Reduced complexity algorithms for high-dimensional statistics

Sasila Ilandarideva

▶ **To cite this version:**

Sasila Ilandarideva. Reduced complexity algorithms for high-dimensional statistics. Statistics [stat]. Université Grenoble Alpes [2020-..], 2024. English. NNT : 2024GRALM024 . tel-04865884

HAL Id: tel-04865884
https://theses.hal.science/tel-04865884v1

Submitted on 6 Jan 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : MSTII - Mathématiques, Sciences et technologies de l'information, Informatique
Spécialité : Mathématiques Appliquées
Unité de recherche : Laboratoire Jean Kuntzmann

Algorithmes à complexité réduite pour statistiques à haute dimension

Reduced complexity algorithms for high-dimensional statistics.

Présentée par :

Sasila ILANDARIDEVA

Direction de thèse :

Anatoli IOUDITSKI PROFESSEUR DES UNIVERSITES, UNIVERSITE GRENOBLE ALPES	Directeur de thèse
Vianney PERCHET PR, ENSAE	Co-directeur de thèse

Rapporteurs :

ANTONIN CHAMBOLLE
DIRECTEUR DE RECHERCHE, CNRS DELEGATION PARIS CENTRE
NIAO HE
ASSISTANT PROFESSOR, ECOLE POLYTECHNIQUE FEDERALE DE ZURICH

Thèse soutenue publiquement le **14 juin 2024**, devant le jury composé de :

JERÔME LELONG, PROFESSEUR DES UNIVERSITES, UNIVERSITE GRENOBLE ALPES	Président
ANATOLI IOUDITSKI, PROFESSEUR DES UNIVERSITES, UNIVERSITE GRENOBLE ALPES	Directeur de thèse
VIANNEY PERCHET, PROFESSEUR, ENSAE PARIS	Co-directeur de thèse
ANTONIN CHAMBOLLE, DIRECTEUR DE RECHERCHE, CNRS DELEGATION PARIS CENTRE	Rapporteur
NIAO HE, ASSISTANT PROFESSOR, ECOLE POLYTECHNIQUE FEDERALE DE ZURICH	Rapporteure
ALEXANDRE D'ASPREMONT, DIRECTEUR DE RECHERCHE, CNRS DELEGATION PARIS CENTRE	Examinateur
JERÔME MALICK, DIRECTEUR DE RECHERCHE, CNRS DELEGATION ALPES	Examinateur
GUANGHUI LAN, FULL PROFESSOR, GEORGIA INSTITUTE OF TECHNOLOGY	Examinateur

Remerciements

Je tiens, dans cette section, à remercier toutes les personnes qui m'ont aidé et soutenu durant ces trois années de thèse. C'est, arrivé au bout du chemin, que je me rend compte la chance que j'ai pu avoir d'être entouré de personnes aussi bienveillantes. Toute les personnes cités dans ces remerciements ont contribué à leur manière à la réussite de cette thèse.

Mes remerciements vont tout d'abord à mes directeurs de thèse. Anatoli, ce n'est que maintenant que je me rend compte de la chance que j'ai pu avoir d'être encadré par un chercheur aussi renommé que vous. Je me souviens comme si c'était hier de notre première discussion pour déterminer la direction de nos recherches, vous m'aviez prévenu ce jour là que ça allait être difficile. Effectivement, difficile, ça l'était, mais sans votre encadrement tout ce travail aurait été impossible. Vous m'avez laissé beaucoup de liberté pour découvrir ce métier de chercheur tout en sachant me donner les connaissances nécessaires pour réussir. Le domaine dans lequel vous m'avez introduit est passionnant et j'espère que par ce manuscrit j'ai pu restituer un peu du savoir que vous m'avez transmis. J'ai beaucoup appris à vos côtés et je vous remercie énormément de m'avoir fait confiance. Vianney, tu m'as aidé à démarrer ma thèse dans les meilleures condition possible malgré la situation de confinement de l'époque. Je me souviens à quel point c'était difficile le début de thèse à distance, mais tu m'as très vite intégré dans ton équipe et c'était un énorme plaisir pour moi de travailler avec de si brillants jeune chercheurs. Je gardes aussi les bons souvenirs des match de foot que l'on a pu faire ensemble avec l'équipe de l'ENSAE et tous les décembres à Luminy. J'espère que l'on pourra remettre ça un jour !

I would like to thank Niao He and Antonin Chambolle for their insightful comments and feedback on this manuscript. It has been a pleasure for me to know that you appreciated my work. I am also very grateful to Jérôme Lelong, Jérôme Malick, Alexandre d'Aspremont and Guanghui Lan for accepting being part of the jury for my thesis defense. It is a source of pride to know that my work has been validated by such eminent researchers as yourselves.

Je voulais également avoir quelques mots pour la géniale équipe de co-thésards et post-docs de Vianney avec lesquels j'ai pu passer de très bons moments durant ces trois années. Merci donc à Flore, Evrard, Reda, Lorenzo, Maria, Côme, Mike, Mathieu, Corentin et Hugo. Merci pour tous ces reading groups et les super goûters qui les suivaient.
Il y a également toutes les personnes que j'ai rencontré au CREST puis que j'ai pu retrouver chaque fin d'année lors du CIRM ou pendant les conférences. Ainsi je tenais à remercier Arnak, Clara, Hugo, Etienne, Evgenii, Arshak, Amir, Julien, Arya, Nicolas et Nayel.

Cette thèse a été pour moi l'occasion de découvrir la magnifique ville de Grenoble et de rencontrer des personnes tout aussi formidable. Je voulais particulièrement remercier Yannis et Alexis. Yannis, sans ton soutiens et nos discussions sur nos recherches respectives, je n'aurais clairement jamais pu

arriver au bout de cette thèse. J'ai beaucoup appris grâce à toi et j'espère que j'ai pu t'apporter autant que ce que tu m'as apporté. Alexis, je te remercie pour les moments de rire que nous avons pu passer au labo ensemble. Je me rappelerai pendant longtemps de tous nos débats autant futile que drôle et nos longues discussion pendant les pauses cafés. Merci de de m'avoir fait découvrir ta belle région. Je vous ai rencontré comme collègue de bureau et je vais quitter Grenoble avec deux frères de plus.
Je voulais aussi remercier mes amis de l'équipe d'optimisation du LJK, merci à Waïss, Victor, Tam, Iéva, Sélim, Florian, Gilles et Yassine. J'ai beaucoup apprécié les moments passés avec vous, les discussions et les débats au CROUS. Vous m'avez plus qu'aidé en cette fin de thèse et je vous en suis très reconnaissant. Je suis persuadé que chacuns de vous à un brillant futur dans la recherche qui vous attends. I would also like to thank you Tianjiao, we first met for a project on zoom and I am glad that we could meet in person during your stay in France. It has been a real pleasure to collaborate with you and wish you the best for your upcoming adventures!
Il y a également tous les autres permanents, post-docs, doctorants ou stagiaire du labo, merci à Sergei, Boris, Jérôme, Kliment, J.B., Margaux, Rémy, Dmitrii, Grégoire, Alexis, Maya, Teddy, Anastasia, Hélène et Alexandre. Merci aussi à l'équipe administrative du LJK, plus particulièrement à Elyssa, Aristide, Cathy, Laurence et Aurore. Votre bonne humeure et vos sourires ont rendu mon passage au LJK plus qu'agréable.

Je tiens à remercier tous mes amis de longue date que j'ai pu rencontrer tout au long de ma vie. Votre amitié à été un moteur de ma réussite.

Tout d'abords je voulais avoir quelques mots pour mes amis d'école, merci à Youssef O., Youssef J., Maxime, Fabiola, Farouk, Mahdi, Ryan, Moha, Elouan, Yazid, Soukaïna, Ghita, Raouf, Abbe et Jordan. Je suis vraiment heureux de vous avoir rencontré et d'avoir pu passer de nombreux moment inoubliables à vos côtés. Fabiola et Elouan je vous remercie énormément pour tous les moments drôle que l'on a pu avoir dans cette fameuse PhDColoc. Je garderai pendant longtemps en mémoire nos soirée SNK, les repas que l'on se preparait ou bien encore de nos dure séance à la salle de sport. La première partie de ma thèse est passé à une vitesse folle grâce à vous. Mahdi et Farouk vous avez toujours été là quand j'avais besoin de vous. Merci pour tous les bon moments passé ensemble. Merci à toi Yazid de m'avoir conseillé de faire une thèse, tu as toujours été de bons conseils. J'espère qu'un jour on trouvera un sujet de collaboration !
J'ai également rencontré de belles personnes pendant mes années en prépa, je vous remercie pour ces discussions interminable à l'internat, nos session de travail intenses dans les salles de kholles et pour tous les autres moments passés ensemble, merci à vous Marouane, Nelson, Othmane, Ramzi et Youcef.
Enfin, Inès, Lisa, Ihab, Nioro, Sohaïb, Robin L., Robin K., Sofiane R. et Sofiane T. Je voulais vous remercier pour avoir sû préserver notre amitié malgré la distance et les années. Je suis fier de savoir que je peux compter sur des personnes aussi admirable que vous. J'espère que notre amitié continuera encore dans le temps.

Je tiens à exprimer toute ma gratitude à mes parents, qui m'ont toujours soutenu dans mes choix et m'ont encouragé tout au long de mes études. Je suis profondément conscient des sacrifices que vous avez faits pour moi, et je vous en serai éternellement reconnaissant. Vous m'avez transmis les valeurs les plus précieuses de la vie, et je vous dois tout. I would also like to thank my family in Sri Lanka, I am aware that I have not been able to see you for a long time but you have always been in my heart and in my thoughts.

Je souhaite dédier ces dernières lignes à la femme qui m'a accompagné depuis tant d'année et qui m'a toujours soutenu dans les moments les plus importants de ma vie. Ma chère Marion, je te remercie pour ton amour, ta patience et ta joie de vivre. Partager une longue partie de ma vie à tes côtés a été un privilège pour moi. Je t'ai vu grandir, évoluer et atteindre tes rêves. Je suis tellement fier de la brillante avocate que tu es devenue. J'espère que tu le seras au moins autant de moi. Malgré les hauts et les bas, tu as toujours été pour moi, une source d'inspiration et de force. Si aujourd'hui je suis arrivé au bout de cette thèse c'est notamment grâce à toi.

Résumé

L'objectif de cette thèse est d'étudier des algorithmes d'optimisation stochastique de premier ordre appliquées à la résolution de problème de récupération statistique en haute dimension. Plus précisément, nous examinerons la mise en œuvre d'algorithmes d'approximation stochastique multi-étape basés sur une stratégie de redémarrage. Lorsqu'ils sont appliqués au problème de récupération parcimonieuse, les estimateurs fournis par nos méthodes stochastiques doivent vérifier des bornes statistiques optimales sur la performance de l'estimation, généralement mesurée par le risque d'estimation ou le risque de prédiction, avec pour objectif final de supprimer les limitations existantes des algorithmes connus. Les algorithmes doivent être robustes aux distributions à queue lourde des observations bruitées et aux observations erronées des régresseurs. Les algorithmes étudiés devront s'adapter à une implémentation parallélisée, leurs performances numériques devant évoluer en fonction de la mémoire disponible ou des ressources de traitement et de l'architecture.

Dans la première partie de ce manuscrit, nous développerons et analyseront une méthode multi-étapes basée sur l'algorithme Stochastique de Descente Miroir basé sur une divergence de Bregman utilisant une fonction potentiel posédant une structure géométrique non Euclidienne. Nous fournirons des bornes pour les larges déviations sous l'hypothèse que le bruit stochastique suit une distribution sous-Gaussienne. Cette méthode est étendue avec des garanties théoriques sous une hypothèse de convexité uniforme autour de l'optimum, incluant également une analyse de robustesse de l'algorithme vis à vis de ses paramètres inconnus basée sur la procédure d'adaptation de Lepski.

Dans la seconde partie, nous explorons une variante de la méthode du gradient accéléré de Nesterov, qui utilise une estimation du vrai gradient par moyennation des gradients stochastiques. Nous montrerons que cette méthode accelerée atteint à la fois la complexité optimale en terme d'itération mais également la complexité optimale en terme d'échantillon, tout celà, sous l'hypothèse d'un bruit stochastique dependant du point de recherche en cours. Notre analyse comprend des vitesses de convergences valide en espérance, ainsi que des bornes pour de larges déviations en présence de bruit stochastic suivant des distributions sous-exponentielle. Cette méthode est ensuite adaptée en une procédure multi-étapes, ciblant spécifiquement les problèmes de récupération de vecteurs parcimonieux en grande dimension.

Tout au long de ce manuscrit, l'efficacité des méthodes proposées est constamment évaluée dans le contexte des modèles de régression linéaire généralisée parcimonieuse, un problème important dans la récupération statistique.

Abstract

The aim of this thesis is to investigate the properties of stochastic optimization procedures applied to high-dimensional statistical recovery problems. Specifically, we concentrate on the implementation of multistage algorithms of stochastic approximation methods. These algorithms are expected to meet optimal statistical bounds regarding estimation or prediction risks thus addressing the limitations of existing algorithms. They should demonstrate robustness to heavy-tailed distributions in observation disturbances and erroneous observations of regressors. Additionally, their design allows heavily parallelized implementation, allowing numerical performance to scale with the available memory, processing resources, and architecture.

In the first part of this work we analyze a multistage method based on a non-Euclidean Stochastic Mirror Descent algorithm. We provide theoretical bounds for large deviations under sub-Gaussian stochastic noise assumption along with numerical validation of the method. This algorithm is extended with theoretical guarantees under uniform convexity assumption around the optimum, including robustness analysis with respect to unknown parameters based on the celebrated Lepski's adaptation procedure.

In the second part, we explore a special version of the accelerated gradient algorithm that employs mini-batch gradient estimation. We demonstrate that this accelerated method achieve both the optimal iteration and optimal sample complexity under the state-dependent noise assumption. Our analysis encompasses bounds for estimation error and prediction error that are valid in expectation as well as for large deviations in the presence of sub-exponential gradient noise. This method is further utilized in multistage procedures designed to solve sparse recovery problems.

Throughout this manuscript, the effectiveness of the proposed methods is consistently evaluated in the context of the sparse Generalized Linear Regression model, an important problem in statistical recovery.

Contents

Remerciements **5**

Résumé **7**

Abstract **9**

Outline **15**

I Foundations **17**

1 Introduction **19**
1.1 Optimization and Machine Learning 19
1.1.1 Mathematical background on Machine Learning 19
1.1.2 Optimization 22
1.2 Stochastic First-order Methods 26
1.2.1 Stochastic Mirror Descent 26
1.2.2 Accelerated Methods 29
1.3 Sparse Recovery 31
1.3.1 Signal recovery basics 32
1.3.2 Sparse Recovery and Stochastic Optimization. 35
1.4 Contributions 38

II Stochastic Mirror Descent and Sparse Recovery **43**

2 Stochastic Mirror Descent for Large Scale Sparse Recovery **45**
2.1 Introduction 45
2.2 Multistage Stochastic Mirror Descent for Sparse Stochastic Optimization 47
2.2.1 Problem statement 47
2.2.2 Composite Stochastic Mirror Descent algorithm 48
2.2.3 Main contribution: a multistage adaptive algorithm 49
2.3 Sparse generalized linear regression by stochastic approximation 51
2.3.1 Problem setting 51
2.3.2 Stochastic Mirror Descent algorithm 54
2.3.3 Numerical experiments 55
2.4 Appendix 58
2.4.1 Proof of Proposition 2.2.1 58

2.4.2 Deviation inequalities . . . 62
2.4.3 Proof of Theorem 2.2.1 . . . 65
2.4.4 Proof of Proposition 2.3.1 . . . 69
2.4.5 Properties of sparsity structures . . . 70
2.4.6 Supplementary numerical experiments . . . 73

3 Extensions **77**
3.1 Adaptive CSMD-SR via Lepski's Procedure . . . 77
3.1.1 Motivation . . . 77
3.1.2 The adaptive CSMD-SR estimate . . . 80
3.2 Analysis under Reduced Uniform Convexity hypothesis (RUC) . . . 84
3.2.1 An example motivating the **RUC** assumption . . . 85
3.2.2 Prescribed choice of parameter and convergence results . . . 86
3.3 Appendix: proofs. . . . 89
3.3.1 Proof of Lemma 3.2.1 . . . 89
3.3.2 Proof of Lemma 3.2.2 . . . 92
3.3.3 Proof of Theorem 3.2.1 . . . 93

III Non-Euclidean Accelerated Methods for Sparse Recovery **95**

4 Accelerated Stochastic Approximation with State-Dependent Noise **97**
4.1 Introduction . . . 97
4.1.1 Contributions and organization . . . 99
4.1.2 Notation . . . 101
4.2 Problem statement . . . 102
4.2.1 Assumptions . . . 102
4.2.2 Mini-batch setup . . . 102
4.3 SAGD for general convex problem with state-dependent noise . . . 104
4.4 SGE for general convex problem with state-dependent noise . . . 107
4.5 SGE for convex problem with quadratic growth condition . . . 110
4.6 SGE for sparse recovery . . . 111
4.6.1 Application: sparse generalized linear regression . . . 112
4.6.2 SGE-SR: stochastic gradient extrapolation for sparse recovery . . . 113
4.7 Numerical experiments . . . 115
4.8 Concluding remarks . . . 118
4.9 Appendix: proofs . . . 119
4.9.1 Proof of Lemmas 4.2.1 and 4.3.1 . . . 119
4.9.2 Proof of Theorem 4.3.1 . . . 120
4.9.3 Proof of Corollary 4.3.1 . . . 122
4.9.4 Proof of Theorem 4.3.2 and Corollary 4.3.2 . . . 123
4.9.5 Proof of Theorem 4.4.1 . . . 125
4.9.6 Proof of Corollary 4.4.1 . . . 129
4.9.7 Proofs of Corollaries 4.5.1 and 4.6.1 . . . 130

5 Large Deviation Bounds, Accuracy Certificates and Composite Algorithm 133
5.1 Introduction . . . 133
5.2 Problem statement . . . 136
5.2.1 Assumptions . . . 136
5.2.2 Mini-batch setup . . . 136
5.2.3 A preliminary large deviation bound . . . 137
5.3 High-probability bounds for SGE . . . 137
5.3.1 General smooth convex problem . . . 138
5.3.2 Smooth convex problem with quadratic growth condition . . . 140
5.4 Accuracy certificate . . . 141
5.5 A Multistage Accelerated Algorithm for Sparse Recovery . . . 144
5.5.1 Local Proximal Setup. . . . 144
5.5.2 Composite Stochastic Gradient Extrapolation (CSGE). . . . 145
5.5.3 Composite Stochastic Gradient Extrapolation for Sparse Recovery . . . 146
5.5.4 Application : Sparse Generalized Linear Regression . . . 148
5.6 Numerical Experiments . . . 151
5.7 Appendix . . . 153
5.7.1 Proof of Lemma 5.2.1 . . . 153
5.7.2 Proof of Lemma 5.2.2 . . . 155
5.7.3 Proof of Lemma 5.3.1 . . . 156
5.7.4 Proof of Theorem 5.3.1 and Corollary 5.3.1 . . . 157
5.7.5 Proof of Corollary 5.3.2 . . . 160
5.7.6 Proof of Proposition 5.4.1, Corollary 5.4.1 and Corollary 5.4.2 . . . 161
5.7.7 Proof of Proposition 5.5.1, Theorem 5.5.1 and Corollary 5.5.1 . . . 167
5.7.8 Proof of Theorem 5.5.2 . . . 172
5.7.9 Proof of Proposition 5.5.2 . . . 174
5.7.10 Condition $\mathbf{Q}(\lambda, \psi)$ . . . 176

Appendix 179
5.8 Regularity in Convex Optimization. . . . 179
5.9 Composite proximal operator . . . 183

List of Algorithms 187

List of Figures 189

Bibliography 190

Outline

Chapter 1: The link between Optimization and Machine Learning is in the focus of the opening chapter. First, we explain how some classical ML tasks can be interpreted as minimization problems of smooth convex functions. We offer an intuitive description of the regularity assumptions frequently found in Optimization literature and introduce the ubiquitous stochastic setting for Optimization problems. Subsequently, we present and analyze three renowned first-order stochastic algorithms: Stochastic Gradient Descent, Stochastic Mirror Descent, and Accelerated Stochastic Gradient Descent. These will serve as the working horses of this manuscript. Lastly, we address a specific machine learning problem that emerges in high-dimensional settings, namely sparse recovery. We explore how first-order optimization methods can leverage sparsity structures, transforming computationally intensive algorithms into more efficient methods.

