sqrt{2\mathcal{M}[u(1)-u_*]},$$

and

$$v-\bar{v}\le(4m)^{-1}[u'(1)-u'(0)]\le\frac{1}{4m}\left[\sqrt{2\mathcal{M}[u(0)-u_*]}+\sqrt{2\mathcal{M}[u(1)-u_*]}\right].$$

The proof of the corresponding bound for $\bar{v}-v$ is completely analogous, implying that

$$|v-\bar{v}|\le\frac{1}{4m}\left[\sqrt{2\mathcal{M}(u(1)-u_*)}+\sqrt{2\mathcal{M}(u(0)-u_*)}\right].$$

When substituting the latter bound and the bound (4.72) into (4.71) we obtain

$$\mathrm{Prob}\{|\widehat{v}-v|\ge 2\upsilon+\upsilon'\}\le\varepsilon$$

for $J\ge 7\log(2/\varepsilon)$, what implies (4.66). When replacing (4.72) with (4.73) in the above derivation we obtain (4.68). □

4.E Proofs for Section 4.4.2

The following statement is essentially well known:

Lemma 4.3. *Let $\phi \in \mathbb{R}^{p\times q}$ with $q \le p$ for the sake of definiteness, be a random sub-Gaussian matrix $\phi \sim \mathcal{SG}(0, S)$ implying that*

$$\forall x \in \mathbb{R}^{p\times q}, \;\; \mathbb{E}\left[e^{\langle x,\phi\rangle}\right] \le e^{\frac{1}{2}\langle x,S(x)\rangle}. \tag{4.74}$$

Suppose that $S \preceq \bar{s}I$; then

$$\mathbb{E}\left[\|\phi\|_*^2\right] \le C\bar{s}(p+q) \quad \textit{and} \quad \mathbb{E}\left[\|\phi\|_*^4\right] \le C'\bar{s}^2(p+q)^2$$

where C and C' are absolute constants.

Proof of the lemma.

1^o. Let $u \in \mathbb{R}^q$ be such that $\|u\|_2 = 1$. Then the random vector $\zeta = \phi u \in \mathbb{R}^p$ is sub-Gaussian with $\zeta \sim \mathcal{SG}(0, Q)$, that is for any $v \in \mathbb{R}^p$

$$\mathbb{E}\left[e^{v^T\zeta}\right] = \mathbb{E}\left[e^{v^T\phi u}\right] = \mathbb{E}\left[e^{\langle uv^T,\phi\rangle}\right] \le e^{\frac{1}{2}\langle uv^T,S(uv^T)\rangle} = e^{\frac{1}{2}v^TQv}$$

where $Q = Q^T \in \mathbb{R}^{p\times p}$. Note that

$$\max_{\|v\|_2=1} v^TQv = \max_{\|v\|_2=1} \left\langle uv^T, S(uv^T)\right\rangle \le \max_{\|w\|_2=1} \langle w, S(w)\rangle .$$

Therefore, we have $Q \preceq \bar{s}I$, and $\mathrm{Tr}(Q) \le \bar{s}p$.

2^o. Let $\Gamma = \{u \in \mathbb{R}^q : \|u\|_2 = 1\}$, and let $\mathcal{D}_\varepsilon$ be a minimal ε-net, w.r.t. $\|\cdot\|_2$, in Γ, and let $\mathcal{N}_\varepsilon$ be the cardinality of $\mathcal{D}_\varepsilon$. We claim that

$$\left\{u^T\phi^T\phi u \le \upsilon \;\forall u \in \mathcal{D}_\varepsilon\right\} \;\Rightarrow\; \left\{\left\|\phi^T\phi\right\|_* \le (1-2\varepsilon)^{-1}\upsilon\right\}. \tag{4.75}$$

Indeed, let the premise in (4.75) hold true; $\phi^T\phi$ is symmetric, so let $\bar{v} \in \Gamma$ be such that $\bar{v}^T\phi^T\phi\bar{v} = \left\|\phi^T\phi\right\|_*$. There exists $u \in \mathcal{D}_\varepsilon$ such that $\|\bar{v} - u\|_2 \le \varepsilon$, whence

$$\left\|\phi^T\phi\right\|_* = |\bar{v}^T\phi^T\phi\bar{v}| \le 2\left\|\phi^T\phi\right\|_* \|\bar{v} - u\|_2 + |u^T\phi^T\phi u| \le 2\left\|\phi^T\phi\right\|_* \varepsilon + \upsilon$$

(note that the quadratic form z^TQz is Lipschitz continuous on Γ, with constant $2\|Q\|_*$ w.r.t. $\|\cdot\|_2$), whence $\left\|\phi^T\phi\right\|_* \le (1-2\varepsilon)^{-1}\upsilon$.