Chapter 2: The second chapter of this manuscript discusses an application of Stochastic Approximation to the estimation of high-dimensional sparse parameter. We consider the problem of sparse *Generalized Linear Regression* (GLR) with random design. We provide a refined analysis and present high-probability convergence rates for the *Composite Stochastic Mirror Descent* (CSMD) algorithm under smoothness and quadratic minoration assumptions on the objective function and sub-Gaussian stochastic perturbations. We subsequently use the aforementionned algorithm as the primary tool to design an adaptive multistage algorithm for the sparse GLR problem. The proposed multistage procedure resolves a *lasso-type* penalized stochastic optimization problem on each stage; each problem is solved up to the desired accuracy by the non-Euclidean CSMD algorithm. It exhibits a linear convergence rate in the initial "preliminary" phase and then follows a sublinear decay during the "asymptotic" phase. We further show that in the setting of interest the proposed algorithm attains the optimal convergence of the estimation error under mild assumptions on the regressors.

This chapter is based on the paper *Stochastic Mirror Descent for Large-Scale Sparse Recovery* [1], published at AISTATS 2023.

Chapter 3: Building upon the results of the previous chapter, this chapter we explore two new ideas. First, we analyze a technique to make the algorithm adaptive to unknown parameters using Lepski's adaptation procedure. This procedure boils down to creating a grid of search for the parameter, launching the algorithm for each elements of the grid, and then selecting the best estimator based on some criterion. The extra "cost" of this adaptation procedure increases the required amount of samples by only a logarithmic factor, making this adaptation technique very valuable. Next, we extend the multistage method derived in the previous chapter under a *quadratic growth condition* into a two phase multistage procedure for objectives which are uniformly convex around the optimum.

Chapter 4: In this chapter we discuss two non-Euclidean accelerated stochastic approximation algorithms, namely *Stochastic Accelerated Gradient Descent* (SAGD) and *Stochastic Gradient*

Extrapolation (SGE). We use these two algorithms to solve a class of smooth convex optimization problems under general assumptions on the gradient stochastic pertubations. Unlike common practices, we do not consider that the variance of the stochastic observations are uniformly bounded across the entire domain. Instead, we assume that it depends on the "sub-optimality" of the approximate solutions delivered by these algorithms. We show that both SAGD and SGE attains the optimal convergence rate $\mathcal{O}(1/k^2)$ while simulteneously reaching the optimal sample complexity. Notably, SGE achieves these optimal complexities under less restrictive assumptions compared to SAGD. We further develop a multistage scheme based on SGE to solve the sparse GLR problem with hard-thresholding steps (ℓ_0-"norm" penalization) to enforce sparsity.

This chapter is based on the paper *Accelerated Stochastic Approximation with State-Dependent Noise* [2], accepted at Mathematical Programming Series B.

Chapter 5: We start this last chapter by first providing a large deviation analysis of the Stochastic Gradient Extrapolation method. We assume that the dual norm of the stochastic gradient noise follows a sub-exponential distribution, with the parameter of sub-exponentiallity depending on the "suboptimality" of the objective function evaluated on the current search point. These assumptions lead to special concentration inequalities for supermartingales which form the probabilitic foundation of our framework. It allows to provide theoretical convergence guarantees that are valid in high-probability. Additionally, our analysis of the SGE algorithm will be complemented by accuracy certificates, providing a robust theoretical strategy for stopping criterion. Next, we build upon the ideas introduced in the second chapter of this manuscript and introduce the *Composite Stochastic Gradient Extrapolation* algorithm (CSGE). This method then serves as the main workhorse for an accelerated multistage approach designed to solve sparse recovery problems. The hard-thresholding step used in the preceding chapter is replaced by a ℓ_1-norm penalization to induce sparsity of the estimate. We conclude by providing a numerical comparison of the algorithm on the sparse Generalized Linear Regression (GLR) problem.

This chapter is based on the working paper entitled *Accelerated stochastic approximation with state-dependent noise: high-probability bounds and accuracy certificate.*

Part I

Foundations

Chapter 1

Introduction

Machine learning models have become essential in various research domains, driven by the recent exponential growth in computing power and the abundance of data in numerous fields. This development has transformed large-scale machine learning problems, once considered intractable, into manageable challenges. The effectiveness of these methods largely stems from their ability to be trained on large datasets, and to continuously improve as data volumes increase.

Key ML applications rely on regression tasks, such as predicting stock prices in finance, by using historical data, trading volumes, and other economic indicators. In healthcare, classification tasks are prominent, e.g., tumor detection through medical imaging, where models analyze features such as texture, shape, or intensity in X-rays or MRIs. The success of these models desmonstrates the transformative impact of machine learning in harnessing data to solve complex, real-world problems.

It is essential to emphasize that many machine learning tasks fundamentally translate into optimization problems. Sophisticated models, used for regression or classification tasks, essentially revolve around optimizing a specific loss function. The model identification process involves fine-tuning a plethora of parameters to minimize error and maximize accuracy. Progression of machine learning is thus intricately tied to advances in optimization techniques, which are crucial for efficiently navigating vast and complex data-driven landscapes. This synergy between ML and optimization is a cornerstone of the field's success and evolution.

One of the major challenges is related to large-scale settings where both the sample size and data dimensions can be exceedingly large. For example, the DOTA dataset [3] comprises over 10,000 aerial images, each with a resolution of 4K $\times$ 4K, equating to 1.6 million pixels per image. In such scenarios, classical deterministic optimization routines may struggle with computational tractability.

To address these challenges, stochastic optimization methods have been developed, offering substantial improvements in handling computational burdens appearing in large-scale contexts. The success of these methods is so pronounced that they have become the backbone of training most renowned deep learning architectures such as transformers [4]. Large Language Models (LLMs) that currently enjoy a huge success, are predominantly trained using stochastic optimization approaches, showcasing their effectiveness in managing the complexities of modern, data-intensive machine learning tasks.

1.1 Optimization and Machine Learning

1.1.1 Mathematical background on Machine Learning

The fundamental goal of machine learning is to approximate an unknown function $f : \mathsf{X} \to \mathsf{Y}$ that maps elements of an input space X to elements of an output space Y. "Learning" is made by

accessing to some data, this can be for instance a finite set of paired elements from both the input space and output space, in other words elements of the input-output space $\mathsf{Z} = \mathsf{X} \times \mathsf{Y}$. The latter framework is known as *supervised learning* [5]. In this paradigm data are typically in the form of input-output pairs $\{\mathsf{z}_i = (\mathsf{x}_i, \mathsf{y}_i)\}_{i=1}^N \in \mathsf{Z}^N$, given an observation $\mathsf{x} \in \mathsf{X}$ the goal is to predict an output $\mathsf{y} = f(\mathsf{x}) \in \mathsf{Y}$. Practionners typically categorize problems based on the nature of the output space. Classification problems arise when the output space is discrete. For instance a classical example is binary classification when $\mathsf{Y} = \{0, 1\}$ or $\mathsf{Y} = \{-1, 1\}$. The setting where there are more than two possible discrete outputs is named multi-class classification problems. On the other hand, regression problems are characterized by a continuous output space, e.g., $\mathsf{Y} = \mathbf{R}$. The inputs are often called *features* or *attributes* and the outputs *labels* or *targets*. Here are two examples of classical supervised learning problems.

- Multi-class classification: Given a set of images of K different animals with its corresponding label you would like to build a model that is able to classify for each new input images the corresponding animal related to this input.
- Regression: A typical example of a regression task is forecasting weather temperatures. Given input data such as historical temperature, humidity levels, and atmospheric pressure, among others, the objective is to predict the temperature for an upcoming time window.

To give a more formal description of the supervised learning framework, we will have to make the assumption that the data points $\mathsf{z}_1, \dots, \mathsf{z}_N$ are in fact N realisations of a random variable Z such that $\mathsf{z}_i = (\mathsf{x}_i, \mathsf{y}_i)$ is drawn from the joint distribution $Z = (X, Y)$ where X and Y are two random variables respectively on the input space X and the output space Y with probability distribution $p(\mathsf{x}, \mathsf{y})$ on $\mathsf{Z} = \mathsf{X} \times \mathsf{Y}$. In order to quantify "how good" a prediction is, we consider a *loss function* $\ell : \mathsf{Y} \times \mathsf{Y} \to \mathbf{R}$, $\ell(\mathsf{y}, \mathsf{w})$ quantifies the loss of predicting w when the true output is y. The choice of the *loss function* is of major importance as it defines how models will be evaluated. Below we present some examples of the commonly chosen loss function for the two classical ML problems.

- For binary classification $\mathsf{Y} = \{-1, 1\}$ or for multi-class classification $\mathsf{Y} = \{1, \dots, K\}$ the usual choice of loss function is the "0-1" loss defined as $\ell(\mathsf{y}, \mathsf{w}) = \mathbb{1}_{\{\mathsf{y} \neq \mathsf{w}\}}$. When the prediction w is correct, i.e., is the same class as y then it is 0 and 1 if the prediction does not belong to the correct class.
- For regression problems the quadratic loss function is widely used, if $\mathsf{Y} = \mathbf{R}$ then $\ell(\mathsf{y}, \mathsf{w}) = (\mathsf{y} - \mathsf{w})^2$.

Expected risk. The goal now is to find the best function, the best "model" f among all the functions that maps elements of from the input space X to the output space Y. In order to formalize this idea we introduce $\mathcal{F}(\mathsf{X}, \mathsf{Y})$ the set of all measurable functions from X to Y. We can now define the *expected risk*

$$\forall f \in \mathcal{F}(\mathsf{X}, \mathsf{Y}), \quad \mathcal{R}(f) = \mathbb{E}_p\big[\ell(\mathsf{y}, f(\mathsf{x}))\big] = \int_{\mathsf{X} \times \mathsf{Y}} \ell(\mathsf{y}, f(\mathsf{x})) dp(\mathsf{x}, \mathsf{y}), \tag{1.1.1}$$

also known as the *testing error*, *population risk* or *generalization error*. The objective of machine learning is then to minimize the latter quantity and to find the optimal model f_* among all the functions in $\mathcal{F}(\mathsf{X}, \mathsf{Y})$. This can be rephrased as

$$\mathcal{R}(f_*) = \inf_{f \in \mathcal{F}(\mathsf{X}, \mathsf{Y})} \mathcal{R}(f). \tag{1.1.2}$$

Empirical risk. In many problems, the expression for the expected risk is not easily tractable due to several challenges. For instance the underlying probability distribution $p(\mathsf{x}, \mathsf{y})$ is unknown and cannot be directly accessed, or the associated integrals are hard to evaluate or approximate. In these cases, *empirical risk*, which is computed using a finite sample of data, may serve as a practical proxy for the true expected risk. The idea is to gather N i.i.d realisations $\mathsf{z}_1, \ldots, \mathsf{z}_N$ of random variable Z and to approximate the integral in (1.1.1) via its finite-sum Monte-Carlo approximation,

$$\widehat{f} \in \underset{f \in \mathcal{F}(\mathsf{X},\mathsf{Y})}{\operatorname{Argmin}} \left\{ \widehat{\mathcal{R}}(f) = \frac{1}{N} \sum_{i=1}^{N} \ell(\mathsf{y}_i, f(\mathsf{x}_i)) \right\}. \tag{1.1.3}$$

This *empirical risk* [6] also known as *training error* converges to the *expected risk* by the law of large number as $N \to \infty$. In many supervised learning settings, the empirical risk is expressed in terms of a parametric family $f_\theta : \mathsf{X} \to \mathsf{Y}$ where $\theta \in \Theta \subset E$. The objective is then to find the best estimator $\widehat{\theta}$ that verifies

$$\widehat{\theta} \in \underset{\theta \in \Theta}{\operatorname{Argmin}} \left\{ \widehat{\mathcal{R}}(f_\theta) = \frac{1}{N} \sum_{i=1}^{N} \ell(\mathsf{y}_i, f_\theta(\mathsf{x}_i)) \right\}. \tag{1.1.4}$$

Regularization. As an illustrative example, consider the challenge often encountered in machine learning of minimizing the empirical risk. Specifically, reducing the training loss to zero might cause the model to *overfit* to the noise in the dataset, resulting in poor generalization performances. To address this, a common practice is to incorporate a penalty term into the empirical risk. This results in what is known as the *regularized empirical risk*. The function to be minimized, as given in (1.1.5), becomes

$$\widehat{\mathcal{R}}(f_\theta) + h(\theta) = \frac{1}{N} \sum_{i=1}^{N} \ell(\mathsf{y}_i, f_\theta(\mathsf{x}_i)) + h(\theta). \tag{1.1.5}$$

Here, the additive function h acts as a regularizer that controls the model's complexity. It is noteworthy to mention that the parameters that are optimal for the empirical risk are not the same as the one optimal for the regularized empirical risk. This shift in optimality underlines the regularization's core idea: by preventing overfitting, we might obtain a model that, while not perfect for the training data, generalizes better to unseen data. A popular choice of penalization function is the ℓ_p-norm ($p \geq 1$) penalization, given by $h_p(\theta) = \lambda \|\theta\|_p^p$, where λ represents the penalty parameter, and

$$\|\theta\|_p := \left(\sum_{i=1}^{n} |\theta_i|^p \right)^{1/p}.$$

Problems that incorporate norm penalization are extensively studied in the literature, especially within the context of linear regression problems. For instance to avoid overfitting in regression analyses, practitioners often employ the *Thikhonov regularization* [7]. In this case, the penalization is introduced through the squared ℓ_2-norm and leads to the *ridge regression* problem. Another popular choice of regularization in the same context is the ℓ_1-norm penalization leading to the famous *LASSO regression* problem [8]. The ℓ_1-norm penalty is known to have interesting sparsity-inducing properties. These properties led to the development of numerous fields in statistics such as Compressed Sensing or feature selection.

Generalized Linear Regression. In this manuscript we will investigate an important problem in statistical learning, namely the problem of sparse parameter estimation. This will be discussed in

much details in Section 1.3. One of the application of this study will be the fundamental problem of parameter estimation in the *Generalized Linear Regression* (GLR) model [9, 10]. The GLR model represents a broad framework wherein the estimation process involves determining an unknown parameter vector $\beta^* \in \mathbf{R}^n$ given observations $(\mathsf{x}_i, \mathsf{y}_i)$. The model is defined as :

$$\mathsf{y}_i = \mathfrak{r}(\mathsf{x}_i^T \beta^*) + \xi_i, \tag{1.1.6}$$

where $\mathsf{y}_i \in \mathbf{R}$ are the responses, $\mathsf{x}_i \in \mathbf{R}^n$ are random regressors and $\xi_i \in \mathbf{R}$ are realizations of a zero-mean random variable. The function $\mathfrak{r} : \mathbf{R} \to \mathbf{R}$ is known as the *activation function*. The type of regression being studied in a Generalized Linear framework is characterized by the choice of the activation function. For instance, in the case where the responses are Gaussian distributed and the link function is the identity function, defined as $\forall \mathsf{w} \in \mathbf{R},\ \mathfrak{r}(\mathsf{w}) = \mathsf{w}$, the model corresponds to linear regression. Alternatively, when the responses follow a Bernoulli distribution and the link function is the logistic function, defined as $\forall \mathsf{w} \in \mathbf{R},\ \mathfrak{r}(\mathsf{w}) = (1 + \exp(-\mathsf{w}))^{-1}$, the model aligns with the framework of logistic regression model.

In the remainder of the manuscript, we transition from the notational conventions commonly used in the ML community to those prevalent in the Optimization literature. In the sections which follow, we will introduce the optimization framework that is crucial for the development of tools aimed at efficiently solving the vector estimation problem, which was briefly discussed in this section.

1.1.2 Optimization

As we have noticed above, solving ML problems amounts to minimize either the expected risk or the (regularized) empirical risk. This task of searching the extremal value of a function falls into the field of *optimization*. The corresponding setup is as follows : we consider a Euclidean space E and a constraint set $X \subset E$, such that our problem is to find either a minimizer or the minimal value of a function $g : X \to \mathbf{R}$. Formally we aim at finding x^* a minimum of the following *constrained minimization* problem

$$x^* \in \operatorname*{Argmin}_{x \in X} g(x). \tag{OPT}$$

The structure of the objective function g can vary depending on the specific framework of the problem. This function can be expressed as a *finite-sum*, $g(x) = \sum_{i=1}^N g_i(x)$, where the sum aggregates individual functions g_i. Alternatively, g may be represented in the form of an expectation $g(x) = \mathbb{E}\{G(x, \omega)\}$, where $\omega \in \Omega$ is a random variable whose probability distribution is supported on Ω. The focus of this thesis will primarily be on the latter representation, which involves the expectation form of g.

Stochastic Optimization basics. The subject of stochastic optimization are optimization problems in which the objective function or the constraint set possesses randomness. Such problems often arise in scenarios where we encounter random observations originating from an unknown distribution, which subsequently give rise to stochastic losses.

Specifically, consider X, a convex compact set within a Euclidean space E. The goal is to minimize a loss function $g : X \to \mathbf{R}$. This function is represented as the expected value of a stochastic loss function $G : X \times \Omega \to \mathbf{R}$. In what follows, ω represents a random variable with an unknown probability distribution P supported on Ω. We also assume that the stochastic loss $G(\cdot, \omega)$ is convex for any ω. The minimization problem that we want to solve writes

$$g^* = \min_{x \in X} \Big\{ g(x) = \mathbb{E}_\omega\{G(x, \omega)\} = \int_\Omega G(x, \omega) dP(\omega) \Big\}. \tag{1.1.7}$$

The individual instances $\omega_1, \omega_2, \omega_3, \ldots$ are considered as realizations of a random variable ω. Given the inherent random structure of the stochastic loss $G(x, \omega)$ and the unaccessibility of the underlying probability distribution a direct minimization of the following problem is impossible.

From now on, and for the remainder of this introduction, we assume that the objective function of the optimization problem, as defined in equation (OPT), is expressed as an expectation, in accordance with the formulation presented in the above equation.

An essential mechanism in stochastic optimization is the *first-order stochastic oracle*. This mechanism takes as input a point $x \in X$ and provides a *stochastic gradient*, i.e., a random variable $\mathcal{G}(x, \omega)$ that is a noisy approximation of the true gradient of the objective function. In other words,

$$\forall (x, \omega) \in X \times \Omega,\ \mathcal{G}(x, \omega) := \nabla g(x) + \zeta(x, \omega),$$

where the random variable $\zeta(x, \omega)$ is commonly referred to as the *stochastic error*. A standard assumption in the literature of stochastic optimization is the unbiased nature of the oracle, which states that $\forall x \in X,\ \mathbb{E}\{\zeta(x, \omega)\} = 0$. However, ensuring unbiasedness alone does not guarantee convergence. It is also essential to control the fluctuations of the random variable $\mathcal{G}(x, \omega)$. This consideration leads to the standard assumption of a finite variance of the stochastic error.

This is summarized in the following assumption.

Assumption 1 *We assume the existence of a stochastic first-order oracle that provides, for any input point $(x, \omega) \in X \times \Omega$, a stochastic observation $\mathcal{G}(x, \omega)$ verifying*

- $\mathbb{E}\{\mathcal{G}(x, \omega)\} = \nabla g(x)$, (1.1.8)
- $\mathbb{E}\left\{\|\mathcal{G}(x, \omega) - \nabla g(x)\|_*^2\right\} \leq \sigma^2$, (1.1.9)

where $\|z\|_* := \sup\{\langle z, x\rangle :\ \|x\| \leq 1\}$ is the conjugate norm associated to some given norm $\|\cdot\|$.

The aforementioned assumption of *uniformly bounded variance* of the stochastic error may be substituted by other assumptions. Specifically, it is common to assume that the probability distribution of the stochastic error exhibits either light-tail or heavy-tail behavior. The standard assumptions in the light-tail case is the sub-Gaussian assumption of the dual norm of the unbiased stochastic gradient error.

Assumption 2 *(sub-Gaussan tail) Let $\mathcal{G}(x, \omega)$ be the unbiased stochastic first-order information provided by the SFO at search point x and with sample ω. The stochastic gradient is said to be sub-Gaussian if for some $\sigma > 0$ we have*

$$\mathbb{E}\Big\{ \exp\left(\|\mathcal{G}(x, \omega) - \nabla g(x)\|_*^2 / \sigma^2\right) \Big\} \leq \exp(1). \tag{$\mathcal{SG}$}$$

Note that the sub-Gaussian tail assumption ($\mathcal{SG}$) is a much more restrictive condition than the finite variance assumption on the stochastic error. Indeed observe that from Jensen inequality and convexity of $x \mapsto \exp(x)$, we have

$$\exp\Big(\mathbb{E}\big\{\|\mathcal{G}(x, \omega) - \nabla g(x)\|_*^2 / \sigma^2\big\}\Big) \leq \mathbb{E}\Big\{ \exp\left(\|\mathcal{G}(x, \omega) - \nabla g(x)\|_*^2 / \sigma^2\right) \Big\} \leq \exp(1),$$

which implies (1.1.9).

Recent research indicates that, in many machine learning problems, the distribution of the stochastic gradient noise exhibits tail behavior which are heavier than those of sub-Gaussian distributions [11–14]. Among various distribution families exhibiting heavier tail than sub-Gaussian tails, the sub-exponential family appears to be interesting to study.

Assumption 3 *(sub-exponential tail) Let $\mathcal{G}(x,\omega)$ be the unbiased gradient estimation at search point x and with sample ω. Then the stochastic gradient noise is assumed to be sub-Gaussian if for some $\sigma > 0$ we have*

$$\mathbb{E}\Big\{ \exp\big(\|\mathcal{G}(x,\omega) - \nabla g(x)\|_*/\sigma\big)\Big\} \leq \exp(1). \tag{$\mathcal{SE}$}$$

It is important to clarify how we quantify the approximation precision of our estimate since in the stochastic optimization setting, the prediction error exhibits randomness. Let $\mathcal{A}$ be a stochastic optimization algorithm that provides an estimate $\widehat{x}_T$ of x^* after T iterations. One may consider to take the full expectation of the suboptimality gap, these types of bounds are known as *in-expectation* bounds. Given a desired inaccuracy $\epsilon > 0$, this can be formalized as controlling the expected suboptimality gap[1]

$$\mathbb{E}\{g(\widehat{x}_T) - g^*\} \leq \epsilon.$$

In-expectation bounds provide insights into the average performance of the stochastic algorithm. It may sometimes be unsatisfactory. Indeed, as stated above, the suboptimality gap is itself a random variable, influenced by various sources of randomness such as noise in measurements or stochasticity in the optimization process. In situations where the suboptimality gap is subject to outliers or extreme values, relying solely on in-expectation bounds may not capture the full variability or risk associated with the problem. Outliers can significantly affect the expected value and lead to misinterpretations on the actual performances of the algorithm. Therefore, it becomes imperative to establish bounds that guarantees with *high-probability* that the estimate produced by the algorithm is of high reliability. We define the estimate solution $\widehat{x}_T$ provided by the stochastic method $\mathcal{A}$ as an (ϵ,δ)-solution for some $\epsilon,\delta \in (0,1)$, if the suboptimality gap verifies

$$\mathrm{Prob}\big(g(\widehat{x}_T) - g^* \leq \epsilon\big) \geq 1 - \delta.$$

This criterion ensures that the algorithm's performance is robust, consistently providing high-precision solutions in most instances. It is noteworthy that the assumption of the uniformly bounded variance is often employed when deriving in-expectation bounds. Such an assumption is usually sufficient for establishing these bounds, as they focus on the average performance. In contrast, high-probability bounds typically require assumptions about the distribution's tail behavior, whether light-tailed or heavy-tailed. These conditions are crucial as high-probability bounds need to account for rare but significant deviations from the mean. This tail behavior consideration is essential for providing a more comprehensive and reliable measure of the algorithm's performance in a single run, especially in scenarios with potentially extreme outcomes.

Next we define the *complexity* of an optimization algorithm as the number of iterations $T(\epsilon)$ or the number of samples $N(\epsilon)$ required to achieve the specified tolerance level ϵ. This measure of complexity provides a quantitative assessment of the algorithm's efficiency in reaching the desired precision.

In the field of stochastic optimization, two primary paradigms emerge. Firstly, there is the method of directly attempting to minimize the problem in its genuine form, known as *Stochastic Approximation* (SA). On the other hand, one can opt to approximate the expectation of the problem using a Monte-Carlo approximation, the approach referred as *Sample Average Approximation* (SAA).

Stochastic Approximation. Stochastic approximation originates from the seminal paper by Robbins and Monro [15]. It is an iterative method originally aimed at finding the zero of a function.

[1] In the context of convex optimization, the suboptimality gap or the prediction error is not the only performance metric that is of interest to study. One can also consider for some norm $\|\cdot\|$ the estimation error $\|\widehat{x}_T - x^*\|$.

When applied to the gradient of an objective function in an unconstrained minimization problem, SA essentially seeks local optima of the objective function. SA can be adapted to solve the stochastic problem (1.1.7). The core idea of the adapted method is straightforward: use individual samples sequentially to compute a stochastic subgradient $\mathcal{G}_t := \mathcal{G}(x_t, \omega_t)$ of the objective function, and update the current search point x_t according to

$$x_{t+1} = x_t - \eta_t \mathcal{G}_t, \tag{1.1.10}$$

where (η_t) is a sequence of stepsizes. This routine can directly address stochastic problems and is thereby applicable for solving the expected risk. For a constrained minimization problem, the projection operator $\Pi_X(\cdot)$ can be used to ensure the updated point remains within the constraint set. Stochastic Approximation (SA) methods and their variants continue to be the predominant approaches in modern large-scale machine learning. Examples of such methods include the Adaptive Gradient Algorithm (AdaGrad) by Duchi et al. (2011) [16], the Stochastic Gradient Descent variant by Ghadimi and Lan (2013)[17], and the Adam optimizer by Kingma and Ba (2014)[18], among others.

Sample Average Approximation. SAA method [19–22] is rather a two stage technique based on sampling and deterministic optimization aiming to solve problem (1.1.7) than an algorithm. The first step is a sampling step where N independent, identically distributed realizations $\omega_1, \ldots, \omega_N$ of the random variable ω are collected so that the expected loss function is approximated through Monte-Carlo approximation by the average of the N realizations,

$$\widehat{g}_N(x) := \frac{1}{N} \sum_{i=1}^{N} G(x, \omega).$$

Secondly, a deterministic minimization algorithm is used to solve the approximate problem of the initial expected problem (1.1.7):

$$g_N^* = \min_{x \in X} \Big\{ \widehat{g}_N(x) = \frac{1}{N} \sum_{i=1}^{N} G(x, \omega) \Big\}.$$

Searching the minimum of this problem can be viewed as seeking the minimum of the empirical risk. SAA has many pros, the main interest is its inherent simplicity, basically it boils down to forming a deterministic objective function using the collected samples $(\omega_i)_{i=1}^N$ and run a deterministic optimization method on the problem. Another interesting aspect is that SAA allows for *multiple passes* through the data. Therefore it can be interesting when the dataset is limited and searching new datapoints can be very costly. There are also few downsides. One obvious problem is that we are no longer tackling the stochastic problem, it is therefore important to quantify and to control the approximation error $|g_N^* - g^*|$. Another issue is that in modern era ML problems the sample size can be huge and computing the true gradient of the deterministic function can take forever to be computed.

Composite Optimization. In the previous section, we introduced the regularized empirical risk (1.1.5). Unlike the classical empirical risk, this formulation includes an additional penalty term. The objective now becomes finding the minimum of this new composite function. From the point of view of optimization, the minimization of *composite functions* falls within the domain of *composite optimization*. In this context, the objective is to minimize the sum of two functions, each possessing different properties. This is formulated as finding the minimum to

$$\min_{x \in X} \Psi(x) := g(x) + h(x). \tag{1.1.11}$$

Composite optimization has received significant attention, leading to the development of numerous algorithms designed to efficiently handle these specialized functions [23–29]. The convergence behaviors of these algorithms vary, largely depending on the regularity of the objective function. Within the scope of our manuscript, our focus will not be on directly minimizing a composite function. Instead, our primary objective is to develop methods aimed at approximating the optimum x^* of Problem (OPT). We will utilize composite problems, together with some theoretical conditions on the structure of the problem, as intermediary steps or proxies to guide us toward our main objective : the estimation problem.

GLR by Stochastic Approximation. As a motivating example, we are now going to see how the problem of vector estimation in the generalized linear regression model can be formulated as a stochastic optimization problem. First, recall that the objective of vector estimation under the GLR model is to approximate an unknown vector $x^* \in \operatorname{int} X$ based on observations (ϕ_i, y_i). These observations are connected through the function $\mathfrak{r} : \mathbf{R} \to \mathbf{R}$, as described by the relation

$$y_i = \mathfrak{r}(\phi_i^T x^*) + \xi_i, \quad \text{for } i = 1, 2, 3, \dots, \tag{1.1.12}$$

where the random regressors $\phi_i \in \mathbf{R}^n$ and the zero-mean noise $\xi_i \in \mathbf{R}$ are assumed to be mutually independent. By introducing a primitive function $\mathfrak{s} : \mathbf{R} \to \mathbf{R}$ of $\mathfrak{r}$, we can recast the estimation problem as

$$\min_{x \in X} \Big\{ g(x) := \mathbb{E}\left[\mathfrak{s}(\phi^T x^*) - \phi^T x y\right] \Big\}. \tag{1.1.13}$$

By accessing some observations $\omega = (\phi, y)$, we can form an unbiased gradient estimate $\mathcal{G}(x, \omega) := \phi\left(\mathfrak{r}(\phi^T x) - y\right)$ of the stochastic loss $G(x, \omega) := \mathfrak{s}(\phi^T x^*) - \phi^T x y$. Consequently, Problem (1.1.13) can be solved by using stochastic approximation methods to iteratively approximate vector x^*.

1.2 Stochastic First-order Methods

In this section, our goal is to offer a comprehensive overview of various stochastic first-order methods commonly utilized in the field of stochastic optimization. We present the main tools and concepts discussed in the core of the manuscript to provide a complete understanding of our contributions. Our focus will be on exploring the theoretical bounds associated with these methods and discussing their applications.

1.2.1 Stochastic Mirror Descent

Introduced in [30, 31], the *Mirror Descent* (MD) algorithm can be viewed as an extension of the projected subgradient descent method solving (OPT). The fundamental concept of this method originates from the observation that in Banach spaces, the first-order information of the objective function are elements of the dual space. The groundbreaking idea proposed by Nemirovski and Yudin is to execute the gradient descent step in the dual space, as opposed to the conventional approach of operating in the primal space. This technique leverages a *mirror map*, defined as the gradient of a distance generating function ϑ, to map the search point into the dual space. This aligns with the underlying geometry of the optimization problem more effectively. The so-called mirror map must satisfy several properties; these are detailed, for instance, in [32]. The Mirror Descent algorithm has two elegant and equivalent formulations, the *proximal formulation* and the *primal-dual formulation*.

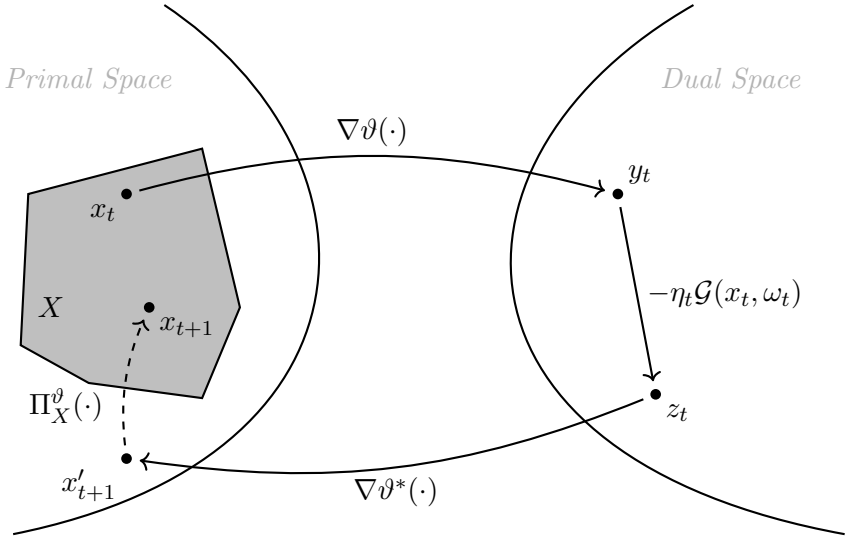


Figure 1.1: Geometric interpretation of Stochastic Mirror Descent algorithm.

We begin by defining the more geometric of the two formulations, namely, the primal-dual formulation. First, we introduce the projection operator onto the space X , which is shaped by a potentially non-Euclidean geometry induced by the strongly convex function ϑ :

$$\Pi_X^\vartheta(x) = \operatorname*{Argmin}_{y\in X} V_\vartheta(x,y),$$

where $V_\vartheta(x,y) := \vartheta(y) - \vartheta(x) - \langle\nabla\vartheta(x), y-x\rangle$ denotes the Bregman divergence associated to the prox-function ϑ. We continue by describing the stochastic variant of the deterministic mirror descent algorithm (SMD), analyzed in [25, 33, 34], where the full gradient is replaced by the stochastic first-order information. Given an initial point $x_0 \in X$, the *stochastic mirror descent* algorithm is defined by the following recursions, for $t = 0, 1, 2, \ldots$

$$y_t = \nabla\vartheta(x_t) \tag{1.2.1}$$
$$z_t = y_t - \eta_t\mathcal{G}_t \tag{1.2.2}$$
$$x_{t+1} = \Pi_X^\vartheta\left(\nabla\vartheta^*(z_t)\right), \tag{1.2.3}$$

where $\mathcal{G}_t$ is an unbiased stochastic subgradient of objective g at search point x_t, i.e., $\mathbb{E}\{\mathcal{G}_t\} \in \partial g(x_t)$. Here, $\nabla\vartheta$ is the mirror map, defined as the gradient of the strongly convex function ϑ which generates the Bregman divergence V_ϑ, and $\nabla\vartheta^*$ is the gradient of the Fenchel conjugate of ϑ. This procedure is illustrated in Figure 1.1.

To derive the second formulation of the SMD method, known as the proximal formulation, we first recall the projected stochastic subgradient descent iterations used to solve the minimization problem in (OPT). It is well-known that the projected subgradient descent method can be rewritten in a proximal form. The iterations then become

$$x_{t+1} = \operatorname*{Argmin}_{x\in X}\left\{\langle\mathcal{G}_t, x\rangle + \frac{1}{2\eta_t}\|x_t - x\|_2^2\right\}. \tag{1.2.4}$$

Later, as shown by Teboulle in [35], the proximal reformulation of the projected subgradient method can be generalized by replacing the squared Euclidean norm in (1.2.4) with a more versatile Bregman Divergence V_ϑ associated with the distance-generating function ϑ. This leads to the mirror descent

recursion emerging as an extension of the Euclidean projected subgradient descent to non-Euclidean geometry. The SMD algorithm can then be summarized by the following recursion, for $t = 0, 1, 2, \ldots$

$$x_{t+1} = \operatorname*{Argmin}_{x \in X} \left\{ \langle \mathcal{G}_t, x \rangle + \frac{1}{\eta_t} V_\vartheta(x_t, x) \right\}. \tag{1.2.5}$$

Observe that when we substitute $\nabla\vartheta^*(y_{t+1})$, as appearing in equation (1.2.3) of the primal-dual formulation, into the expression of the non-Euclidean projection onto the set X, $\Pi_X^\vartheta(\cdot)$, we obtain the same definition of x_{t+1} as in (1.2.5). Indeed, for the sake of conciseness by denoting $\alpha = \nabla\vartheta^*(y_{t+1})$, we have

$$\begin{aligned}
x_{t+1} &= \operatorname*{Argmin}_{x \in X} \{V_\vartheta(\alpha, x)\} \\
&= \operatorname*{Argmin}_{x \in X} \{\vartheta(x) - \vartheta(\alpha) - \langle \nabla\vartheta(\alpha), x - \alpha \rangle\} \\
&= \operatorname*{Argmin}_{x \in X} \{\vartheta(x) - \langle \nabla\vartheta(x_t) - \eta_t \mathcal{G}_t, x - \alpha \rangle\} \\
&= \operatorname*{Argmin}_{x \in X} \{\langle \eta_t \mathcal{G}_t, x \rangle + \vartheta(x) - \vartheta(x_t) - \langle \nabla\vartheta(x_t), x \rangle + \langle \nabla\vartheta(x_t), x_t \rangle\} \\
&= \operatorname*{Argmin}_{x \in X} \left\{ \langle \mathcal{G}_t, x \rangle + \frac{1}{\eta_t} V_\vartheta(x_t, x) \right\}.
\end{aligned}$$

This compact *proximal formulation* of the SMD algorithm is adopted in the remainder of the manuscript.

The SMD method serves as a foundational framework from which many famous stochastic algorithms arise as special cases. Below, we highlight some interesting examples of these methods.

- For $X \subset \mathbf{R}^n$, when selecting $\vartheta(x) = \frac{1}{2}\|x\|_2^2$, the Bregman Divergence becomes $V_\vartheta(x, y) = \frac{1}{2}\|y - x\|_2^2$, which essentially reduces to the stochastic subgradient descent algorithm.
- On the n-dimensional probability simplex $X = \Delta_n := \{x \in \mathbf{R}^n : \sum_{i=1}^n x_i = 1,\ x_i \geq 0,\ \forall i = 1, 2, \ldots, n\}$, using the negative entropy as distance generating function $\vartheta(x) = \sum_{i=1}^n x_i \ln(x_i)$, the Bregman divergence becomes the relative entropy, also known as the *KL-divergence*, $V_\vartheta(x, y) = \sum_{i=1}^n y_i \ln(y_i/x_i)$. This results in the closed-form update of the Exponentiated Gradient algorithm [36] for stochastic optimization.
- When X is a subset of the *spectrahedron* defined as $\Sigma_n = \{Z \in \mathcal{S}_n^{++} : \operatorname{tr}(Z) = 1\}$, using matrix entropy as a distance-generating $\vartheta(Z) = \operatorname{tr}(Z \ln Z)$, we obtain the Von Neumann relative entropy $V_\vartheta(Z, Y) = \operatorname{tr}(Y \ln Y - Y \ln Z)$, which leads to the SMD update presented in [37].

Due to its remarkable versatility and its inherent simplicity, SMD method and its variants are arguably among the most extensively studied and widely used methods in practice. For instance in deep learning [38], the method is used to optimize complex *overparameterized* neural networks [39, 40]. In Reinforcement Learning, especially in policy optimization, it is used to improve decision-making algorithms [41–44]. Online Learning can also benefit from SMD's adaptability and low computation complexity, where it has been extensively studied under the name of Online Mirror Descent [36, 45–47]. Consequently, many theoretical results have been derived to underline the practical utility of the SMD algorithm. For the general non-smooth convex case, Nemirovsky et al. [33] proved that SMD has the optimal $\mathcal{O}(1/\sqrt{T})$ convergence rate. In the non-smooth strongly convex case, Juditsky and Nesterov [48] improved the convergence rate to $\mathcal{O}(1/T)$. Lan [49] studied the L-smooth convex case and provided in-expectation and high-probability bounds for a modified version of the SMD

algorithm where the output is the Polyak-Ruppert average [50, 51] over the iterates for a fixed horizon T, defined as

$$\widehat{x}_T := \frac{1}{T}\sum_{t=1}^{T} x_{t+1}.$$

A typical in-expectation result is presented in Theorem 1.2.1.

Theorem 1.2.1 (Lan, 2020 [49]) *Let $\widehat{x}_T$ be the output of the SMD algorithm applied over a finite horizon $T \in \mathbb{N}^*$. Let the stepsize η be defined as $\eta = \min\left(\frac{1}{2L}, \frac{D_X}{\sigma\sqrt{T}}\right)$, where L represents the Lipschitz constant of ∇g and $D_X^2 = \max_{x,y\in X} V_\vartheta(x,y)$. We also assume that the stochastic gradient noise verifies Assumption 1 with parameter $\sigma > 0$. Under these conditions the expected suboptimality is bounded as :*

$$\mathbb{E}[g(\widehat{x}_T) - g(x^*)] \leq \frac{2LD_X^2}{T} + \frac{2D_X\sigma}{\sqrt{T}}. \tag{1.2.6}$$

The high-probability result is provided under the sub-Gaussian assumption ($\mathcal{SG}$) on the dual norm of the stochastic error.