3⁰. We can straightforwardly build an ε-net $\mathcal{D}'$ in Γ in such a way that the $\|\cdot\|_2$-distance between every two distinct points of the net is $> \varepsilon$, so that the balls $B_v = \{z \in \mathbb{R}^p : \|z - v\|_2 \le \varepsilon/2\}$ with $v \in \mathcal{D}'$ are mutually disjoint. Since the union of these balls belongs to $B = \{z \in \mathbb{R}^q : \|z\|_2 \le 1 + \varepsilon/2\}$, we get $\mathrm{Card}(\mathcal{D}')(\varepsilon/2)^q \le (1 + \varepsilon/2)^q$, that is, $\mathcal{N}_\varepsilon \le \mathrm{Card}(\mathcal{D}') \le (1 + 2/\varepsilon)^q$.

Now we need the following well-known result (we present its proof at the end of this section for the sake of completeness).

Lemma 4.4. *Let $\zeta \sim \mathcal{SG}(0, Q)$ be a sub-Gaussian random vector in $\mathbb{R}^n$, i.e.*

$$\forall t \in \mathbb{R}^n \quad \mathbb{E}\left[e^{t^T\zeta}\right] \le e^{\frac{1}{2}t^TQt} \tag{4.76}$$

where $Q = Q^T \in \mathbb{R}^{n\times n}$. Then for all $x \ge 0$

$$\mathrm{Prob}\left\{\|\zeta\|_2^2 \ge \mathrm{Tr}(Q) + 2\sqrt{xv} + 2x\bar{q}\right\} \le e^{-x} \tag{4.77}$$

where $\bar{q} = \max_i \sigma_i(Q)$ is the principal eigenvalue of Q and $v = \|Q\|_2^2 = \sum_i \sigma_i^2(Q)$ is the squared Frobenius norm of Q. Thus, for any $\alpha > 0$

$$\mathrm{Prob}\left\{\|\zeta\|_2^2 \ge \mathrm{Tr}(Q)(1 + \alpha^{-1}) + (2 + \alpha)x\bar{q}\right\} \le e^{-x}. \tag{4.78}$$

Utilizing (4.78) with $\alpha = 1$ we conclude that $\forall u \in \Gamma$ the random vector $\zeta = \phi u$ satisfies

$$\mathrm{Prob}\left\{\|\zeta\|_2^2 \ge 2\bar{s}p + 3\bar{s}x\right\} \le e^{-x}. \tag{4.79}$$

Let us set $\varepsilon = \frac{1}{4}$; utilizing (4.79), we conclude that the probability of violating the premise in (4.75) with $\upsilon = 2\bar{s}p + 3\bar{s}x$ does not exceed $\exp\left(-x + q\log[1 + 2\varepsilon^{-1}]\right) = \exp\left(-x + q\log 9\right)$, so that

$$\mathrm{Prob}\left\{\left\|\phi^T\phi\right\|_* \ge 2\bar{s}(2p + 3x)\right\} \le \exp\left(-x + q\log 9\right).$$

Finally, we obtain

$$\begin{aligned}
\mathbb{E}\left[\|\phi\|_*^4\right] = \mathbb{E}\left[\left\|\phi^T\phi\right\|_*^2\right] &= 2\int_0^\infty \mathrm{Prob}\left\{\left\|\phi^T\phi\right\|_* \ge u\right\} u\,du \\
&\le 2\int_0^\infty u \min\left\{\exp\left(\frac{4\bar{s}p - u}{6\bar{s}} + q\log 9\right), 1\right\} du \\
&\le 2\int_0^{\bar{s}(4p+6q\log 9)} u\,du + 2\int_{\bar{s}(4p+6q\log 9)}^\infty u\exp\left(\frac{4\bar{s}p - u}{6\bar{s}} + q\log 9\right) du \\
&\le \bar{s}^2(4p + 6q\log 9)^2 + 12\bar{s}^2(4p + 6q\log 9) + 72\bar{s}^2 \le C'\bar{s}^2(p + q)^2.
\end{aligned}$$

Similarly we get $\mathbb{E}\left[\|\phi\|_*^2\right] \le C\bar{s}(p + q)$ for an appropriate C.

4ᵒ. Let us now prove Lemma 4.4.

Note that for $t < 1/(2\bar{s})$ and $\eta \in \mathbb{R}^n$, $\eta \sim \mathcal{N}(0, I)$ independent of ζ we have by (4.74)

$$\begin{aligned}
\mathbb{E}\left[e^{t\langle\zeta,\zeta\rangle}\right] = \mathbb{E}\left[\mathbb{E}_\eta\left[e^{\sqrt{2t}\langle\zeta,\eta\rangle}\right]\right] = \mathbb{E}_\eta\left[\mathbb{E}\left[e^{\sqrt{2t}\langle\zeta,\eta\rangle}\right]\right] &\le \mathbb{E}_\eta\left[e^{t\langle\eta,S\eta\rangle}\right] = \mathbb{E}_\eta\left[e^{t\langle\eta,D\eta\rangle}\right] \\
&= \prod_i \mathbb{E}_{\eta_i}\left\{e^{t\eta_i^2 s_i}\right\} = \prod_i (1 - 2ts_i)^{-1/2}
\end{aligned}$$

where $D = \mathrm{Diag}(s_i)$ is the diagonal matrix of eigenvalues. Recall that one has, cf. [Birgé et al., 1998, Lemma 8],

$$-\frac{1}{2}\log(1-2ts_i) - ts_i \le \frac{t^2 s_i^2}{1-2ts_i} \le \frac{t^2 s_i^2}{1-2t\bar{s}}$$

for $t < 1/(2\bar{s})$. On the other hand, $\forall t < 1/(2\bar{s})$

$$\begin{aligned}\mathrm{Prob}\left\{\|\zeta\|_2^2 - \mathrm{Tr}(S) \ge u\right\} &\le \mathbb{E}\left[\exp\left(t\left[\|\zeta\|_2^2 - \sum_i s_i - u\right]\right)\right] \\ &\le \exp\left(-tu + \frac{t^2}{1-2t\bar{s}}\sum_i s_i^2\right) = \exp\left(-tu + \frac{t^2 v}{1-2t\bar{s}}\right).\end{aligned}$$

When choosing $t = \frac{\sqrt{x}}{v+2\bar{s}\sqrt{x}}$ $\left(< \frac{1}{2\bar{s}}\right)$ and $u = 2\sqrt{xv} + 2x\bar{s}$ we obtain

$$\mathrm{Prob}\left(\|\zeta\|_2^2 \ge \mathrm{Tr}(S) + 2\sqrt{xv} + 2x\bar{s}\right) \le e^{-x}$$

what is (4.77). Because $v \le \mathrm{Tr}(S)\bar{s}$ the latter bound also implies (4.78). □

Chapter 5

Conclusion and Perspectives

In this thesis, we discussed several frameworks addressing the issues of robustness and acceleration of iterative algorithms in stochastic optimization.

Our first contribution consists of the development of a framework which provides a unified way for analyzing many incremental approaches, including several variance-reduced algorithms. First, this technique allows to make all these methods robust to stochastic noise. Second, it also naturally allows to build new algorithms with theoretical guarantees which are similar to those obtained for the existing methods. Finally, we have introduced an accelerated stochastic gradient descent algorithm and a new accelerated SVRG algorithm that is robust to stochastic noise. All of the developed algorithms support non-uniform sampling strategies. We have also developed versions of these algorithms for the strongly convex problems, which are adaptive to the value of the strong convexity parameter.

In our second contribution, we provide a solution to the problem of generic acceleration generalizing the multi-stage approach called Catalyst, which was aimed originally to accelerate deterministic methods. We improve its flexibility with respect to the choice of surrogate functions minimized at different stages and successfully address the acceleration of the stochastic methods, including the stochastic variance-reduced algorithms introduced in the first part of this thesis. Besides this, we provide a unified analysis which sheds new lights on the problem of robustness of stochastic algorithms when the proximal operator cannot be exactly computed.

Finally, we have focused on stochastic sparse optimization, and our third contribution amounts to development of a multi-stage procedure that effectively exploits the sparsity of a problem. This procedure has theoretical guarantees, which are much better in terms of the variance compared to the standard algorithms relying on Euclidean geometry. It improves on the best known existing solutions in several aspects, including linear convergence at initial iterations and enlarged class of considered models.

Perspectives The developed schemes and frameworks provide several perspectives for future work.

Given the framework developed in Chapter 2, the first interesting application, which was not covered in the thesis, amounts to considering optimization problems with a *drifting optimum*. Specifically, time-varying optimization problems require algorithms which converge to a changing solution by catching the objective drift over time. This general setting, though notably not the most frequent in the literature, can be found in various applications, see [Simonetto, 2017, Hall and Willett, 2015, Dall'Anese et al., 2020] and references therein. This setting was especially well studied in the field of *online learning* that typically deals with regret minimization [Hazan, 2019, Shalev-Shwartz et al., 2012, Wilson et al., 2018]. Another popular criterion to be minimized is called *dynamic regret* [Shahrampour and Jadbabaie, 2017], and we can potentially address its optimization in the large scale setting within our frameworks. The motivation lies in the simple intuitive observation that the drifts could be seen as inexact computation of gradients. Moreover, the construction of estimate sequences in non-composite cases is essentially quadratic, so that the drift could be easily treated. Intuitively, we can expect that the time-varying setting can be treated in a framework similar to the one developed in Chapter 2. Both frameworks of Chapters 2 and 3 could be used to generalize existing time-varying first-order optimization methods to the composite optimization setting.

Unfortunately, consideration of variance-reduced methods, like SVRG, SAGA, SDCA, Finito or MISO, in the time-varying setting does not seem to be productive. The reason for it lies in their common motivation, which was explained after Equation (2.11), which is to define a new gradient estimate $Z = X - Y + \mathbb{E}[Y]$ which has the same expectation as X but potentially a lower variance if Y is positively correlated with X. When there is a drift in the objective, the positive correlation between Y and X is hard to maintain, because it suffers from a quadratic dependency on the dimension p of the problem.