Theorem 1.2.2 (Lan, 2020 [49]) *Let $\widehat{x}_T$ be the output of the SMD algorithm applied over a finite horizon $T \in \mathbb{N}^*$. Let the stepsize η be defined as $\eta = \min\left(\frac{1}{2L}, \frac{D_X}{\sigma\sqrt{T}}\right)$, where L represents the Lipschitz constant of ∇g, and $D_X^2 = \max_{x,y\in X} V_\vartheta(x,y)$. We also assume that the stochastic gradient noise verifies Assumption 2 with parameter $\sigma > 0$. Under these conditions for $\Lambda > 0$ we have :*

$$\mathrm{Prob}\left\{ g(\widehat{x}_T) - g(x^*) \geq \frac{2LD_X^2}{T} + \frac{2(1+\Lambda)D_X\sigma}{\sqrt{T}} \right\} \leq \exp\{-\Lambda^2/3\} + \exp\{-\Lambda\}. \tag{1.2.7}$$

1.2.2 Accelerated Methods

Designing fast optimization methods has always been a cornerstone of research in optimization. This focus has led to the groundbreaking development of accelerated optimization algorithms.

Momentum methods and acceleration. The first wave of these advanced algorithms emerged in deterministic optimization, marked by the development of momentum methods. In his groundbreaking paper of 1964 [52], Polyak proposed the *heavy-ball method* (HB) for solving smooth and strongly convex minimization problems. The update rule of the latter algorithm is very similar to the update rule of the gradient descent algorithm and is defined as follows :

$$x_{t+1} = x_t - \alpha\nabla g(x_t) + \underbrace{\beta(x_t - x_{t-1})}_{\text{momentum}}. \tag{1.2.8}$$

The key idea behind this added momentum term is to utilize the history of past iterates to predict the future trajectory. This method has been proven to achieve the optimal asymptotic complexity $\mathcal{O}(\sqrt{L/\mu}\ln(1/\epsilon))$ in the L-smooth and μ-strongly convex setting for twice continuously differentiable functions. Later, Nemirovski and Yudin, in their seminal paper [31] proposed an algorithm with a complexity of $\mathcal{O}(\sqrt{L/\mu}\ln(1/\epsilon))$ by drawing inspiration from the *conjugate gradient method* applied for solving convex quadratic minimization problems [53]. In the same paper, focusing on the smooth convex setting, they demonstrated a lower bound of $\Omega(1/T^2)$ for first-order methods, indicating that no method within this class could achieve a faster convergence rate. This opened a gap as the best

convergence rate achieved by first-order methods in the smooth convex setting at that time was $\mathcal{O}(1/T)$. In 1983, Nesterov [54] closes the gap by introducing a simpler method for solving convex minimization problems with Lipschitz-continuous gradients, achieving the optimal rate $\mathcal{O}(1/T^2)$. This method is known as the *accelerated gradient algorithm* (NAG). In the L-smooth and μ-strongly convex setting, this method achieve also the optimal iteration complexity $\mathcal{O}(\sqrt{L/\mu}\ln(1/\epsilon))$. Nesterov's accelerated algorithm operates by starting with two initialization points $x_0 = y_0 \in \mathbf{R}^n$ and updating them as follows :

$$x_{t+1} = y_t - \frac{1}{L}\nabla g(y_t) \tag{1.2.9}$$

$$y_{t+1} = x_{t+1} + \frac{t-1}{t+2}(x_{t+1} - x_t). \tag{1.2.10}$$

This method stems from Polyak's foundational concept of incorporating an additive momentum term into the gradient descent stage. It significantly enhances this approach by a clever idea, which consists in evaluating the gradient at a forward-looking point, a "predictive step" that anticipates the future position. This forward step, based on the momentum term, allows for a better update, thus accelerating the convergence. Figure 1.2 illustrates the update rule of the NAG method.

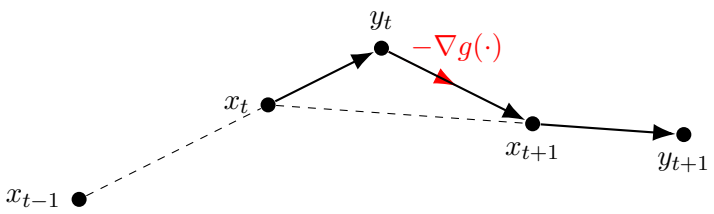


Figure 1.2: Visualization of the update of NAG algorithm.

Following this discovery, the idea of acceleration was extended by numerous authors to design new algorithms for solving composite problems with potentially nonsmooth regularization term [23, 55, 56]. Beck and Teboulle [56], for instance, proposed a reformulation of Nesterov's fast algorithm, known as *Fast Iterative Shrinkage-Thresholding Algorithm* (FISTA), which utilizes the proximal operator to effectively tackle linear inverse problems.

Multistage algorithms. Another widely used technique to enhance the convergence behavior of nearly any first-order method is the *restart scheme*. Under certain conditions on the objective function, it is possible to improve the *sublinear* convergence rate of an optimization algorithm $\mathcal{A}$ to achieve *linear* convergence, as illustrated in Figure 1.3. This technique stems from the observation that even with sublinear convergence, the optimization method may exhibit phases of rapid decrease. The approach can be succinctly described as follows: we run algorithm $\mathcal{A}$ for a predetermined number of iterations, and once the phase of rapid convergence diminishes, we restart algorithm $\mathcal{A}$ using the last output as the new initial point. This process is repeated until a specific termination criterion is satisfied. These strategies have been extensively studied and come with strong theoretical and empirical guarantees [23, 48, 57–59]. For instance O'Donoghue and Candès [58] adapted a restarted version of an accelerated method with great empirical results.

Stochastic methods. In the stochastic setting accelerated algorithms were also developed taking inspirations from the work done in the deterministic setting. To solve composite problems with a smooth convex component and a potentially nonsmooth component, Lan [24] developed a generalization of the NAG method, known as the AC-SA algorithm achieving the theoretical optimal iteration

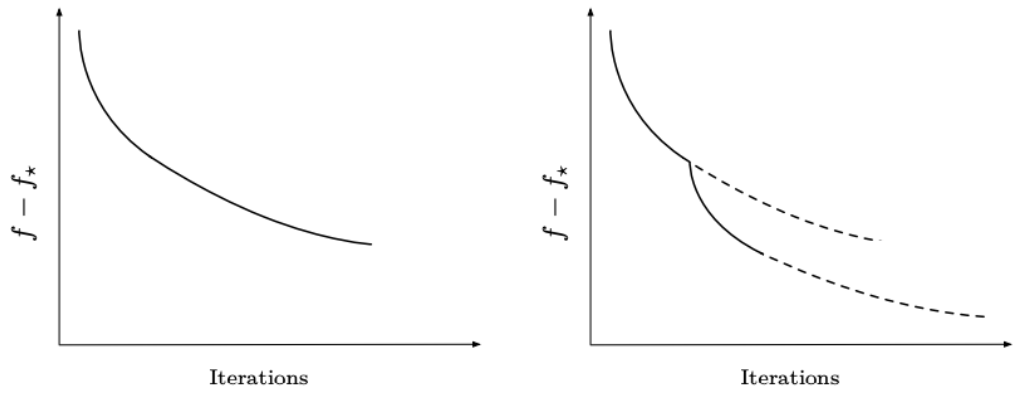


Figure 1.3: *Left:* Sublinear convergence of an optimization method. *Right:* Faster convergence thanks to restart scheme. From d'Aspremont et al. [62].

complexity for solving stochastic composite problems :

$$\mathcal{O}\left(\sqrt{\frac{LD_X^2}{\epsilon}} + \frac{\sigma^2 D_X^2}{\epsilon^2}\right),$$

where $D_X^2 = \max_{x,y\in X} V_\vartheta(x,y)$ and σ is the variance of the stochastic noise. Later, Ghadimi and Lan [25] utilized a modified version of the fast AC-SA method to develop an optimal method for composite problems with smooth strongly convex objectives with potentially nonsmooth regularizer. Their modified version achieves the following, nearly optimal, iteration complexity for finding and (ϵ, Λ)-solution for a composite problem

$$\mathcal{O}\left(\sqrt{\frac{LD_X^2}{\epsilon}} + \frac{\sigma^2}{\mu\epsilon}\ln\left(\frac{1}{\Lambda}\right) + \left(\frac{\sigma R_X(x^*)}{\epsilon}\ln\left(\frac{1}{\Lambda}\right)\right)^2\right),$$

where $R_X(x^*) := \max_{x\in X}\|x - x^*\|$ and for some $\Lambda \in (0,1)$. They present in [26] a better rate for the smooth and strongly convex case. The rate is improved thanks to a multistage strategy and achieve the optimal iteration complexity of

$$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\max\left(1, \ln\left(\frac{\Delta_0}{\epsilon}\right)\right) + \left[\ln\left(\frac{\ln(\Delta_0/\epsilon)}{\Lambda}\right)\right]^2\frac{\sigma^2}{\mu\epsilon}\right),$$

where for the last two bounds, σ denotes the sub-Gaussianity parameter and Δ_0 denotes the initial suboptimality gap.

Stochastic methods with momentum are still currently the go-to methods for practitionners in modern machine learning. For example, in deep learning, a variety of algorithms based on the momentum concept have gained widespread popularity within the community. The most commonly used algorithms include SGD with momentum, Adadelta [60], Adam [18], and Nadam [61], among others

1.3 Sparse Recovery

In this section, motivated by the study of high-dimensional statistical estimation problems, we explore fundamental concepts within the rapidly evolving field of Statistics and Signal Processing

known as *Compressed Sensing* (CS). CS is a technique in Signal Processing that efficiently acquires and reconstructs a signal by finding sparse solutions to underdetermined linear systems. Specifically, Compressed Sensing is concerned with the recovery of a signal $x \in \mathbf{R}^n$ from observations $y \in \mathbf{R}^N$ obtained through a linear model $y = \Phi x + \zeta$, where Φ is the sensing matrix and ζ represents noise. This approach is applicable when the number of observations N is significantly less than the signal's dimension n, while being sufficiently large compared to the signal's "true" dimension, the count of its nonzero entries. Compressed sensing (CS) is not limited to exploiting 'vanilla' sparsity structures; it also addresses other types of low-dimensional structures inherently present in high-dimensional data, such as *block sparsity* and *low-rank matrices*. We begin with defining the problem of signal recovery.

1.3.1 Signal recovery basics

In the context of sparse signal recovery, three primary scenarios emerge based on the characteristics of the sensing matrix Φ and the presence of noise. The deterministic scenario occurs in noise-free environments ($\zeta = 0$) with a full-rank sensing matrix, typically when the number of observations N is greater than or equal to the number of variables n, allowing for the unique recovery of the signal through matrix inversion. Conversely, in the underdetermined scenario, when $N < n$, the resulting system has more variables than observations. In this setting, additional assumptions, such as the effective dimension $s < N$, are required to ensure recoverability. Here, sparsity assumptions and techniques like the nullspace condition are crucial for signal recovery. In the noisy scenario ($\zeta \neq 0$), which reflects real-world conditions with measurement noise, the distinctions between having more observations than unknowns ($N > n$) and fewer observations than unknowns ($N < n$) remain significant. This affects the choice of recovery techniques, with robust approaches such as regularization being employed to approximate the original signal while mitigating the effects of noise.

Compressed sensing and ℓ_0-formulation: In the CS [63–68] setup we consider the case $N \ll n$ and assume that the signal we want to recover can be well approximated by an s-sparse representation. To quantify the sparsity of a vector $z \in \mathbf{R}^n$, we consider its ℓ_0- "norm" defined as

$$\|z\|_0 := \mathrm{Card}(\{i : z_i \neq 0\}).$$

In the CS setup the a priori information is added by considering the following set

$$X = \{z \in \mathbf{R}^n : \|z\|_0 \leq s\}.$$

So the CS framework for the signal recovery problem in the noiseless case ($\zeta = 0$) can be express by the following combinatorial minimization problem :

$$\min_{z \in \mathbf{R}^n} \{\|z\|_0 : \Phi z = y\}. \tag{1.3.1}$$

This can also be extended for noisy observations, given some norm $\|\cdot\|$ and $\rho > 0$, the combinatorial minimization problem becomes:

$$\min_{z \in \mathbf{R}^n} \{\|z\|_0 : \|\Phi z - y\| \leq \rho\}. \tag{1.3.2}$$

Unfortunately, the problems described in (1.3.1) and (1.3.2) are computationally intractable without specific information about the support of the targeted signal. An exhaustive search over all possible subsets of size s is not feasible due to the prohibitively large number of possibilities $\binom{n}{s}$. Furthermore,

these minimization problems have been proven to be NP-hard, as established by Natarajan [69] in the context of sparse signal recovery.

A viable solution to this complex issue exists, and it boils down to "approximating" the ℓ_0-"norm" by a convex surrogate. The ℓ_1 objective is a good proxy to replace the ℓ_0 minimization problem.

ℓ_1-relaxation: The most natural reformulation of problems (1.3.1) and (1.3.2) appears to be the ℓ_1-norm minimization also known as *basis pursuit* problem [70, 71]. In the noiseless case we obtain

$$\min_{z\in\mathbf{R}^n} \{\|z\|_1 : \Phi z = y\}, \tag{1.3.3}$$

and in the presence of additive noise this is reformulated as

$$\min_{z\in\mathbf{R}^n} \{\|z\|_1 : \|\Phi z - y\| \leq \rho\}. \tag{1.3.4}$$

These two problems are now convex and can be efficiently solved via convex programming. One important question to answer is how good the solutions for these convex problems are compared to their ℓ_0-"norm" minimization counterparts. Many studies [72, 73] point out that the ability to recover exactly s-sparse solutions to (1.3.3) depends on properties of the sensing matrix Φ. Juditsky and Nemirovski [73] characterize the notion of *s-goodness* verified by the sensing matrix Φ by observing that whenever the observations are generated in a noiseless model by an s-sparse signal $x \in \mathbf{R}^n$, then it should be the unique solution to (1.3.3).

Definition 1.3.1 *(s-goodness) Let $\Phi \in \mathbf{R}^{N\times n}$ be a sensing matrix and consider an integer $0 \leq s \leq n$. Then Φ is said to be s-good if for every s-sparse vector $x \in \mathbf{R}^n$, x is the unique optimal solution to the problem*

$$\min_{z\in\mathbf{R}^n} \{\|z\|_1 : \Phi z = \Phi x\}. \tag{1.3.5}$$

We may now define a *necessary* and *sufficient* condition that the sensing matrix has to verify to be s-good. Given an index set $I \subset \{1,\ldots,n\}$, we denote by x_I the vector obtained by zeroing all coefficient of x with indices outside of I. First we define some notation. For vector $x \in \mathbf{R}^n$ and an integer $0 \leq s \leq n$, we denote by x^s the vector obtained by setting to 0 all entries except for the s entries of largest magnitudes. Note that x^s is the best s-sparse approximation of x in all ℓ_p-norms $1 \leq p \leq \infty$. For $p \in [1,\infty]$ and $s \leq n$, we introduce the norm

$$\|x\|_{s,p} := \|x^s\|_p = \max_{\mathrm{Card}(I)\leq s} \left\{ \left(\sum_{i\in I} |x_i|^p \right)^{1/p} \right\}.$$

Let now I^{c} be the complementary set $I^{\mathsf{c}} := \{1,\ldots,n\} \setminus I$. The *nullspace condition* of order s is a necessary and sufficient condition of exact recovery in the noiseless problem. The condition is stated as follows

$$\forall I \subset \{1,\ldots,n\} \text{ such that } \mathrm{Card}(I) \leq s,\ \forall w \in \mathrm{Ker}(\Phi) \setminus \{0\} : \|w_I\|_1 < \|w_{I^{\mathsf{c}}}\|_1, \tag{1.3.6}$$

where $\mathrm{Ker}(\Phi) := \{w \in \mathbf{R}^n : \Phi w = 0\}$. Invoking a compactness argument, the previous condition is the same as

$$\exists \kappa \in (0, 1/2),\ \forall w \in \mathrm{Ker}(\Phi) : \|w\|_{s,1} \leq \kappa \|w\|_1. \tag{1.3.7}$$

Whenever the design matrix Φ satisfies the nullspace condition, the exact recovery of s-sparse vector in the noiseless case is assured.

Proposition 1.3.1 *Let $\Phi \in \mathbf{R}^{N\times n}$ be a sensing matrix. Then Φ is s-good if and only if it verifies the nullspace condition of order s.*

There exists other sufficient conditions which guarantee s-goodness of the sensing matrix. One of the most studied condition is the *Restricted Isometry Property* (RIP) introduced by Candès and Tao [74].

Definition 1.3.2 *(RIP) Consider an integer $s \leq N$ and $\delta \in (0,1)$. A $N \times n$ sensing matrix Φ is said to possess RIP(δ, s), when for every s-sparse vector $x \in \mathbf{R}^n$*

$$(1-\delta)\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1+\delta)\|x\|_2^2. \tag{1.3.8}$$

In the case of noiseless observations, we have established that exact recovery of the true signal is guaranteed under specific conditions related to the *nullspace* of the sensing matrix. However, in real-world scenarios, observations are often corrupted with noise, rendering exact recovery unfeasible. Consequently, a practical objective shifts in obtaining an estimator $\widehat{x}$ such that the ℓ_2-error $\|\widehat{x}-x^*\|_2$ is well controlled. To this end and to overcome the challenge appearing in the noisy setting, Bickel et al. [75] introduced the *restricted eigenvalue* condition.

Definition 1.3.3 *(Restricted eigenvalue RE (s,α)) A sensing matrix Φ is said to verify the restricted eigenvalue condition with parameter $1 \leq s \leq n$ and positive number α, when the following condition holds*

$$\kappa(s,\alpha) := \min_{\substack{I\subseteq\{1,\dots,n\},\\ |I|\leq s}} \min_{\substack{\Delta\neq 0,\\ \|\Delta_{I^c}\|_1\leq\alpha\|\Delta_I\|_1}} \frac{\|\Phi\Delta\|_2}{\sqrt{N}\|\Delta\|_2} > 0. \tag{1.3.9}$$

This condition is less restrictive than RIP condition, therefore it has been extensively utilized to prove accuracy bounds for the *Dantzig Selector* estimator [67, 75, 76] and the *Lasso* estimator [75–79]. First, recall that the Dantzig Selector estimator is defined as follows

$$\widehat{x}_{\mathrm{DS}} \in \operatorname*{Argmin}_{z\in\mathbf{R}^n}\{\|z\|_1 : \|\Phi^T(\Phi z - y)\|_\infty \leq \rho\}, \tag{1.3.10}$$

and the Lasso estimator has the form

$$\widehat{x}_{\mathrm{lasso}} \in \operatorname*{Argmin}_{z\in\mathbf{R}^n}\{\tfrac{1}{2N}\|\Phi z - y\|_2^2 + \lambda\|z\|_1\}. \tag{1.3.11}$$

The following theorem provides a typical example of the type of bounds that can be achieved when employing the Restricted Eigenvalue condition, specifically in the context of the Lasso estimator.

Theorem 1.3.1 (Wainwright, 2019 [80]) *Let $\Phi \in \mathbf{R}^{N\times n}$ be a deterministic design matrix verifying the restricted eigenvalue property with parameters $1 \leq s \leq n$ and $\alpha = 3$, and assume also that it is C-column normalized, meaning that $\max_{j=1,\dots,n}\frac{\|\Phi_j\|_2}{\sqrt{N}} \leq C$. Under the sparse linear model with an s-sparse signal x^* and Gaussian random noise, i.e., $\zeta \in \mathbf{R}^N$ with i.i.d. centered Gaussian entries with variance σ^2, when choosing the penalty parameter such that $\lambda = 2C\sigma\left(\sqrt{\frac{2\ln(n)}{N}}+\delta\right)$, we have for any $\delta > 0$, that*

$$\mathrm{Prob}\left\{\|\widehat{x}_{lasso} - x^*\|_2 \leq 6C\frac{\sigma\sqrt{s}}{\kappa(s,3)}\left(\sqrt{\frac{2\ln(n)}{N}}+\delta\right)\right\} \geq 1 - 2\exp\left(-\frac{N\delta^2}{2}\right),$$

The Restricted Eigenvalue condition, has been used for deriving theoretical bounds, notably in scenarios involving random design matrices. For instance, Raskutti et al. [81] demonstrated that the RE condtion and the nullspace condition are valid with high-probability for random design matrix with correlated covariates, where the Restricted Isometry Property fails to hold. Specifically, in the noisy linear model with Gaussian additive noise, they proved that the RE condition and nullspace condition are both valid for a class of correlated Gaussian design. Under these conditions, they established that using either the Lasso or the Dantzig selector estimator yields a solution $\widehat{x}$ that achieves $\|\widehat{x}-x^*\|_2 = \mathcal{O}\left(\sqrt{s\ln(n)/N}\right)$, provided that the number of samples scales as $N = \Omega(s\ln(n))$. Another analysis of the RE condition and its variants can be found in [82].