The framework developed in Chapter 4 may be useful in the analysis of the time-varying setting as well. Although we do not state it explicitly, the SMD-SR algorithm is applicable to optimization problems with *approximately* sparse solutions. Therefore, although the drift of the solution might spoil the sparsity, we will still be able to address such cases through approximate sparse optimization, for instance, when the drift is bounded in the ℓ_2-norm. If successful, the SMD-SR algorithm could also potentially decrease the influence of the drift by constraining it only to nonvanishing components of the optimum, being especially useful in the high-dimensional setting.

We may also raise the question of generalization of the results of other chapters to the non-Euclidean setting. This question might be especially relevant in the context of Chapter 3, because the corresponding accelerating framework does not depend on the inner structure of the base method, which thus might be non-Euclidean one. then, we could apply the SMD-SR algorithm from Chapter 4, which itself does not depend on the inner construction of the base method. This would lead to a natural application of variance-reduced approaches to stochastic sparse optimization.

The variance-reduced algorithms can be analyzed in the context of online learning algorithms. Although this idea has already appeared in [Cutkosky and Orabona, 2019, Cutkosky and Busa-Fekete, 2018], it might be interesting to analyze this setting using the unified framework of Chapter 2 to provide a joint treatment to various variance-reduced algorithms at once.

In Chapter 2, we were mainly interested in the case of bounded variance σ. However,

the proposed algorithm might be used in the case of unbounded variance of the gradient estimations.

In Chapter 3 we have provided the generic acceleration framework with theoretical guarantees that suffer from an extra logarithmic factor in the condition number. The task of getting rid from this factor for SAGA/MISO/SDCA/Finito is open (the problem of acceleration of the SVRG approach was solved in Chapter 2). We believe that estimate sequences may be useful to obtain the optimal complexity without this logarithmic term, but the construction would be non-trivial and would rely on a different lower bound than the one we used in Section 2.4.

Note that the original Catalyst approach [Lin et al., 2015] was successfully generalized to quasi-Newton methods [Lin et al., 2019]. It is an open question if the analysis of 3 could provide a generalization of [Lin et al., 2019] to the stochastic case.

Bibliography

A. Agarwal, M. J. Wainwright, P. L. Bartlett, and P. K. Ravikumar. Information-theoretic lower bounds on the oracle complexity of convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012a.

Alekh Agarwal, Sahand Negahban, and Martin J Wainwright. Stochastic optimization and sparse statistical recovery: Optimal algorithms for high dimensions. In *Advances in Neural Information Processing Systems 25*, pages 1538–1546. Curran Associates, Inc., 2012b.

Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2017, pages 1200–1205. ACM, 2017.

Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. *arXiv preprint arXiv:1407.1537*, 2014.

Zeyuan Allen-Zhu and Yang Yuan. Improved SVRG for non-strongly-convex or sum-of-non-convex objectives. In *International conference on machine learning*, pages 1080–1089, 2016.

Yossi Arjevani and Ohad Shamir. Dimension-free iteration complexity of finite sum optimization problems. In *Advances in Neural Information Processing Systems*, pages 3540–3548, 2016.

Hilal Asi and John C Duchi. Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM Journal on Optimization*, 29(3):2257–2290, 2019.

Necdet Serhat Aybat, Alireza Fallah, Mert Gurbuzbalaban, and Asuman Ozdaglar. A universally optimal multistage accelerated stochastic gradient method. In *Advances in Neural Information Processing Systems (NIPS)*, pages 8525–8536, 2019.

Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. *CoRR*, abs/1306.2119, 2013.

M. Baes. Estimate sequence methods: extensions and approximations. *ETH technical report*, 2009.

Pierre Baldi. Gradient descent learning algorithm overview: A general dynamical systems perspective. *IEEE Transactions on neural networks*, 6(1):182–195, 1995.

Rina Foygel Barber and Wooseok Ha. Gradient descent with non-convex constraints: local concavity determines convergence. *Information and Inference: A Journal of the IMA*, 7 (4):755–806, 03 2018.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Stephen Becker, Jérôme Bobin, and Emmanuel J Candès. Nesta: A fast and accurate first-order method for sparse recovery. *SIAM Journal on Imaging Sciences*, 4(1):1–39, 2011.

Dimitri P Bertsekas. Incremental proximal methods for large scale convex optimization. *Mathematical Programming*, 129(2):163, 2011.

Kush Bhatia, Prateek Jain, and Purushottam Kar. Robust regression via hard thresholding. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

Peter J Bickel, Ya'acov Ritov, Alexandre B Tsybakov, et al. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

A. Bietti and J. Mairal. Stochastic optimization with variance reduction for infinite datasets with finite-sum structure. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

Lucien Birgé, Pascal Massart, et al. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998.

Doron Blatt, Alfred O Hero, and Hillel Gauchman. A convergent incremental gradient method with a constant step size. *SIAM Journal on Optimization*, 18(1):29–51, 2007.

Thomas Blumensath and Mike E Davies. Iterative hard thresholding for compressed sensing. *Applied and computational harmonic analysis*, 27(3):265–274, 2009.

L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *arXiv preprint arXiv:1405.4980*, 2014.

Sébastien Bubeck, Yin Tat Lee, and Mohit Singh. A geometric alternative to Nesterov's accelerated gradient descent. *arXiv preprint arXiv:1506.08187*, 2015.

Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

James V Burke and Michael C Ferris. A Gauss—Newton method for convex composite optimization. *Mathematical Programming*, 71(2):179–194, 1995.

Emmanuel Candes. Compressive sampling. In *Proceedings of the International Congress of Mathematicians*, volume 3, pages 1433–1452. Madrid, August 22-30, Spain, 2006.

Emmanuel Candes. The restricted isometry property and its implications for compressed sensing. *Comptes rendus de l'Académie des Sciences, Mathématique*, 346(9-10):589–592, 2008.

Emmanuel Candes, Terence Tao, et al. The dantzig selector: Statistical estimation when p is much larger than n. *The annals of Statistics*, 35(6):2313–2351, 2007.

Emmanuel J Candes and Yaniv Plan. A probabilistic and ripless theory of compressed sensing. *IEEE transactions on information theory*, 57(11):7235–7254, 2011a.

Emmanuel J Candes and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011b.

Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717, 2009.

Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(8): 1207–1223, 2006.

Emmanuel J Candès, Yaniv Plan, et al. Near-ideal model selection by ℓ_1 minimization. *The Annals of Statistics*, 37(5A):2145–2177, 2009.

A. Chambolle and T. Pock. A remark on accelerated block coordinate descent for computing the proximity operators of a sum of convex functions. *SMAI Journal of Computational Mathematics*, 1:29–54, 2015.

Gong Chen and Marc Teboulle. Convergence analysis of a proximal-like minimization algorithm using bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, 1993.

Kai Lai Chung. On a stochastic approximation method. *The Annals of Mathematical Statistics*, pages 463–483, 1954.

Michael B Cohen, Jelena Diakonikolas, and Lorenzo Orecchia. On acceleration with noise-corrupted gradients. In *Proceedings of the International Conferences on Machine Learning (ICML)*, 2018.

Ashok Cutkosky and Róbert Busa-Fekete. Distributed stochastic optimization via adaptive SGD. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex sgd. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Arnak Dalalyan and Philip Thompson. Outlier-robust estimation of a sparse linear model using ℓ_1-penalized Huber's M-estimator. In *Advances in Neural Information Processing Systems*, pages 13188–13198, 2019.

Emiliano Dall'Anese, Andrea Simonetto, Stephen Becker, and Liam Madden. Optimization and learning with information streams: Time-varying algorithms and applications. *IEEE Signal Processing Magazine*, 37(3):71–83, 2020.

Alexandre d'Aspremont. Smooth optimization with approximate gradient. *SIAM Journal on Optimization*, 19(3):1171–1183, October 2008. ISSN 1052-6234.

Jon Dattorro. *Convex optimization & Euclidean distance geometry*. Lulu. com, 2010.

A. Defazio, F. Bach, and S. Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems (NIPS)*, 2014a.

A. Defazio, T. Caetano, and J. Domke. Finito: A faster, permutable incremental gradient method for big data problems. In *Proceedings of the International Conferences on Machine Learning (ICML)*, 2014b.

Frederik Michel Dekking, Cornelis Kraaikamp, Hendrik Paul Lopuhaä, and Ludolf Erwin Meester. *A Modern Introduction to Probability and Statistics: Understanding why and how*. Springer Science & Business Media, 2005.

Olivier Devolder et al. Stochastic first order methods in smooth convex optimization. Technical report, CORE, 2011.

Aymeric Dieuleveut, Nicolas Flammarion, and Francis Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *Journal of Machine Learning Research (JMLR)*, 18:101:1–101:51, 2017.

David L Donoho et al. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS math challenges lecture*, 1(2000):32, 2000.

Dmitriy Drusvyatskiy, Maryam Fazel, and Scott Roy. An optimal first order method based on optimal quadratic averaging. *SIAM Journal on Optimization*, 28(1):251–271, 2018.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(Jul): 2121–2159, 2011.

Torbjørn Eltoft, Taesu Kim, and Te-Won Lee. On the multivariate Laplace distribution. *IEEE Signal Processing Letters*, 13(5):300–303, 2006.

Maryam Fazel, E Candes, Benjamin Recht, and P Parrilo. Compressed sensing and robust recovery of low rank matrices. In *2008 42nd Asilomar Conference on Signals, Systems and Computers*, pages 1043–1047. IEEE, 2008.

Rina Foygel Barber and Haoyang Liu. Between hard and soft thresholding: optimal iterative thresholding algorithms. *Information and Inference: A Journal of the IMA*, 12 2019.