In Juditsky and Nemirovski [83], a related condition named $\mathbf{Q}_q(s,\kappa,\widehat{\lambda})$, is proposed to facilitate the analysis of the *Dantzig Selector* estimator and the *Lasso* estimator. It requires that for some $s \leq N$, $q \in [1,\infty]$ and $\widehat{\lambda} > 0$ it holds

$$\forall w \in \mathbf{R}^n : \|w\|_{s,q} \leq \widehat{\lambda} s^{\frac{1}{q}} \|\Phi w\|_2 + \kappa s^{1-\frac{1}{q}} \|w\|_1. \qquad \mathbf{Q}_q(s,\kappa,\widehat{\lambda})$$

An important aspect of this condition is that if it is satisfied by a sensing matrix, then the nullspace property (1.3.7) also holds. Conversly, if a design matrix verifies the nullspace property, then there exists a parameter $\widehat{\lambda}$ for which the condition $\mathbf{Q}_q(s,\kappa,\widehat{\lambda})$ is satisfied.

1.3.2 Sparse Recovery and Stochastic Optimization.

Consder the problem of recovery of the signal $x^* \in \mathbf{R}^n$ from observations

$$y_i = \phi_i^T x^* + \sigma\zeta_i, \quad i = 1,2,\ldots, \tag{1.3.12}$$

where $x^* \in X$ is s-sparse, while the regressors $\phi_i \in \mathbf{R}^n$ and noises $\zeta_i \in \mathbf{R}$ are i.i.d. random variables according to some respective distributions. Then the recovery problem can be formulated as a stochastic minimization problem as follows

$$\min_{x\in X} \left\{ g(x) = \mathbb{E}\left\{ \tfrac{1}{2}(y - \phi^T x)^2 \right\} \right\}. \tag{1.3.13}$$

In the following development, we describe two famous approaches that aims to tackle this problem.

SAA approaches. When given a collection of N observations $y = (y_1,\ldots,y_N)^T$ and N regressors represented as a sensing matrix $\Phi^T = (\phi_1;\ldots;\phi_N)$ (C.f. figure 1.4), one can consider the problem of minimize the sample average approximation (SAA) problem :

$$\operatorname*{Argmin}_{x\in X} \left\{ \widehat{g}_N(x) = \tfrac{1}{2N} \sum_{i=1}^{N} \left(\phi_i^T x - y_i\right)^2 = \tfrac{1}{2N}\|\Phi x - y\|_2^2 \right\}. \tag{1.3.14}$$

This Least-Square minimization problem is typically addressed using deterministic algorithms. However, simply solving the Sample Average Approximation (SAA) problem does not guarantee a sparse solution. It is a common practice to enforce sparsity through the *iterative hard thresholding* technique [84–86]. This method operates as follows: a deterministic minimization algorithm first performs a gradient step to minimize the SAA objective $\widehat{g}_N$. Subsequently, the resulting estimator is sparsified by retaining only the s largest components in magnitude and setting all others to 0. This process is then repeated until a specified termination criterion is met.

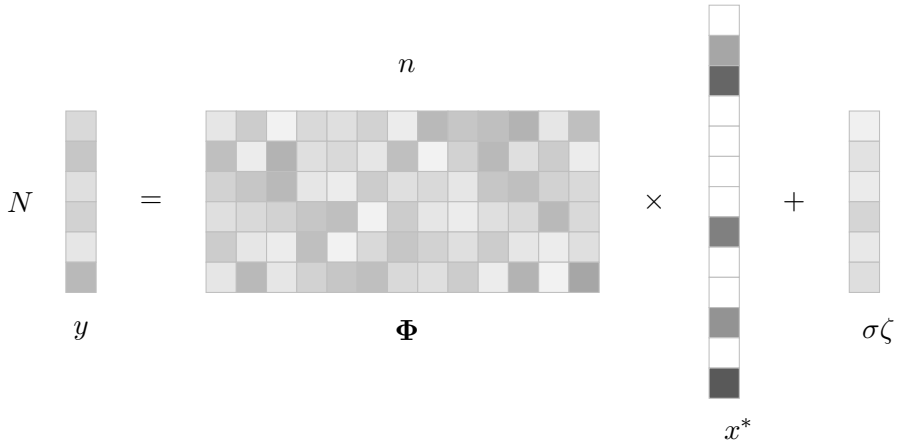


Figure 1.4: Sparse linear recovery problem.

As mentioned in the previous section, instead of working with the ℓ_0-"norm", one way consider its ℓ_1-relaxation. Accordingly, the minimization problem (1.3.14) is substituted by its ℓ_1-penalized problem:

$$\operatorname*{Argmin}_{x\in\mathbf{R}^n} \left\{ \tfrac{1}{2}\|\Phi x - y\|_2^2 + \lambda\|x\|_1 \right\}. \tag{1.3.15}$$

This relaxed problem has been the focus of numerous algorithmic developments. For instance, Beck and Teboulle [56] introduced the FISTA (Fast Iterative Shrinkage-Thresholding Algorithm) method, which efficiently solves problem (1.3.15) and achieves a fast convergence rate of $\mathcal{O}(1/T^2)$. Many theoretical results guarantee a sparse solution for the aforementioned problem, relying on the design matrix Φ meeting some conditions such as those discussed in the last section.
One of the main drawbacks, despite the simplicity of these techniques, is that both the sample size N and the dimension n can be very large. This scale can make the use of deterministic methods solving problem (1.3.15) computationally expensive.

Stochastic Approximation methods. An increasingly popular approach to solving sparse regression problems is the application of Stochastic Approximation algorithms. These methods address the limitation of having to store a potentially very large sensing matrix Φ in memory and also benefit from lower computational costs. SA algorithms are particularly useful when observations are received in an online manner, and when it is possible to form an unbiased estimate of the gradient. However, despite their numerous benefits, these algorithms face a significant challenge in high-dimensional sparse linear regression problems due to their bad scaling in the problem's dimension. Indeed, the usual bounds on the expected suboptimality gap for the SA methods contain the term proportional to the expected squared Euclidean norm of the regressors $\mathbb{E}\{\|\phi\|_2^2\}$. For instance, for regressors with i.i.d. Gaussian entries, unless the regressors possess a sparse structure, the upper bound is inevitably proportional to $\mathbb{E}\{\|\phi\|_2^2\} = \mathcal{O}(n)$. Therefore, to achieve a milder dependency on the problem's dimension, particularly in sparse regression, there is a growing interest in exploring non-Euclidean SA algorithms.
Non-Euclidean SMD have been extensively studied for solving (1.3.13) in the context of stochastic optimization. For instance, Srebro et al. [87] analyzes the SMD algorithm in a sparse linear regression context and proved the "slow" convergence rate

$$\mathbb{E}\{g(\widehat{x})\} - g(x^*) \leq \mathcal{O}\Big(\sqrt{\frac{s\sigma^2\ln(n)}{N}}\Big).$$

multistage procedures, also known as algorithms with restarting schemes, discussed in Section 1.2.2, are known to improve the convergence rates under some conditions. These procedures are particularly valuable in the sparse recovery context, where they have been used to improve the statistical rates for strongly-convex objectives to achieve the optimal minmax rate for sparse estimation : $\mathcal{O}(s\sigma^2/N)$ (up to logarithmic factors). In studies such as those by Agarwal et al. [88], the authors propose an online multistage approach that involves solving a sequence of constrained ℓ_1-penalized problems. Specifically, at each stage i of the procedure, a new minimization problem is defined and solved. This problem is characterized by a sequentially decreasing penalization parameter $(\lambda_i)_{i\geq 1}$ and an adjusted constraint set X_i. Agarwal et al. [88] utilize a variant of the SMD algorithm, namely, the non-Euclidean *regularized dual averaging* method [23, 89], as a subroutine within their main multistage solution solving the newly defined problem at the stage i. In the local Lipschitz and local strong convexity setting, along with sub-Gaussian stochastic gradients and bounded regressors ($\|\phi\|_\infty \leq B$), they achieve, for the sparse linear regression problem, the following statistical rate in high-probability :

$$\|\widehat{x} - x^*\|_2^2 \leq \mathcal{O}\Big(\frac{sB^2\sigma^2 \ln(n)}{\kappa_\Sigma^2 N}\Big),$$

where κ_Σ denotes the smallest eigenvalue of the population covariance matrix Σ of the regressors. It is important to notice that at each stage, the number of iterations needed to provide this rate is of order $\mathcal{O}(s^2\ln(n)/\kappa_\Sigma^2)$. For a fixed budget N, this implies that their multistage method can achieve the sparse vector estimation only when the sparsity level is not exceeding $\mathcal{O}(\kappa_\Sigma\sqrt{N/\ln(n)})$. On the other hand, Raskutti et al. [81] showed that the admissible sparsity level for Lasso problem cannot exceed $\mathcal{O}(\kappa_\Sigma N/\ln(n))$.

Recently, Juditsky et al. [59] employ a multistage method incorporating hard-thresholding steps at the end of each stage to enforce sparsity on the estimator. Their analysis focuses on the study of smooth convex objective functions that verify also the *μ-quadratic growth condition* w.r.t the *Euclidean norm*,

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

By using the SMD method as the workhorse, the multistage method obtains the linear convergence rate (with the exponent proportional to $\frac{\kappa_\Sigma}{s\ln(n)}$) for the deterministic term and obtain a similar statistical rate

$$\mathbb{E}\{\|\widehat{x} - x^*\|_2^2\} \leq \mathcal{O}\left(\frac{s\sigma^2\ln(n)}{\kappa_\Sigma^2 N}\right).$$

The proposed algorithm performs a fixed number of iterations per stages, that is bounded by $\mathcal{O}(s\ln(n)/\kappa_\Sigma)$. Therefore this algorithm can be used whenever the sparsity level is of order $\mathcal{O}(\kappa_\Sigma N/\ln(n))$ thus representing an improvement compared to approach presented in [88].

In this thesis, our objective is to develop and provide rigourous analysis of new fast stochastic optimization algorithms for general smooth convex stochastic opyimization problems in high dimensional settings. Specifically, we compare our methods for problems of the form (1.3.13) in the context of sparse recovery. The algorithms we develop are designed in the stochastic optimization paradigm, where sample availability occurs sequentially. This scenario of online data acquisition renders traditional conditions on the design matrix, as typically seen in deterministic settings,

inappropriate. Consequently, it necessitates the formulation of new conditions for sparse estimation on the population covariance matrix of the regressors. Following the work of [26, 48, 59, 88, 90] we develop fast algorithms based on multistage method. We explore, through both theoretical analysis and numerical experments, how these methods can be effectively employed to tackle sparse recovery problems within the GLR framework.

1.4 Contributions

The rest of the manuscript is divided in two parts. The first part studies a non-Euclidean CSMD algorithm with restarts. We provide in-depth analysis of the main method in the context of stochastic optimization with an application to the problem of sparse GLR. In the second part of the manuscript we analyze a non-Euclidean stochastic accelerated method using mini-batch approximation.

Part I of the manuscript is organized in two chapters.

- In Chapter 2, we present an algorithm that aims to solve smooth convex stochastic optimization problems of the type
$$\min_{x\in X}\left\{g(x) := \mathbf{E}\big\{G\big(x,\omega\big)\big\}\right\},$$
where the solution x_* has a sparse structure. We begin with a refined analysis for the Composite Stochastic Mirror Descent(CSMD) algorithm for solving a norm-penalized auxiliary composite problem of the form
$$\min_{x\in X}\left\{g(x)+\kappa\|x\|\right\}.$$
We assume that the stochastic gradients evaluated at x_* are sub-Gaussian. Using new results on large deviations of sub-Gaussian supermartingales, we derive high-probability convergence guarantees for the CSMD algorithm. The latter allows us to propose the multistage CSMD-SR algorithm, a two phase multistage procedure. Phases are repetions of stages, and each stage is a specific instantiation of the non-Euclidean CSMD algorithm solving a composite subproblem. In the first phase, a fixed step size with iteration count per stage of order $\mathcal{O}(\frac{s\ln(n)}{\kappa_\Sigma})$ results in an estimation error decreasing linearly with the total number of stochastic oracle calls. The phase terminates when the estimation error is of the same order as the statistical error $\mathcal{O}(\frac{\sigma s}{\sqrt{\kappa_\Sigma}})$. During the second phase, at each stage, the step size decreases and the length of the stages increases linearly, leading to the standard "sublinear" rate $\mathcal{O}\Big(\frac{\sigma s}{\kappa_\Sigma}\sqrt{\frac{\ln(n)}{N}}\Big)$. We prove that, with high-probability, this routine achieve the sample complexity
$$\widetilde{\mathcal{O}}\left(\frac{s\nu}{\kappa_\Sigma}\ln\left(\frac{R^2}{\epsilon}\right)+\frac{\sigma^2 s}{{\kappa_\Sigma}^2\epsilon}\right)$$
where $\widetilde{\mathcal{O}}(\cdot)$ hides logarithmic terms, R is the upper bound for the initial distance to the optimum, and ν is the expected Lipschitz constant of the stochastic gradient. These results stems from the Reduced Strong Convexity (RSC) assumption, a new theoretical framework allowing to study recovery problems with diverse sparsity structures. We show that, in the GLR model, this assumption is verified when the objective function leverage a quadratic growth structure and when the population covariance matrix Σ of the regressors verifies the $\mathbf{Q}(\lambda,\psi)$ condition :
$$\forall z\in\mathbf{R}^n,\quad \|z_I\|_1\le\sqrt{\frac{s}{\lambda}}\|z\|_\Sigma+\tfrac{1}{2}(1-\psi)\|z\|_1,$$

for some $\lambda > 0$ and $0 < \psi \leq 1$. This condition is reminiscent of the condition $\mathbf{Q}_q(s, \kappa, \widehat{\lambda})$ presented earlier in the introduction. However, condition $\mathbf{Q}(\lambda, \psi)$ is less restrictive beacause it imposes a condition on the distribution of the regressors rather than on the design matrix Φ. Finally we present a numerical validation of the CSMD-SR algorithm on a high-dimensional sparse estimation problem under the GLR model.

- In Chapter 3, we provide two extensions to the preceding work. We first define a procedure based on Lepski's method [91–93] to turn CSMD-SR algorithm adaptive to unknown values of the sparsity parameter s and the parameter ρ of the RSC assumption, which is the inverse of the quadratic growth condition parameter. This procedure is based on the observation that, the parameterization of the algorithm and the estimation risk depend only on the product $\beta := \rho s$ and are monotonic in this parameter. The adaptive procedure is as follows. We build a grid $\{\beta_1, \beta_2, \ldots, \beta_I\}$ and then run the CSMD-SR algorithm with parameters $\beta = \beta_i$. Thus producing I estimators. Lepski's procedure then amounts to comparing the I estimation risks associated with these estimators and selecting the best among them using some criterion. The additional "cost" of this procedure only deteriorates the bounds given in the last chapter by a logarithmic factor in I, the size of the grid. The CSMD-SR algorithm has the advantage to be adaptive to both ρ and s, which is an improvement over the SMD-SR algorithm proposed in [59], where such adaptation strategies only work for either s or ρ individually. Indeed, the sparsification step in the latter algorithm necessitates the knowledge of the sparsity parameter s alone, making such adaptation procedure impossible.

  We also extend the analysis of the multistage algorithm to the situation where the RSC assumption is replaced with the Reduced Uniform Convexity (RUC) assumption. The latter assumption is verified in the GLR model when two conditions are met. First, the objective function suboptimality must verify a higher-order polynomial lower bound, expressed as:

$$g(x) - g(x_*) \geq \frac{\mu}{p}\|x - x_*\|_\Sigma^p,$$

  where $\mu > 0$, $p \geq 2$, and $\|x\|_\Sigma := \sqrt{x^T \Sigma x}$. Second, the population covariance matrix of the regressors, denoted by Σ, must satisfy the condition $\mathbf{Q}(\lambda, \psi)$, with $\lambda, \psi > 0$. Building on the RUC assumption, we provide a convergence analysis of the CSMD-SR method with new prescribed parameters.

The second part of the manuscript is motivated by the observation that the CSMD-SR routine does not achieve the optimal sample complexity. This discrepancy arises from the fact that the latter multistage algorithm uses the CSMD algorithm as a subroutine, and this algorithm possesses the suboptimal rate of convergence $\mathcal{O}(1/T)$ in the smooth convex setting (cf. Theorem 1.2.2). To address this issue, in the second part of the manuscript we discuss accelerated methods using mini-batch approximation, that achieve the optimal rate $\mathcal{O}(1/T^2)$ in the smooth convex setting.

Part II of the manuscript is structured as follows :

- In Chapter 4, we focus on analyzing two accelerated non-Euclidean stochastic approximation algorithms for smooth convex stochastic optimization. We provide a *in-expectation* analysis of the mini-batch versions of the non-Euclidean *Stochastic Accelerated Gradient Descent* and the *Stochastic Gradient Extrapolation* algorithms for minimizing L-smooth convex objectives. The analysis is performed under the assumption of a state-dependent variance bound on the stochastic gradient noise :

$$\mathbf{E}_{\xi_t}\left[\|\zeta(x_t, \xi_t)\|_*^2\right] \leq \sigma_t^2(x_t) = \mathcal{L}[f(x_t) - f^*] + \sigma_*^2,$$

where $\zeta(x_t, \xi_t)$ is the stochastic gradient noise and $\mathcal{L}, \sigma_* > 0$. We first provide an analysis for the SAGD algorithm. We prove that in order to achieve the optimal iteration complexity $\mathcal{O}(\sqrt{LD_X^2/\epsilon})$, the latter algorithm exhibits a sub-optimal sample complexity

$$\mathcal{O}\left(\sqrt{\frac{LD_X^2}{\epsilon}} + \frac{\sqrt{L}\mathcal{L}D_X^3}{\epsilon^{3/2}} + \frac{\sigma_*^2 D_X^2}{\epsilon^2}\right).$$

However, we show that by adding a condition on the second moment of the Lipschitz constant of the stochastic gradient we improve the second term in the above bound to $\mathcal{O}(1/\epsilon)$. In a second analysis, we study the SGE algorithm and show that, in contrast to the SAGD algorithm, this methods achieves the optimal iteration complexity $\mathcal{O}(\sqrt{LD_X^2/\epsilon})$ while also exhibiting the optimal sample complexity

$$\mathcal{O}\left(\sqrt{\frac{LD_X^2}{\epsilon}} + \frac{\mathcal{L}D_X^2}{\epsilon} + \frac{\sigma_*^2 D_X^2}{\epsilon^2}\right),$$

without adding more restrictive conditions. Next, by exploiting the μ-quadratic growth condition w.r.t. some norm $\|\cdot\|$, we show that by using the SGE algorithm as the workhorse within a multistage procedure, we derive an algorithm that achieves, again, the optimal iteration complexity $\mathcal{O}(\sqrt{L/\mu}\ln(1/\epsilon))$ while simultaneously achieving the optimal sample complexity

$$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\ln\left(\frac{R}{\epsilon}\right) + \frac{\mathcal{L}}{\mu}\ln\left(\frac{R}{\epsilon}\right) + \frac{\sigma_*^2}{\mu\epsilon}\right).$$

Finally, to address smooth convex stochastic problems with sparse minimizers, we draw inspiration from the work [59] and propose a modified multistage procedure. This modification incorporates a hard-thresholding step at the end of each stage to enforce sparsity of the estimators. We finally prove that this method achieves the same optimal complexities.

- Finally, in the Chapter 5 of this manuscript we focus on providing *high-probability* guarantees for the SGE method and its multistage variant. To this end, we introduce new assumptions and new tools. We make the assumption that the dual norm of the stochastic error has a sub-exponential tail with a state-dependant parameter. This assumption is similar to Assumption 3 where the constant parameter $\sigma > 0$ is replaced by a state-dependent counterpart. It is stated as follow

$$\mathbf{E}_{\xi_t}\left[\exp\left(\frac{\|\zeta(x_t, \xi_t)\|_*}{\sigma_t(x_t)}\right)\right] \leq \exp(1),$$

where $\sigma_t^2(x_t) = \overline{\mathcal{L}}[f(x_t) - f^*] + \overline{\sigma}_*^2$ for some $\overline{\mathcal{L}}, \overline{\sigma}_* > 0$. To build high-probability bounds, we introduce new Bernstein types concentration inequalities for sequences of sub-exponential random variables. We prove that both the SGE algorithm and the multistage routine attain the optimal iteration complexity and sample complexity, up to some logarithmic terms, in their respective framework of study.