Alexander Vladimirovich Gasnikov and Yu E Nesterov. Universal method for stochastic composite optimization problems. *Computational Mathematics and Mathematical Physics*, 58(1):48–64, 2018.

S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization I: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization II: Shrinking procedures and optimal algorithms. *SIAM Journal on Optimization*, 23(4):2061–2089, 2013.

Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1-2):267–305, 2016.

O. Güler. New proximal point algorithms for convex minimization. *SIAM Journal on Optimization*, 2(4):649–664, 1992.

Eric C Hall and Rebecca M Willett. Online convex optimization in dynamic environments. *IEEE Journal of Selected Topics in Signal Processing*, 9(4):647–662, 2015.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., 2001.

Elad Hazan. Introduction to online convex optimization. *arXiv preprint arXiv:1909.05207*, 2019.

Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization. *J. Mach. Learn. Res.*, 15:2489–2512, 2010.

J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex analysis and minimization algorithms. II.* Springer, 1996.

T. Hofmann, A. Lucchi, S. Lacoste-Julien, and B. McWilliams. Variance reduced stochastic gradient descent with neighbors. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

Daniel Hsu and Sivan Sabato. Heavy-tailed regression with a generalized median-of-means. In *International Conference on Machine Learning*, pages 37–45, 2014.

C. Hu, W. Pan, and J. T. Kwok. Accelerated gradient methods for stochastic optimization and online learning. In *Advances in Neural Information Processing Systems (NIPS)*. 2009.

Anatoli Iouditski and Yuri Nesterov. Primal-dual subgradient methods for minimizing uniformly convex functions. *arXiv preprint arXiv:1401.1792*, 2014.

Prateek Jain, Ambuj Tewari, and Purushottam Kar. On iterative hard thresholding methods for high-dimensional m-estimation. In *Advances in Neural Information Processing Systems*, pages 685–693, 2014.

R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

A Juditsky and A Nemirovski. First order methods for nonsmooth convex large-scale optimization, I: general purpose methods. In S Sra, S Nowozin, and S J Wright, editors, *Optimization for Machine Learning*. MIT Press Cambridge, 2011a.

Anatoli Juditsky and Arkadi Nemirovski. Accuracy guarantees for ℓ_1-recovery. *IEEE Transactions on Information Theory*, 57(12):7818–7839, 2011b.

Anatoli Juditsky and Yuri Nesterov. Primal-dual subgradient methods for minimizing uniformly convex functions. 2010.

Anatoli Juditsky and Yuri Nesterov. Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization. *Stochastic Systems*, 4(1):44–80, 2014.

Anatoli Juditsky, Fatma Kılınç Karzan, and Arkadi Nemirovski. On a unified view of nullspace-type conditions for recoveries associated with general sparsity structures. *Linear Algebra and its Applications*, 441:124–151, 2014.

Anatoli Juditsky, Alexander Nazin, Arkadi Nemirovsky, and Alexandre Tsybakov. Algorithms of robust stochastic optimization based on mirror descent method. *Automation and Remote Control*, 80(9):1607–1627, 2019.

Anatoli B. Juditsky and Arkadi Nemirovski. Accuracy guarantees for l_1-recovery. *IEEE Trans. Information Theory*, 57:7818–7839, 2011c.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *preprint arXiv:1412.6980*, 2014.

Anton J Kleywegt, Alexander Shapiro, and Tito Homem-de Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12(2):479–502, 2002.

Vladimir Koltchinskii, Karim Lounici, Alexandre B Tsybakov, et al. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

Samuel Kotz and Saralees Nadarajah. *Multivariate t-distributions and their applications*. Cambridge University Press, 2004.

Dmitry Kovalev, Samuel Horváth, and Peter Richtárik. Don't jump through hoops and remove those loops: SVRG and Katyusha are better without the outer loop. In *Algorithmic Learning Theory*, pages 451–467, 2020.

Brian Kulis and Peter L Bartlett. Implicit online learning. In *Proceedings of the International Conferences on Machine Learning (ICML)*, 2010.

Simon Lacoste-Julien, Mark Schmidt, and Francis Bach. A simpler approach to obtaining an $O(1/t)$ convergence rate for the projected stochastic subgradient method. *arXiv preprint arXiv:1212.2002*, 2012.

G. Lan and Y. Zhou. An optimal randomized incremental gradient method. *Mathematical Programming*, 171(1–2):167–215, 2018a.

G. Lan and Y. Zhou. Random gradient extrapolation for distributed and stochastic optimization. *SIAM Journal on Optimization*, 28(4):2753–2782, 2018b.

Guanghui Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1-2):365–397, 2012.

Guillaume Lecué, Shahar Mendelson, et al. Regularization and the small-ball method i: sparse recovery. *The Annals of Statistics*, 46(2):611–641, 2018.

Guillaume Lecué, Matthieu Lerasle, et al. Robust machine learning by median-of-means: theory and practice. *Annals of Statistics*, 48(2):906–931, 2020.

Sangyeol Lee, Jeongcheol Ha, Okyoung Na, and Seongryong Na. The cusum test for parameter change in time series models. *Scandinavian Journal of Statistics*, 30(4): 781–796, 2003.

Matthieu Lerasle and Roberto I Oliveira. Robust empirical mean estimators. *arXiv preprint arXiv:1112.3914*, 2011.

Guoyin Li and Ting Kei Pong. Global convergence of splitting methods for nonconvex composite optimization. *SIAM Journal on Optimization*, 25(4):2434–2460, 2015.

Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A universal catalyst for first-order optimization. In *Advances in neural information processing systems*, pages 3384–3392, 2015.

Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. Catalyst acceleration for first-order convex optimization: from theory to practice. *Journal of Machine Learning Research (JMLR)*, 18(212):1–54, 2018.

Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. An inexact variable metric proximal point algorithm for generic quasi-Newton acceleration. *SIAM Journal on Optimization*, 29(2):1408–1443, 2019.

Q. Lin, X. Chen, and J. Peña. A sparsity preserving stochastic gradient methods for sparse regression. *Computational Optimization and Applications*, 58(2):455–482, 2014.

Jeff Linderoth, Alexander Shapiro, and Stephen Wright. The empirical behavior of sampling methods for stochastic programming. *Annals of Operations Research*, 142(1):215–241, 2006.

L. Liu, Yanyao Shen, Tianyang Li, and C. Caramanis. High dimensional robust sparse regression. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

Liu Liu, Tianyang Li, and Constantine Caramanis. High dimensional robust estimation of sparse models via trimmed hard thresholding. *ArXiv*, abs/1901.08237, 2019.

Po-Ling Loh and Martin J Wainwright. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *The Journal of Machine Learning Research*, 16(1):559–616, 2015.

Z. Lu and L. Xiao. On the complexity analysis of randomized block-coordinate descent methods. *Mathematical Programming*, 152(1):615–642, 2015.

J. Mairal. Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization*, 25(2):829–855, 2015.

J. Mairal. End-to-end kernel learning with supervised convolutional kernel networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

J. Mairal, F. Bach, and J. Ponce. Sparse modeling for image and vision processing. *Foundations and Trends in Computer Graphics and Vision*, 8(2-3):85–283, 2014.

N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

Stanislav Minsker. Geometric median and robust estimation in banach spaces. *Bernoulli*, 21(4):2308–2335, 2015.

J.-J. Moreau. Fonctions convexes duales et points proximaux dans un espace hilbertien. *CR Acad. Sci. Paris Sér. A Math*, 255:2897–2899, 1962.

J.-J. Moreau. Proximité et dualité dans un espace hilbertien. *Bulletins de la Socitété Mathématique de France*, 93(2):273–299, 1965.

Tomoya Murata and Taiji Suzuki. Sample efficient stochastic gradient iterative hard thresholding method for stochastic sparse linear regression with limited attribute observation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Ion Necoara and Andrei Patrascu. A random coordinate descent algorithm for optimization problems with composite objective function and linear coupled constraints. *Computational Optimization and Applications*, 57(2):307–337, 2014.

Sahand Negahban and Martin J Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, pages 1069–1097, 2011.

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, Bin Yu, et al. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, January 2009.

A. Nemirovsky and D. Yudin. Problem complexity and method efficiency in optimization. 1983.

Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Publishing Company, Incorporated, 2014.

Y. Nesterov and B. T. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

Yu Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

Yu Nesterov and J-Ph Vial. Confidence level solutions for stochastic programming. *Automatica*, 44(6):1559–1568, 2008.

Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. In *Doklady an ussr*, volume 269, pages 543–547, 1983.

Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120:221–259, 2009.

Yurii Nesterov and Arkadi Nemirovski. On first-order algorithms for ℓ_1/nuclear norm minimization. *Acta Numerica*, 22:509–575, 2013.

L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *Proceedings of the International Conferences on Machine Learning (ICML)*, 2017a.

Lam M Nguyen, Phuong Ha Nguyen, Marten van Dijk, Peter Richtárik, Katya Scheinberg, and Martin Takáč. SGD and Hogwild! convergence without the bounded gradients assumption. In *Proceedings of the International Conferences on Machine Learning (ICML)*, 2018.

Nam Nguyen, Deanna Needell, and Tina Woolf. Linear convergence of stochastic iterative greedy algorithms with sparse constraints. *IEEE Transactions on Information Theory*, 63:6869–6895, 2014.

Nam Nguyen, Deanna Needell, and Tina Woolf. Linear convergence of stochastic iterative greedy algorithms with sparse constraints. *IEEE Transactions on Information Theory*, 63(11):6869–6895, 2017b.

Courtney Paquette, Hongzhou Lin, Dmitriy Drusvyatskiy, Julien Mairal, and Zaid Harchaoui. Catalyst acceleration for gradient-based non-convex optimization. *preprint arXiv:1703.10993*, 2018.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):127–239, 2014.