As a by-product of this analysis we derive accuracy certificates which allows to compute on-the-fly stochastic upper bounds on the suboptimal gap $f(\widehat{x}_T) - f^*$, thereby leading to a stopping criterion for the SGE algorithm.

Finally, we draw inspirations on the work presented in Chapter 2 to solve smooth convex stochastic optimization problems with sparse solutions and introduce the Composite Stochastic

Gradient Extrapolation (CSGE) algorithm. This method is then used in a multistage algorithm for sparse recovery based on an non-Euclidean version of the CSGE algorithm. The idea is to use the CSGE method as the workhorse for solving auxiliary composite subproblems where each stages represents a problem of the form:

$$\min_{x \in X} \{f(x) + \kappa \|x\|\}.$$

This has the effect of incorporating soft-thresholding at each stage leading to sparse estimators. We then use the theoretical tools introduced in Chapter 2 along with the setting of this chapter to provide high-probability guarantees for the CSGE-SR multistage routine. We terminate by providing a theoretical and numerical analysis of the developed algorithm when applied to a sparse GLR problem.

Part II

Stochastic Mirror Descent and Sparse Recovery

Chapter 2

Stochastic Mirror Descent for Large Scale Sparse Recovery

Chapter Abstract

In this paper we discuss an application of Stochastic Approximation to statistical estimation of high-dimensional sparse parameters. The proposed solution reduces to resolving a penalized stochastic optimization problem on each stage of a multistage algorithm; each problem being solved to a prescribed accuracy by the non-Euclidean Composite Stochastic Mirror Descent (CSMD) algorithm. Assuming that the problem objective is smooth and quadratically minorated and stochastic perturbations are sub-Gaussian, our analysis prescribes the method parameters which ensure fast convergence of the estimation error (the radius of a confidence ball of a given norm around the approximate solution). This convergence is linear during the first "preliminary" phase of the routine and is sublinear during the second "asymptotic" phase. We consider an application of the proposed approach to sparse Generalized Linear Regression problem. In this setting, we show that the proposed algorithm attains the optimal convergence of the estimation error under weak assumptions on the regressor distribution. We also present a numerical study illustrating the performance of the algorithm on high-dimensional simulation data.

2.1 Introduction

Our original motivation is the well known problem of (generalized) linear high-dimensional regression with random design. Formally, consider a dataset of N points $(\phi_i, \eta_i), i \in \{1, \ldots, N\}$, where $\phi_i \in \mathbf{R}^n$ are (random) features and $\eta_i \in \mathbf{R}$ are observations, linked by the following equation

$$\eta_i = \mathfrak{r}(\phi_i^T x_*) + \sigma\xi_i, \quad i \in [N] := \{1, \ldots, N\} \tag{2.1.1}$$

where $\xi_i \in \mathbf{R}$ are i.i.d. observation noises. The standard objective is to recover the unknown parameter $x_* \in \mathbf{R}^n$ of the Generalized Linear Regression (2.1.1) – which is assumed to belong to a given convex closed set X and to be *s-sparse,* i.e., to have at most $s \ll n$ non-vanishing entries from the data-set.

As mentioned before, we consider random design, where ϕ_i are i.i.d. random variables, so that the estimation problem of x_* can be recast as the following generic Stochastic Optimization problem:

$$g_* = \min_{x \in X} g(x), \quad \text{where} \quad g(x) = \mathbf{E}\big\{G\big(x, (\phi, \eta)\big)\big\}, \quad G(x, (\phi, \eta)) = \mathfrak{s}(\phi^T x) - \phi^T x \eta, \tag{2.1.2}$$

with $\mathfrak{s}(\cdot)$ any primitive of $\mathfrak{r}(\cdot)$, i.e., $\mathfrak{r}(t) = \mathfrak{s}'(t)$. The equivalence between the original and the stochastic optimization problems comes from the fact that x_* is a critical point of $g(\cdot)$, i.e., $\nabla g(x_*) = 0$ since, under mild assumptions, $\nabla g(x) = \mathbf{E}\{\phi[\mathfrak{r}(\phi^T x) - \mathfrak{r}(\phi^T x_*)]\}$. Hence, as soon as g has a unique minimizer (say, g is strongly convex over X), solutions of both problems are identical.

As a consequence, we shall focus on the generic problem (2.1.2), that has already been widely tackled. For instance, when given an observation sample (ϕ_i, η_i), $i \in [N]$, one may build a Sample Average Approximation (SAA) of the objective $g(x)$

$$\widehat{g}_N(x) = \frac{1}{N}\sum_{i=1}^{N} G(x, (\phi_i, \eta_i)) = \frac{1}{N}\sum_{i=1}^{N} [\mathfrak{s}(\phi_i^T x) - \phi_i^T x \eta_i] \tag{2.1.3}$$

and then solve the resulting problem of minimizing $\widehat{g}_N(x)$ over sparse x's. The celebrated ℓ_1-norm minimization approach allows to reduce this problem to convex optimization. We will provide a new algorithm adapted to this high-dimensional case, and instantiating it to the original problem 2.1.1.

Existing approaches and related works. Sparse recovery by Lasso and Dantzig Selector has been extensively studied [65, 67, 75, 77, 94, 95]. It computes a solution $\widehat{x}_N$ to the ℓ_1-penalized problem $\min_x \widehat{g}_N(x) + \lambda\|x\|_1$ where $\lambda \geq 0$ is the algorithm parameter [96]. This delivers "good solutions", with high probability for sparsity level s as large as $\mathcal{O}\left(\frac{N\kappa_\Sigma}{\ln n}\right)$, as soon as the random regressors (the ϕ_i) are drawn independently from a normal distribution with a covariance matrix Σ such that $\kappa_\Sigma I \preceq \Sigma \preceq \rho\kappa_\Sigma I$[1], for some $\kappa_\Sigma > 0, \rho \geq 1$. However, computing this solution may be challenging in a very high-dimensional setting: even popular iterative algorithms, like coordinate descent, loops over a large number of variables. To mitigate this, randomized algorithms [97, 98], screening rules and working sets [99–101] may be used to diminish the size of the optimization problem at hand, while iterative thresholding [84–86, 102, 103] is a "direct" approach to enhance sparsity of the solution.

Another approach relies on Stochastic Approximation (SA). As $\nabla G(x, (\phi_i, \eta_i)) = \phi_i(\mathfrak{r}(\phi_i^T x) - \eta_i)$ is an unbiased estimate of $\nabla g(x)$, iterative Stochastic Gradient Descent (SGD) algorithm may be used to build approximate solutions. Unfortunately, unless regressors ϕ are sparse or possess a special structure, standard SA leads to accuracy bounds for sparse recovery proportional to the dimension n which are essentially useless in the high-dimensional setting. This motivates non-Euclidean SA procedures, such as Stochastic Mirror Descent (SMD) [104], its application to sparse recovery enjoys almost dimension free convergence and it has been well studied in the literature. For instance, under bounded regressors and with sub-Gaussian noise, SMD reaches "slow rate" of sparse recovery of the type $g(\widehat{x}_N) - g_* = \mathcal{O}\left(\sigma\sqrt{s\ln(n)/N}\right)$ where $\widehat{x}_N$ is the approximate solution after N iterations [87, 105]. Multistage routines may be used to improve the error estimates of SA under strong or uniform convexity assumptions [26, 37, 48]. However, they do not always hold, as in sparse Generalized Linear Regression, where they are replaced by Restricted Strong Convexity conditions. In that setting, the multistage procedure by [88] attains the rate $\mathcal{O}\left(\frac{\sigma}{\kappa_\Sigma}\sqrt{\frac{s\ln n}{N}}\right)$ for the ℓ_2-error $\|\widehat{x}_N - x_*\|_2$ with high probability.[2] This is the best "asymptotic" rate attainable when solving (2.1.2). However, those algorithms have two major limitations. They both need a number of iterations to reach a given accuracy proportional to the initial error $R = \|x_* - x_0\|_1$ and the sparsity level s must be of order $\mathcal{O}\left(\kappa_\Sigma\sqrt{\frac{N}{\ln n}}\right)$ for the sparse linear regression. These limits may be seen as a consequence of dealing with *non-smooth* objective $g(x)$. Although it slightly restricts the scope of corresponding algorithms,

[1] We use $A \preceq B$ for two symmetric matrices A and B if $B - A \succeq 0$, i.e. $B - A$ is positive semidefinite.

[2] Some flaws in the proofs in [88] we fixed by [106].

we shall consider smooth objectives and algorithm for minimizing composite objectives (cf. [23, 107, 108]) to mitigate the aforementioned drawbacks of the multistage algorithms from [88, 106].

Principal contributions. We provide a refined analysis of *Composite Stochastic Mirror Descent (CSMD)* algorithms for computing sparse solutions to Stochastic Optimization problem leveraging smoothness of the objective. This leads to a new "aggressive" choice of parameters in a multistage algorithm with significantly improved performances compared to those in [88]. We summarize below some properties of the proposed procedure for problem (2.1.2).

Each stage of the algorithm is a specific CSMD recursion; They fall into two phases. During the first (preliminary) phase, the estimation error decreases linearly with the exponent proportional to $\frac{\kappa_\Sigma}{s\ln n}$. When it reaches the value $\mathcal{O}\Big(\frac{\sigma s}{\sqrt{\kappa_\Sigma}}\Big)$, the second (asymptotic) phase begins, and its stages contain exponentially increasing number of iterations per stage, hence the estimation error decreases as $\mathcal{O}\Big(\frac{\sigma s}{\kappa_\Sigma}\sqrt{\frac{\ln n}{N}}\Big)$ where N is the total iteration count.

Organization and notation The remaining of the paper is organized as follows. In Section 2.2, the general problem is set, and the multistage optimization routine and the study of its basic properties are presented. Then, in Section 2.3, we discuss the properties of the method and conditions under which it leads to "small error" solutions to sparse GLR estimation problems. Finally, a small simulation study illustrating numerical performance of the proposed routines in high-dimensional GLR estimation problem is presented in Section 2.3.3.

In the following, E is a Euclidean space and $\|\cdot\|$ is a norm on E; we denote $\|\cdot\|_*$ the conjugate norm (i.e., $\|x\|_* = \sup_{\|y\|\le 1}\langle y, x\rangle$). Given a positive semidefinite matrix $\Sigma \in \mathbf{S}_n$, for $x \in \mathbf{R}^n$ we denote $\|x\|_\Sigma = \sqrt{x^T\Sigma x}$ and for any matrix Q, we denote $\|Q\|_\infty = \max_{ij}|[Q]_{ij}|$. We use a generic notation c and C for absolute constants; a shortcut notation $a \lesssim b$ ($a \gtrsim b$) means that the ratio a/b (ratio b/a) is bounded by an absolute constant; the symbols $\bigvee,\bigwedge$ and the notation $(.)_+$ respectively refer to "maximum between", "minimum between" and "positive part".

2.2 Multistage Stochastic Mirror Descent for Sparse Stochastic Optimization

This section is dedicated to the formulation of the generic stochastic optimization problem, the description and the analysis of the generic algorithm.

2.2.1 Problem statement

Let X be a convex closed subset of an Euclidean space E and (Ω, P) a probability space. We consider a mapping $G : X \times \Omega \to \mathbf{R}$ such that, for all $\omega \in \Omega$, $G(\cdot,\omega)$ is convex on X and smooth, meaning that $\nabla G(\cdot,\omega)$ is Lipschitz continuous on X with a.s. bounded Lipschitz constant,

$$\forall x, x' \in X, \quad \|\nabla G(x,\omega) - \nabla G(x',\omega)\|_* \le \mathcal{L}(\omega)\|x - x'\|, \qquad \mathcal{L}(\omega) \le \nu \quad a.s.. \tag{2.2.1}$$

We define $g(x) := \mathbf{E}\{G(x,\omega)\}$, where $\mathbf{E}\{\cdot\}$ stands for the expectation with respect to ω, drawn from P. We shall assume that the mapping $g(\cdot)$ is finite, convex and differentiable on X and we aim at solving the following stochastic optimization problem

$$\min_{x\in X}[g(x) = \mathbf{E}\{G(x,\omega)\}], \tag{2.2.2}$$

assuming it admits an s-sparse optimal solution x_* for some sparsity structure.

To solve this problem, stochastic oracle can be queried: when given at input a point $x \in X$, generates an $\omega \in \Omega$ from P and outputs $G(x,\omega)$ and $\nabla G(x,\omega) := \nabla_x G(x,\omega)$ (with a slight abuse of notations). We assume that the oracle is *unbiased*, i.e.,

$$\mathbf{E}\{\nabla G(x,\omega)\} = \nabla g(x), \qquad \forall x \in X.$$

To streamline presentation, we assume, as it is often the case in applications of stochastic optimization problem (2.2.2), that x_* is unconditional, i.e., $\nabla g(x_*) = 0$ or stated otherwise $\mathbf{E}\{\nabla G(x_*,\omega)\} = 0$; we also suppose the sub-Gaussianity of $\nabla G(x_*,\omega)$, namely that, for some $\sigma_* < \infty$

$$\mathbf{E}\Big\{ \exp\Big(\|\nabla G(x_*,\omega)\|_*^2/\sigma_*^2 \Big) \Big\} \le \exp(1). \tag{2.2.3}$$

2.2.2 Composite Stochastic Mirror Descent algorithm

As mentioned in the introduction, (stochastic) optimization over the set of sparse solutions can be done through "composite" techniques. We take a similar approach here, by transforming the generic problem (2.2.2) into the following *composite Stochastic Optimization problem*, adapted to some norm $\|\cdot\|$, and parameterized by $\kappa \ge 0$,

$$\min_{x\in X} \big[F_\kappa(x) := \tfrac{1}{2} g(x) + \kappa\|x\| = \tfrac{1}{2}\mathbf{E}\{G(x,\omega)\} + \kappa\|x\| \big]. \tag{2.2.4}$$

The purpose of this section is to derive a new (proximal) algorithm. We first provide necessary backgrounds and notations.

Proximal setup, Bregman divergences and Proximal mapping. Let B be the unit ball of the norm $\|\cdot\|$ and $\theta : B \to \mathbf{R}$ be a *distance-generating function (d.-g.f.)* of B, i.e., a continuously differentiable convex function which is strongly convex with respect to the norm $\|\cdot\|$,

$$\langle \nabla\theta(x) - \nabla\theta(x'), x - x' \rangle \ge \|x - x'\|^2, \quad \forall x, x' \in X.$$

We assume w.l.o.g. that $\theta(x) \ge \theta(0) = 0$ and denote $\Theta = \max_{\|z\|\le 1} \theta(z)$.

We now introduce a local and renormalized version of the d.-g.f. θ.

Definition 2.2.1 *For any $x_0 \in X$, let $X_R(x_0) := \{z \in X : \|z - x_0\| \le R\}$ be the ball of radius R around x_0. It is equipped with the d.-g.f. $\vartheta_{x_0}^R(z) := R^2\theta\left((z - x_0)/R\right)$.*

Note that $\vartheta_{x_0}^R(z)$ is strongly convex on $X_R(x_0)$ with modulus 1, $\vartheta_{x_0}^R(x_0) = 0$, and $\vartheta_{x_0}^R(z) \le \Theta R^2$.

Definition 2.2.2 *Given $x_0 \in X$ and $R > 0$, the Bregman divergence V associated to ϑ is defined by*

$$V_{x_0}(x,z) = \vartheta_{x_0}^R(z) - \vartheta_{x_0}^R(x) - \langle \nabla\vartheta_{x_0}^R(x), z - x\rangle, \quad x, z \in X.$$

We can now define *composite proximal mapping* on $X_R(x_0)$ [23, 109] with respect to some convex and continuous mapping $h : X \to \mathbf{R}$.

Definition 2.2.3 *The composite proximal mapping with respect to h and x is defined by*

$$\begin{aligned} \mathrm{Prox}_{h,x_0}(\zeta, x) &:= \operatorname*{argmin}_{z\in X_R(x_0)} \big\{ \langle \zeta, z\rangle + h(z) + V_{x_0}(x,z) \big\} \\ &= \operatorname*{argmin}_{z\in X_R(x_0)} \big\{ \langle \zeta - \nabla\vartheta_{x_0}^R(x), z\rangle + h(z) + \vartheta_{x_0}^R(z) \big\} \end{aligned} \tag{2.2.5}$$

If (2.2.5) can be efficiently solved to high accuracy and Θ is "not too large" (we refer to [33, 37, 109]); those setups will be called "prox-friendly". We now introduce the main building block of our algorithm, the Composite Stochastic Mirror Descent.

Composite Stochastic Mirror Descent algorithm. Given a sequence of positive *step sizes* $\gamma_i > 0$, the *Composite Stochastic Mirror Descent* (CSMD) algorithm is defined by the following recursion

$$x_i \quad = \mathrm{Prox}_{\gamma_i h, x_0}(\gamma_{i-1}\nabla G(x_{i-1}, \omega_i), x_{i-1}), \quad x_0 \in X. \tag{2.2.6}$$

After m steps of CSMD, the final output is $\widehat{x}_m$ (approximate solution) defined by

$$\widehat{x}_m = \frac{\sum_{i=0}^{m-1} \gamma_i x_i}{\sum_{i=0}^{m-1} \gamma_i} \tag{2.2.7}$$

For any integer $L \in \mathbf{N}$, we can also define the L-minibatch CSMD. Let $\omega_i^{(L)} = [\omega_i^1, ..., \omega_i^L]$ be i.i.d. realizations of ω_i. The associated (average) stochastic gradient is then simply defined as

$$H\left(x_{i-1}, \omega_i^{(L)}\right) = \frac{1}{L}\sum_{\ell=1}^{L} \nabla G(x_{i-1}, \omega_i^\ell),$$

which yields the following recursion for the L-minibatch CSMD recursion:

$$x_i^{(L)} \quad = \mathrm{Prox}_{\gamma_i h, x_0}\left(\gamma_{i-1} H\left(x_{i-1}, \omega_i^{(L)}\right), x_{i-1}^{(L)}\right), \quad x_0 \in X, \tag{2.2.8}$$

with its approximate solution $\widehat{x}_m^{(L)} = \sum_{i=0}^{m-1} \gamma_i x_i^{(L)} / \sum_{i=0}^{m-1} \gamma_i$ after m iterations.

From now on, we set $h(x) = \kappa\|x\|$. The next proposition provides an important result for the convergence of the CSMD algorithm

Proposition 2.2.1 *If step-sizes are constant, i.e., $\gamma_i \equiv \gamma \leq (4\nu)^{-1}$, $i = 0, 1, ...,$ and the initial point $x_0 \in X$ such that $x_* \in X_R(x_0)$ then for any $t \gtrsim \sqrt{1 + \ln m}$, with probability at least $1 - 4e^{-t}$*

$$F_\kappa(\widehat{x}_m) - F_\kappa(x_*) \lesssim m^{-1}\left[\gamma^{-1}R^2(\Theta + t) + \kappa R + \gamma\sigma_*^2(m + t)\right], \tag{2.2.9}$$

and the approximate solution $\widehat{x}_m^{(L)}$ of the L-minibatch CSMD satisfies

$$F_\kappa(\widehat{x}_m^{(L)}) - F_\kappa(x_*) \lesssim m^{-1}\left[\gamma^{-1}R^2(\Theta + t) + \kappa R + \gamma\sigma_*^2\Theta L^{-1}(m + t)\right]. \tag{2.2.10}$$

For the sake of clarity and conciseness, we denote CSMD$(x_0, \gamma, \kappa, R, m, L)$ the approximate solution $\widehat{x}_m^{(L)}$ computed after m iterations of L-minibatch CSMD algorithm with initial point x_0, step-size γ, and radius R using recursion (2.2.8).