Werner Ploberger and Walter Krämer. The CUSUM test with ols residuals. *Econometrica: Journal of the Econometric Society*, pages 271–285, 1992.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Optimization*, 30(4):838–855, July 1992.

B.T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1 – 17, 1964.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over l_q-balls. *IEEE Transactions on Information Theory*, 57:6976–6994, 2009.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated Gaussian designs. *Journal of Machine Learning Research*, 11(Aug):2241–2259, 2010.

Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *Ann. Math. Statist.*, 22(3):400–407, 09 1951.

Mark Rudelson and Shuheng Zhou. Reconstruction from anisotropic random measurements. In *Conference on Learning Theory, Workshop and Conference Proceedings*, volume 23, pages 10.1–10.28, 2012.

Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

Jerome Sacks. Asymptotic distribution of stochastic approximation procedures. *The Annals of Mathematical Statistics*, 29(2):373–405, 1958.

M. Schmidt, R. Babanezhad, M. Ahmed, A. Defazio, A. Clifton, and A. Sarkar. Non-uniform stochastic average gradient method for training conditional random fields. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015.

M. Schmidt, N. Le Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017.

Mark Schmidt, Nicolas L Roux, and Francis R Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In *Advances in neural information processing systems*, pages 1458–1466, 2011.

Hanie Sedghi, Anima Anandkumar, and Edmond A. Jonckheere. Multi-step stochastic admm in high dimensions: Applications to sparse optimization and matrix decomposition. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

Shahin Shahrampour and Ali Jadbabaie. Distributed online optimization in dynamic environments using mirror descent. *IEEE Transactions on Automatic Control*, 63(3): 714–725, 2017.

S. Shalev-Shwartz. SDCA without duality, regularization, and individual convexity. In *Proceedings of the International Conferences on Machine Learning (ICML)*, 2016.

S. Shalev-Shwartz and T. Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *Mathematical Programming*, 155(1):105–145, 2016.

Shai Shalev-Shwartz and Ambuj Tewari. Stochastic methods for ℓ_1-regularized loss minimization. *Journal of Machine Learning Research*, 12(Jun):1865–1892, 2011.

Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

Alexander Shapiro. Monte carlo sampling methods. *Handbooks in operations research and management science*, 10:353–425, 2003.

Alexander Shapiro and Arkadi Nemirovski. On complexity of stochastic programming problems. In *Continuous optimization*, pages 111–146. Springer, 2005.

Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. A proximal gradient algorithm for decentralized composite optimization. *IEEE Transactions on Signal Processing*, 63(22): 6013–6023, 2015.

Andrea Simonetto. Time-varying convex optimization via time-varying averaged operators. *arXiv: Optimization and Control*, 2017.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems 23*, pages 2199–2207. Curran Associates, Inc., 2010.

N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research (JMLR)*, 15(1):1929–1958, 2014.

Jacob Steinhardt, Stefan Wager, and Percy Liang. The statistics of streaming sparse regression. *arXiv preprint arXiv:1412.4182*, 2014.

Panos Toulis, Dustin Tran, and Edo Airoldi. Towards stability and optimality in stochastic gradient descent. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2016.

Panos Toulis, Thibaut Horel, and Edoardo M Airoldi. Stable Robbins-Monro approximations through stochastic proximal updates. *preprint arXiv:1510.00967*, 2018.

Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization*, 2(3), 2008.

Sara Van De Geer and Peter Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

V. Vapnik. *The nature of statistical learning theory*. Springer, 2000.

M. J. Wainwright, M. I. Jordan, and J. C. Duchi. Privacy aware learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

Craig Wilson, Venugopal V Veeravalli, and Angelia Nedić. Adaptive sequential stochastic optimization. *IEEE Transactions on Automatic Control*, 64(2):496–509, 2018.

L. Xiao and T. Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

Lin Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research (JMLR)*, 11(Oct):2543–2596, 2010.

Lin Xiao and Tong Zhang. A proximal-gradient homotopy method for the sparse least-squares problem. *SIAM Journal on Optimization*, 23(2):1062–1091, 2013.

Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, pages 116–. ACM, 2004.

Yuchen Zhang, Martin J. Wainwright, and Michael I. Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *COLT*, 2014.

Yun-Bin Zhao and Zhi-Quan Luo. Analysis of optimal thresholding algorithms for compressed sensing. *arXiv preprint arXiv:1912.10258*, 2019.

S. Zheng and J. T. Kwok. Lightweight stochastic optimization for minimizing finite sums with infinite data. In *Proceedings of the International Conferences on Machine Learning (ICML)*, 2018.

S. Zheng, Y. Song, T. Leung, and I. Goodfellow. Improving the robustness of deep neural networks via stability training. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

K. Zhou. Direct acceleration of SAGA using sampled negative momentum. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

K. Zhou, F. Shang, and J. Cheng. A simple stochastic variance reduced algorithm with fast convergence rates. In *Proceedings of the International Conferences on Machine Learning (ICML)*, 2018.