2.2.3 Main contribution: a multistage adaptive algorithm

Our approach to find sparse solution to the original stochastic optimization problem (2.2.2) consists in solving a sequence of auxiliary composite problems (2.2.4), with their sequence of parameters (κ, x_0, R) defined recursively. For the latter, we need to infer the quality of approximate solution to (2.2.2). To this end, we introduce the following *Reduced Strong Convexity* (RSC) assumption, satisfied in the motivating example (it is discussed in the appendix for the sake of fluency):

Assumption [RSC] There exist some $\delta, \upsilon > 0$ and $\rho < \infty$ such that for any feasible solution $\widehat{x} \in X$ to the composite problem (2.2.4) satisfying, with probability at least $1 - \varepsilon$,

$$F_\kappa(\widehat{x}) - F_\kappa(x_*) \leq \upsilon,$$

it holds, with probability at least $1 - \varepsilon$, that

$$\|\widehat{x} - x_*\| \leq \delta \left[\rho s \kappa + \upsilon \kappa^{-1}\right]. \tag{2.2.11}$$

Given the different problem parameters $s, \nu, \delta, \rho, \kappa, R$ and some initial point $x_0 \in X$ such that $x_* \in X_R(x_0)$ Algorithm 1 works in stages. Each stage represents a run of CSMD algorithm with properly set penalty parameter κ. More precisely, at stage $k+1$, given the approximate solution $\widehat{x}_m^k$ of stage k, a new instance of CSMD is initialized on $X_{R_{k+1}}(x_0^{k+1})$ with $x_0^{k+1} = \widehat{x}_m^k$ and $R_{k+1} = R_k/2$.

Furthermore, those stages are divided into two phases which we refer to as *preliminary* and *asymptotic*:

Preliminary phase: During this phase, the step-sizes γ and the number of CSMD iterations per stage are fixed; the error of approximate solutions converges linearly with the total number of calls to stochastic oracle. This phase terminates when the error of approximate solution becomes independent of the initial error of the algorithm; then the asymptotic phase begins.

Asymptotic phase: In this phase, the step-size decreases and the length of the stage increases linearly; the solution converges sublinearly, with the "standard" rate $O\big(N^{-1/2}\big)$ where N is the total number of oracle calls. When expensive proximal computation (2.2.5) results in high numerical cost of the iterative algorithm, minibatches are used to keep the number of iterations per stage fixed.

In the algorithm description, $\overline{K}_1$ and $\overline{K}_2 \asymp 1 + \log(\frac{N}{m_0})$ stand for the respective maximal number of stages of the two phases of the method, here, $m_0 \asymp s\rho\nu\delta^2(\Theta + t)$ is the length of stages of the first (preliminary) phase. The pseudo-code for the variant of the asymptotic phase with minibatches is given in Algorithm 2.

The following theorem states the main result of this paper, an upper bound on the precision of the estimator computed by our multistage method.

Theorem 2.2.1 *Assume that the total sample budget satisfies $N \geq m_0$, so that at least one stage of the preliminary phase of Algorithm 1 is completed, then for $t \gtrsim \sqrt{\ln N}$ the approximate solution $\widehat{x}_N$ of Algorithm 1 satisfies, with probability at least $1 - C(\overline{K}_1 + \overline{K}_2)e^{-t}$,*

$$\|\widehat{x}_N - x_*\| \lesssim R \exp\left\{-\frac{c}{\delta^2 \rho s} \frac{N}{\nu(\Theta + t)}\right\} + \delta^2 \rho s \sigma_* \sqrt{\frac{\Theta + t}{N}}.$$

The corresponding solution $\widehat{x}_N^{(b)}$ of the minibatch Algorithm 2 satisfies with probability $\geq 1 - C(\overline{K}_1 + \widetilde{K}_2)e^{-t}$

$$\|\widehat{x}_N^{(b)} - x_*\| \lesssim R \exp\left\{-\frac{c}{\delta^2 \rho s} \frac{N}{\nu\,(\Theta + t)}\right\} + \delta^2 \rho s \sigma_* \sqrt{\frac{\Theta\,(\Theta + t)}{N}}.$$

where $\widetilde{K}_2 \asymp 1 + \ln\left(\frac{N}{\Theta m_0}\right)$ is the bound for the number of stages of the asymptotic phase of the minibatch algorithm and $c, C > 0$ are absolute constant.

Algorithm 1 Composite Stochastic Mirror Descent for Sparse Recovery (CSMD-SR)

Initialization : Initial point $x_0 \in X$, step-size $\gamma = (4\nu)^{-1}$, initial radius R_0, confidence level t, total budget N.
Set $m_0 \asymp s\rho\nu\delta^2(\Theta + t)$, $\overline{K}_1 \asymp \ln\left(\frac{R_0^2\nu}{\delta^2\sigma_*^2\rho s}\right) \wedge \frac{N}{m_0}$, $L = 1$

if $R_0 \gtrsim \sigma_*\delta\sqrt{\frac{\rho s}{\nu}}$ **continue** with preliminary stage,
else proceed directly to asymptotic phase
end
for stage $k = 1, \ldots, \overline{K}_1$ **do** ▷ Preliminary Phase
  Set $\kappa_k \asymp R_{k-1}(\delta\rho s)^{-1}$
  Compute approximate solution $\widehat{x}^k_{m_0} = \mathrm{CSMD}(x_0, \gamma, \kappa_k, R_k, m_0, L)$ at stage k
  Reset the prox-center $x_0 = \widehat{x}^k_{m_0}$
  Set $R_k = R_{k-1}/2$
end for
Set $\widehat{x}_N = \widehat{x}^{\overline{K}_1}_{m_0}$, $B = N - m_0\overline{K}_1$, $m_1 \asymp m_0$
if $m_1 > B$ **output :** $\widehat{x}_N$ and **return**; **endif** ▷ Asymptotic Phase
Set $r_0 = R_{\overline{K}_1}$
Set $k = 1$
while $m_k \le B$ **do**
  Set $\kappa_k \asymp 2^{-k}\sigma_*(\rho\nu s)^{-1/2}$, $\gamma_k \asymp 4^{-k}\nu^{-1}$
  Compute approximate solution $\widehat{x}^k_{m_k}$ =CSMD$(x_0, \gamma_k, \kappa_k, r_k, m_k, L)$ at stage k
  Reset the prox-center $x_0 = \widehat{x}^k_{m_k}$
  Set $B = B - m_k$, $k = k + 1$, $r_k = r_{k-1}/2$, $m_k \asymp 4^k m_0$
end while
output : $\widehat{x}_N = \widehat{x}^{k-1}_{m_{k-1}}$

Remark 2.2.1 *Along with the oracle computation, proximal computation to be implemented at each iteration of the algorithm is an important part of the computational cost of the method. It becomes even more important during the asymptotic phase when number of iterations per stage increases exponentially fast with the stage count, and may result in poor real-time convergence. The interest of minibatch implementation of the second phase of the algorithm is in reducing drastically the number of iterations per asymptotic stage. The price to be paid is an extra factor* $\sqrt{\Theta}$ *that could also theoretically hinder convergence. However, in the problems of interest (sparse and group-sparse recovery, low rank matrix recovery)* Θ *is logarithmic in problem dimension. Furthermore, in our numerical experiments we did not observe any accuracy degradation when using the minibatch variant of the method.*

2.3 Sparse generalized linear regression by stochastic approximation

2.3.1 Problem setting

We now consider again the original problem of recovery of a s-sparse signal $x_* \in X \subset \mathbf{R}^n$ from random observations defined by

$$\eta_i = \mathfrak{r}(\phi_i^T x_*) + \sigma\xi_i, \quad i = 1, 2, ..., N, \tag{2.3.1}$$

Algorithm 2 Asymptotic phase of CSMD-SR with minibatch

Input : The approximate solution $\widehat{x}_{m_0}^{\overline{K}_1}$ at the end of the preliminary stage, step-size parameter γ, radius at the end of the preliminary phase $R_{\overline{K}_1}$, initial batch size $\ell_1 \asymp \Theta$

Set $r_0 = R_{\overline{K}_1}$, $x_0 = \widehat{x}_{m_0}^{\overline{K}_1}$, $B = N - m_0\overline{K}_1$ ▷ Asymptotic Phase
$k = 1$
while $m_0\ell_k \leq B$ **do**
  $\kappa_k \asymp 2^{-k}\sigma_*(\rho\nu s)^{-1/2}$
  Compute approximate solution $\widehat{x}_{m_0}^k$ =CSMD$(x_0, \gamma_k, \kappa_k, r_k, m_0, L = \ell_k)$ at stage k
  Reset the prox-center $x_0 = \widehat{x}_{m_0}^k$
  Set $B = B - m_0\ell_k$, $k = k+1$, $r_k = r_{k-1}/2$, $\ell_k \asymp 4^k\ell_1$
end while

output: $\widehat{x}_N^{(b)} = \widehat{x}_{m_1}^k$

where $\mathfrak{r} : \mathbf{R} \to \mathbf{R}$ is some non-decreasing and continuous "activation function", and $\phi_i \in \mathbf{R}^n$ and $\xi_i \in \mathbf{R}$ are mutually independent. We assume that ξ_i are sub-Gaussian, i.e., $\mathbf{E}\{e^{\xi_i^2}\} \leq \exp(1)$, while regressors ϕ_i are bounded, i.e., $\|\phi_i\|_\infty \leq \overline{\nu}$. We also denote $\Sigma = \mathbf{E}\{\phi_i\phi_i^T\}$, with $\Sigma \succeq \kappa_\Sigma I$ with some $\kappa_\Sigma > 0$, and $\|\Sigma_j\|_\infty \leq \upsilon < \infty$.

We will apply the machinery developed in Section 2.2, with respect to

$$g(x) = \mathbf{E}\{\mathfrak{s}(\phi^T x) - x^T\phi\eta\}$$

where $\mathfrak{r}(t) = \mathfrak{s}'(t)$ for some convex and continuously differentiable $\mathfrak{s}$, applied with the norm $\|\cdot\| = \|\cdot\|_1$ (hence $\|\cdot\|_* = \|\cdot\|_\infty$), from some initial point $x_0 \in X$ such that $\|x_* - x_0\|_1 \leq R$. It remains to prove that the different assumptions of Section 2.2 are satisfied.

Proposition 2.3.1 *Assume that $\mathfrak{r}$ is $\overline{r}$-Lipschitz continuous and $\underline{r}$-strongly monotone (i.e., $|\mathfrak{r}(t) - \mathfrak{r}(t')| \geq \underline{r}|t - t'|$ which implies that $\mathfrak{s}$ is $\underline{r}$-strongly convex) then*

1. *[Smoothness] $G(\cdot, \omega)$ is $\mathcal{L}(\omega)$-smooth with $\mathcal{L}(\omega) \leq \overline{r}\overline{\nu}^2$.*
2. *[Quadratic minoration] g satisfies*

$$g(x) - g(x_*) \geq \tfrac{1}{2}\underline{r}\|x - x_*\|_\Sigma^2. \tag{2.3.2}$$

3. *[Reduced Strong Convexity] Assumption [RSC] holds with $\delta = 1$ and $\rho = (\kappa_\Sigma\underline{r})^{-1}$.*
4. *[Sub-Gaussianity] $\nabla G(x_*, \omega_i)$ is $\sigma^2\overline{\nu}^2$-sub Gaussian.*

The proof is postponed to the appendix. The last point is a consequence of a generalization of the Restricted Eigenvalue property [75], that we detail below (as it gives insight on why Proposition 2.3.1 holds).

This condition, that we state and call $\mathbf{Q}(\lambda, \psi)$ in the following Lemma 2.3.1, and is reminiscent of [83] with the corresponding assumptions of [81, 110].

Lemma 2.3.1 *Let $\lambda > 0$ and $0 < \psi \leq 1$, and suppose that for all subsets $I \subset \{1, ..., n\}$ of cardinality smaller than s the following property is verified:*

$$\forall z \in \mathbf{R}^n \quad \|z_I\|_1 \leq \sqrt{\frac{s}{\lambda}}\|z\|_\Sigma + \tfrac{1}{2}(1 - \psi)\|z\|_1 \qquad \mathbf{Q}(\lambda, \psi)$$

where z_I is obtained by zeroing all its components with indices $i \notin I$.

If $g(\cdot)$ satisfies the quadratic minoration condition, *i.e., for some $\mu > 0$,*

$$g(x) - g(x_*) \geq \tfrac{1}{2}\mu\|x - x_*\|_\Sigma^2, \tag{2.3.3}$$

and that $\widehat{x}$ is an admissible solution to (2.2.4) satisfying, with probability at least $1-\varepsilon$,

$$F_\kappa(\widehat{x}) \leq F_\kappa(x_*) + \upsilon.$$

Then, with probability at least $1-\varepsilon$,

$$\|\widehat{x} - x_*\|_1 \leq \frac{s\kappa}{\lambda\mu\psi} + \frac{\upsilon}{\kappa\psi}. \tag{2.3.4}$$

Remark 2.3.1 *Condition $\mathbf{Q}(\lambda,\psi)$ generalizes the classical Restricted Eigenvalue (RE) property [75] and Compatibility Condition [77], and is the most relaxed condition under which classical bounds for the error of ℓ_1-recovery routines were established. Validity of $\mathbf{Q}(\lambda,\psi)$ with some $\lambda > 0$ is necessary for Σ to possess the celebrated* null-space property *[111]*

$$\exists \psi > 0 : \max_{I,\,|I|\leq s} \|z_I\|_1 \leq \tfrac{1}{2}(1-\psi)\|z\|_1 \ \ \forall z \in \mathrm{Ker}(\Sigma)$$

which is necessary and sufficient for the s-goodness of Σ (i.e., $\widehat{x} \in \mathrm{Argmin}_u \{\|u\| : \Sigma u = \Sigma x_*\}$ reproduces exactly every s-sparse signal x_* in the noiseless case).

When Σ possesses the nullspace property, $\mathbf{Q}(\lambda,\psi)$ may hold for Σ with nontrivial kernel; this is typically the case for random matrices [81, 112] such as rank deficient Wishart matrices, etc. When Σ is a regular matrix, condition $\mathbf{Q}(\lambda,\psi)$ may also holds with constant λ which is much higher than the minimal eigenvalue of Σ when the eigenspace corresponding to small eigenvalues of Σ does not contain vectors z with $\|z_I\|_1 > \frac{1}{2}(1-\psi)\|z\|_1$.

Special cases. The quadratic minoration bound (2.3.2) for $g(x)-g(x_*)$ is usually overly pessimistic. Indeed, consider for instance, Gaussian regressor $\phi \sim \mathcal{N}(0,\Sigma)$ (even if they are not a.s. bounded, this is for illustration purposes) and activation $\mathfrak{r}$, define for some $0 \leq \alpha \leq 1$ (with the convention, $0/0 = 0$)

$$\mathfrak{r}(t) = \begin{cases} t, & |t| \leq 1, \\ \mathrm{sign}(t)[\alpha^{-1}(|t|^\alpha - 1) + 1], & |t| > 1. \end{cases} \tag{2.3.5}$$

When passing from ϕ to $\varphi = \Sigma^{-1/2}\phi$ and from x to $z = \Sigma^{1/2}x$ and using the fact that

$$\varphi = \frac{zz^T}{\|z\|_2^2}\varphi + \underbrace{\left(I - \frac{zz^T}{\|z\|_2^2}\right)\varphi}_{=:\chi}$$

with independent $\frac{zz^T}{\|z\|_2^2}\varphi$ and χ, with $\mathbf{E}\{\chi\} = 0$, we obtain

$$\begin{aligned} H(z) =& \mathbf{E}\{\varphi[\mathfrak{r}(\varphi^T z)]\} = \mathbf{E}\left\{\frac{zz^T}{\|z\|_2^2}\varphi\,\mathfrak{r}(\varphi^T z)\right\} \\ =& \frac{z}{\|z\|_2}\mathbf{E}\left\{\varsigma\mathfrak{r}(\varsigma\|z\|_2)\right\} = \frac{\Sigma^{1/2}x}{\|x\|_\Sigma}\mathbf{E}\left\{\varsigma\mathfrak{r}(\varsigma\|x\|_\Sigma)\right\} \end{aligned}$$

where $\varsigma \sim \mathcal{N}(0,1)$. Thus, $H(\Sigma^{1/2}x)$ is proportional to $\frac{\Sigma^{1/2}x}{\|x\|_\Sigma}$ with coefficient

$$h\big(\|x\|_\Sigma\big) = \mathbf{E}\left\{\varsigma\mathfrak{r}(\varsigma\|x\|_\Sigma)\right\}.$$

Figure 2.1 represents the mapping h for different values of α (on the left), along with the dependence on r of moduli of strong monotonicity of corresponding mappings H on the centered at the origin $\|\cdot\|_2$-ball of radius r (on the right).

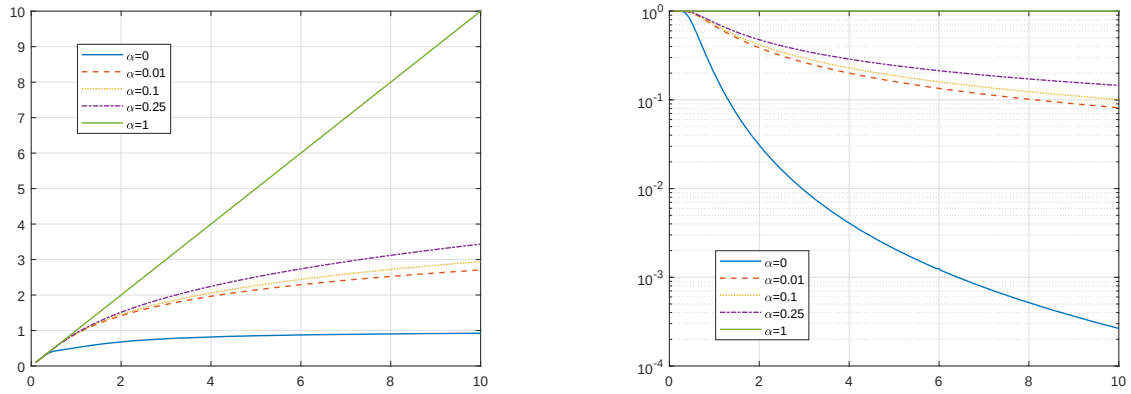


Figure 2.1: Given the activation function $\mathfrak{r}$ in (2.3.5) and $\alpha = (0, 0.01, 0.1, 0.25, 1)$; left plot: mappings h; right plot: moduli of strong monotonicity of mappings H on $\{z : \|z\|_2 \le r\}$ as function of r.

In the case of linear regression where $\mathfrak{r}(t) = t$, it holds

$$\begin{array}{rcl} g(x) & = & \mathbf{E}\big\{\frac{1}{2}(\phi^T x)^2 - x^T\phi\eta\big\} \\ & = & \frac{1}{2}\mathbf{E}\big\{(\phi^T(x_* - x))^2 - (\phi^T x_*)^2\big\} \\ & = & \frac{1}{2}(x - x_*)^T\Sigma(x - x_*) - \frac{1}{2}x_*^T\Sigma x_* \\ & = & \frac{1}{2}\|x - x_*\|_\Sigma^2 - \frac{1}{2}\|x_*\|_\Sigma^2 \end{array}$$

and $\nabla G(x,\omega) = \phi\phi^T(x - x_*) - \sigma\xi\phi$. In this case $\mathcal{L}(\omega) \le \|\phi\phi^T\|_\infty \le \overline{\nu}^2$.

2.3.2 Stochastic Mirror Descent algorithm

In this section, we describe the statistical properties of approximate solutions of Algorithm 1 when applied to the sparse recovery problem. We shall use the following distance-generating function of the ℓ_1-ball of $\mathbf{R}^n$ (cf. [37, Section 5.7.1])

$$\theta(x) = \frac{c}{p}\|x\|_p^p, \quad p = \begin{cases} 2, & n = 2 \\ 1 + \frac{1}{\ln(n)}, & n \ge 3, \end{cases} \qquad c = \begin{cases} 2, & n = 2, \\ e\ln n, & n \ge 3. \end{cases} \tag{2.3.6}$$

It immediately follows that θ is strongly convex with modulus 1 w.r.t. the norm $\|\cdot\|_1$ on its unit ball, and that $\Theta \le e\ln n$. In particular, Theorem 2.2.1 entails the following statement.

Proposition 2.3.2 *For $t \gtrsim \sqrt{\ln N}$, assuming the samples budget is large enough, i.e., $N \ge m_0$ (so that at least one stage of the preliminary phase of Algorithm 1 is completed), the approximate solution $\widehat{x}_N$ output satisfies with probability at least $1 - Ce^{-t}\ln N$,*

$$\|\widehat{x}_N - x_*\|_1 \lesssim R\exp\left\{-c\frac{\underline{r}\kappa_\Sigma}{\overline{r}\overline{\nu}^2}\frac{N}{s(\ln n + t)}\right\} + \frac{\sigma\overline{\nu}s}{\underline{r}\kappa_\Sigma}\sqrt{\frac{\ln n + t}{N}} \tag{2.3.7}$$

The corresponding solution $\widehat{x}_N^{(b)}$ *of the minibatch variant of the algorithm satisfies with probability* $\geq 1 - Ce^{-t}\ln N$,

$$\|\widehat{x}_N^{(b)} - x_*\|_1 \lesssim R\exp\left\{-c\frac{\underline{r}\kappa_\Sigma}{\overline{r}\overline{\nu}^2}\frac{N}{s\,(\ln n + t)}\right\} + \frac{\sigma\overline{\nu}s}{\underline{r}\kappa_\Sigma}\sqrt{\frac{\ln n\,(\ln n + t)}{N}}$$

Remark 2.3.2 *Bounds for the* ℓ_1*-norm of the error* $\widehat{x}_N - x_*$ *(or* $\widehat{x}_N^{(b)} - x_*$*) established in Proposition 2.3.2 allows us to quantify prediction error* $g(\widehat{x}_N) - g(x_*)$ *(and* $g(\widehat{x}_N^{(b)}) - g(x_*)$*), and also lead to bounds for* $\|\widehat{x}_N - x_*\|_\Sigma$ *and* $\|\widehat{x}_N - x_*\|_2$ *(respectively, for* $\|\widehat{x}_N^{(b)} - x_*\|_\Sigma$ *and* $\|\widehat{x}_N^{(b)} - x_*\|_2$*). For instance, Proposition 2.2.1 in the present setting implies the bound on the prediction error after* N *steps of the algorithm that reads*

$$g(\widehat{x}_N) - g(x_*) \lesssim \frac{R^2\kappa_\Sigma\underline{r}}{s}\exp\left\{-\frac{c\kappa_\Sigma\underline{r}}{\overline{r}\overline{\nu}^2}\frac{N}{s(\Theta + t)}\right\} + \frac{\sigma^2\overline{\nu}^2 s(\Theta + t)}{\kappa_\Sigma\underline{r}N}$$

with probability $\geq 1 - C\ln N e^{-t}$. *We conclude by (2.3.2) that*

$$\begin{aligned}\|\widehat{x}_N - x_*\|_2^2 &\leq \kappa_\Sigma^{-1}\|\widehat{x}_N - x_*\|_\Sigma^2 \leq 2\kappa_\Sigma^{-1}\underline{r}^{-1}[g(\widehat{x}_N) - g(x_*)]\\ &\lesssim \frac{R^2}{s}\exp\left\{-\frac{c\kappa_\Sigma\underline{r}}{\overline{r}\overline{\nu}^2}\frac{N}{s(\Theta + t)}\right\} + \frac{\sigma^2\overline{\nu}^2 s(\Theta + t)}{\kappa_\Sigma^2\underline{r}^2 N}.\end{aligned}$$

In other words, the error $\|\widehat{x}_N - x_*\|_2$ *converges geometrically to the "asymptotic rate"* $\frac{\sigma\overline{\nu}}{\kappa_\Sigma\underline{r}}\sqrt{\frac{s(\Theta+t)}{N}}$ *which is the "standard" rate established in the setting (cf. [75, 96, 102], etc).*

Remark 2.3.3 *The proposed approach allows also to address the situation in which regressors are not a.s. bounded. For instance, consider the case of random regressors with i.i.d sub-Gaussian entries such that*

$$\forall j \leq n, \quad \mathbf{E}\left[\exp\left(\frac{[\phi_i]_j^2}{\varkappa^2}\right)\right] \leq 1.$$

Using the fact that the maximum of uniform norms $\|\phi_i\|_\infty$, $1 \leq i \leq m$, *concentrates around* $\varkappa\sqrt{\ln mn}$ *along with independence of noises* ξ_i *of* ϕ_i, *the "smoothness" and "sub-Gaussianity" assumptions of Proposition 2.3.2 can be stated "conditionally" to the event* $\{\omega : \max_{i\leq m}\|\phi_i\|_\infty^2 \lesssim \varkappa^2(\ln[mn] + t)\}$ *of probability greater than* $1 - e^{-t}$. *For instance, when replacing the bound for the uniform norm of regressors with* $\varkappa^2(\ln[mn] + t)$ *in the definition of algorithm parameters and combining with appropriate deviation inequality for martingales (cf., e.g., [113]), one arrives at the bound for the error* $\|\widehat{x}_N - x_*\|_1$ *of Algorithm 1 which is similar to (2.3.7) of Proposition 2.3.2 in which* $\overline{\nu}$ *is replaced with* $\varkappa\sqrt{\ln[mn] + t}$.

2.3.3 Numerical experiments

In this section, we present results of a small simulation study illustrating the theoretical part of the previous section.[3] We consider the GLR model (2.3.1) with activation function (2.3.5) where $\alpha = 1/2$. In our simulations, x_* is an s-sparse vector with s nonvanishing components sampled independently from the standard s-dimensional Gaussian distribution; regressors ϕ_i are sampled from a multivariate Gaussian distribution $\phi \sim \mathcal{N}(0, \Sigma)$, where Σ is a diagonal covariance matrix with diagonal entries $\Sigma_{1,1} \leq \cdots \leq \Sigma_{n,n}$. In Figure 2.2 we report on the experiment in which we compare the performance

[3]The reader is invited to check Section 2.4.6 of the supplementary material for more experimental results.

of the CSMD-SR algorithm from Section 2.2.3 to that of four other methods. The contenders are (1) "vanilla" non-Euclidean SMD algorithm constrained to the ℓ_1-ball equipped with the distance generating function (2.3.6), (2) composite non-Euclidean dual averaging algorithm (p-Norm RDA) from [89], (3) multistage SMD-SR of [59], and (4) "vanilla" Euclidean SGD. The regularization parameter of the ℓ_1 penalty in (2) is set to the theoretically optimal value $\lambda = 2\sigma\sqrt{2\log(n)/T}$. The corresponding dimension of the parameter space is $n = 500000$, the sparsity level of the optimal point x_* is $s = 200$, and the "total budget" of oracle calls is $N = 250000$; we use the identity regressor covariance matrix ($\Sigma = I_n$) and $\sigma \in \{0.001, 0.1\}$. To reduce computation time we use the minibatch versions of the multi-stage algorithms—CSMD-SR and algorithm (3)), the data to compute stochastic gradient realizations $\nabla G(x_i, \omega) = \phi(\mathfrak{r}(\phi^T x_i) - \eta)$ at the current search point x_i being generated "on the fly." We repeat simulations 20 times and plot the median value along with the first and the last deciles of the error $\|\widehat{x}_i - x_*\|_1$ at each iteration of the algorithm against the number of oracle calls.

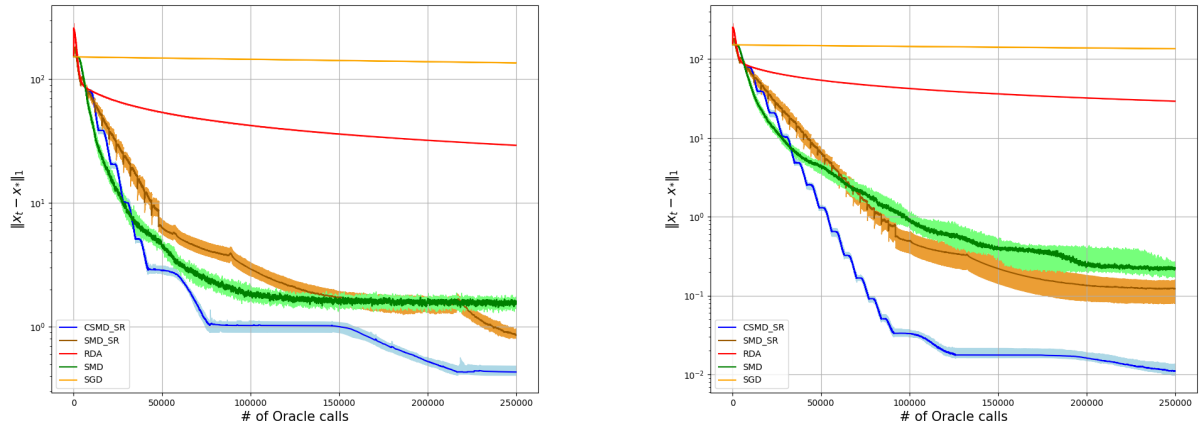


Figure 2.2: Comparison between CSMD-SR and baseline algorithms in Generalized Linear Regression problem: ℓ_1 error as a function of the number of oracle calls

The proposed method outperforms other algorithms which struggle to reach the regime where the stochastic noise is dominant.

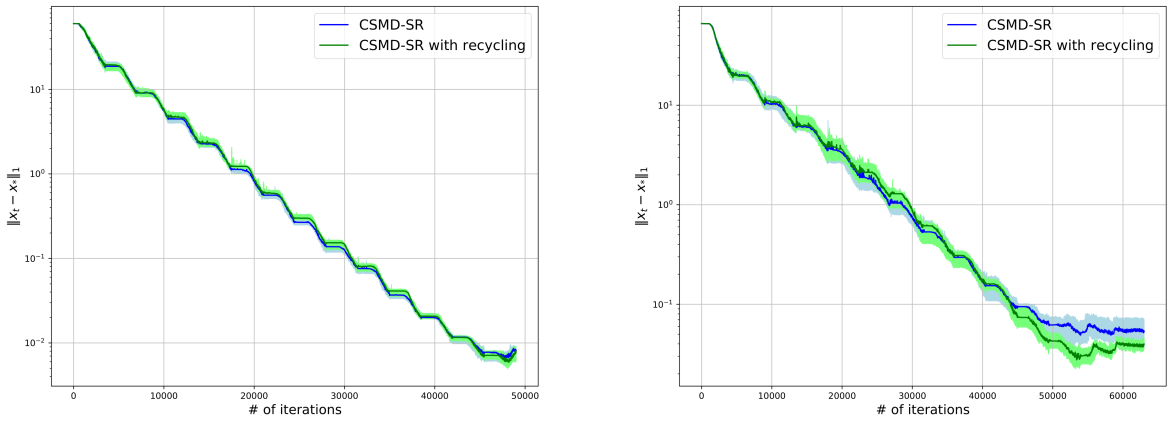


Figure 2.3: Preliminary stages of the CSMD-SR and its variant with data recycling: linear regression experiment (left pane), GLR with activation $\mathfrak{r}_{1/10}(t)$ (right pane).

In the second experiment we report on here, we study the behavior of the multistage algorithm derived from Algorithm 2 in which, instead of using independent data samples, we reuse the same data at each stage of the method. In Figure 2.3 we present results of comparison of the CSMD-SR algorithm with its variant with data recycle. This version is of interest as it attains fast the noise regime while using limited amount of samples. In our first experiment, we consider linear regression problem with parameter dimension $n = 100\,000$ and sparsity level $s = 75$ of the optimal solution; we consider the GLR model (2.3.1) with activation function $\mathfrak{r}_{1/10}(t)$ in the second experiment. We choose $\Sigma = I_n$ and $\sigma = 0.001$; we run 14 (preliminary) stages of the algorithm with $m_0 = 3500$ in the first simulation and $m_0 = 4500$ in the second. We believe that the results speak for themselves.

2.4 Appendix

We use notation $\mathbf{E}_i$ for conditional expectation given x_0 and $\omega_1, ..., \omega_i$.

2.4.1 Proof of Proposition 2.2.1

The result of Proposition 2.2.1 is an immediate consequence of the following statement.

Proposition 2.4.1 *Let*

$$f(x) = \tfrac{1}{2}g(x) + h(x), \quad x \in X.$$

In the situation of Section 2.2.2, let $\gamma_i \le (4\nu)^{-1}$ for all $i = 0, 1, ...$, and let $\widehat{x}_m$ be defined in (2.2.7), where x_i are iterations (2.2.6). Then for any $t \ge 2\sqrt{2 + \ln m}$ there is $\overline{\Omega}_m \subset \Omega$ such that $\mathrm{Prob}(\overline{\Omega}_m) \ge 1 - 4e^{-t}$ *and for all $\omega^m = [\omega_1, ..., \omega_m] \in \overline{\Omega}_m$,*

$$\begin{aligned}\left(\sum_{i=0}^{m-1}\gamma_i\right)[f(\widehat{x}_m) - f(x_*)] &\le \sum_{i=0}^{m-1}\left[\tfrac{1}{2}\gamma_i\langle\nabla g(x_i), x_i - x_*\rangle + \gamma_{i+1}(h(x_{i+1}) - h(x_*))\right]\\ &\le V(x_0, x_*) + \gamma_0[h(x_0) - h(x_*)] - \gamma_m[h(x_m) - h(x_*)]\\ &\quad + V(x_0, x_*) + 15tR^2 + \sigma_*^2\left[7\sum_{i=0}^{m-1}\gamma_i^2 + 24t\overline{\gamma}^2\right]. \end{aligned} \tag{2.4.1}$$

In particular, when using the constant stepsize strategy with $\gamma_i \equiv \gamma$, $0 < \gamma \le (4\nu)^{-1}$, one has

$$\begin{aligned}&\tfrac{1}{2}[g(\widehat{x}_m) - g(x_*)] + [h(\widehat{x}_m) - h(x_*)]\\ &\quad\le \frac{V(x_0, x_*) + 15tR^2}{\gamma m} + \frac{h(x_0) - h(x_m)}{m} + \gamma\sigma_*^2\left(7 + \frac{24t}{m}\right).\end{aligned} \tag{2.4.2}$$

Proof. Denote $H_i = \nabla G(x_{i-1}, \omega_i)$. In the sequel, we use the shortcut notation $\vartheta(z)$ and $V(x, z)$ for $\vartheta_{x_0}^R(z)$ and $V_{x_0}(x, z)$ when exact values x_0 and R are clear from the context.

1^o. From the definition of x_i and of the composite prox-mapping (2.2.5) (cf. Lemma A.1 of [109]), we conclude that there is $\eta_i \in \partial h(x_i)$ such that

$$\langle\gamma_{i-1}H_i + \gamma_i\eta_i + \nabla\vartheta(x_i) - \nabla\vartheta(x_{i-1}), z - x_i\rangle \ge 0, \ \ \forall\, z \in \mathcal{X},$$

implying, as usual [114], that $\forall z \in X$

$$\langle\gamma_{i-1}H_i + \gamma_i\eta_i, x_i - z\rangle \le V(x_{i-1}, z) - V(x_i, z) - V(x_{i-1}, x_i).$$

In particular,

$$\begin{aligned}&\gamma_{i-1}\langle H_i, x_{i-1} - x_*\rangle + \gamma_i\langle\eta_i, x_i - x_*\rangle\\ &\le V(x_{i-1}, x_*) - V(x_i, x_*) - V(x_{i-1}, x_i) + \gamma_{i-1}\langle H_i, x_{i-1} - x_i\rangle\\ &\le V(x_{i-1}, x_*) - V(x_i, x_*) + \tfrac{1}{2}\gamma_{i-1}^2\|H_i\|_*^2.\end{aligned}$$

Observe that due to the Lipschitz continuity of $\nabla G(\cdot, \omega)$ one has

$$\nu\langle\nabla G(x, \omega) - \nabla G(x', \omega), x - x'\rangle \ge \|\nabla G(x, \omega) - \nabla G(x', \omega)\|_*^2, \quad \forall x, x' \in \mathcal{X}, \tag{2.4.3}$$

so that

$$\begin{aligned}\|\nabla G(x,\omega)\|_*^2 &\le 2\|\nabla G(x,\omega)-\nabla G(x_*,\omega)\|_*^2+2\|\nabla G(x_*,\omega)\|_*^2\\ &\le 2\nu\langle\nabla G(x,\omega)-\nabla G(x_*,\omega),x-x_*\rangle+2\|\nabla G(x_*,\omega)\|_*^2\\ &= 2\nu\langle\nabla G(x,\omega),x-x_*\rangle-2\nu\langle\nabla G(x_*,\omega),x-x_*\rangle+2\|\nabla G(x_*,\omega)\|_*^2\end{aligned}$$

so that

$$\begin{aligned}&\gamma_{i-1}\langle H_i,x_{i-1}-x_*\rangle+\gamma_i\langle\eta_i,x_i-x_*\rangle\\ &\le V(x_{i-1},x_*)-V(x_i,x_*)+\gamma_{i-1}^2[\nu\langle H_i,x_{i-1}-x_*\rangle-\nu\zeta_i+\tau_i]\end{aligned}$$

where $\zeta_i=\langle\nabla G(x_*,\omega_i),x_{i-1}-x_*\rangle$ and $\tau_i=\|\nabla G(x_*,\omega)\|_*^2$. As a result, by convexity of h we have for $\gamma_i\le(4\nu)^{-1}$

$$\begin{aligned}&\tfrac{3}{4}\gamma_{i-1}\langle\nabla g(x_{i-1}),x_{i-1}-x_*\rangle+\gamma_i[h(x_i)-h(x_*)]\\ &\le(\gamma_{i-1}-\gamma_{i-1}^2\nu)\langle\nabla g(x_{i-1}),x_{i-1}-x_*\rangle+\gamma_i\langle\eta_i,x_i-x_*\rangle\\ &\le V(x_{i-1},x_*)-V(x_i,x_*)+(\gamma_{i-1}-\gamma_{i-1}^2\nu)\langle\xi_i,x_{i-1}-x_*\rangle+\gamma_{i-1}^2[\tau_i-\nu\zeta_i]\end{aligned}$$

where we put $\xi_i=H_i-\nabla g(x_{i-1})$. When summing from $i=1$ to m we obtain

$$\begin{aligned}&\sum_{i=1}^m\gamma_{i-1}\Big(\tfrac{3}{4}\langle\nabla g(x_{i-1}),x_{i-1}-x_*\rangle+[h(x_{i-1})-h(x_*)]\Big)\\ &\qquad\le V(x_0,x_*)+\underbrace{\sum_{i=1}^m[\gamma_{i-1}^2(\tau_i-\nu\zeta_i)+\gamma_{i-1}(1-\gamma_{i-1}\nu)\langle\xi_i,x_{i-1}-x_*\rangle]}_{=:R_m}\\ &\qquad\quad+\gamma_0[h(x_0)-h(x_*)]-\gamma_m[h(x_m)-h(x_*)].\end{aligned}\tag{2.4.4}$$

2^o. We have

$$\begin{aligned}\gamma_{i-1}\langle\xi_i,x_{i-1}-x_*\rangle &= \gamma_{i-1}\overbrace{\langle[\nabla G(x_{i-1},\omega_i)-\nabla G(x_*,\omega_i)]-\nabla g(x_{i-1}),x_{i-1}-x_*\rangle}^{\upsilon_i}\\ &\quad+\gamma_{i-1}\langle\nabla G(x_*,\omega_i),x_{i-1}-x_*\rangle\\ &= \gamma_{i-1}[\upsilon_i+\zeta_i],\end{aligned}$$

so that

$$R_m=\sum_{i=1}^m\gamma_{i-1}^2\tau_i+\sum_{i=1}^m(\gamma_{i-1}-\gamma_{i-1}^2\nu)\upsilon_i+\sum_{i=1}^m(\gamma_{i-1}-2\nu\gamma_{i-1}^2)\zeta_i=:r_m^{(1)}+r_m^{(2)}+r_m^{(3)}.\tag{2.4.5}$$

Note that $r_m^{(3)}$ is a sub-Gaussian martingale. Indeed, one has $\mathbf{E}_{i-1}\{\zeta_i\}=0$ a.s.,[4] and

$$|\zeta_i|\le\|x_{i-1}-x_*\|\,\|\nabla G(x_*,\omega)\|_*,$$

so that by the sub-Gaussian hypothesis (2.2.3), $\mathbf{E}_{i-1}\Big\{\exp\Big(\underbrace{\frac{\zeta_i^2}{4R^2\sigma_*^2}}_{\nu_*^2}\Big)\Big\}\le\exp(1)$. As a result (cf. the proof of Proposition 4.2 in [115]),

$$\forall t\quad\mathbf{E}_{i-1}\left\{e^{t\zeta_i}\right\}\le\exp\left(t\mathbf{E}_{i-1}\{\zeta_i\}+\tfrac{3}{4}t^2\nu_*^2\right)=\exp\left(3t^2R^2\sigma_*^2\right),$$

[4] We use notation $\mathbf{E}_{i-1}$ for the conditional expectation given $x_0,\omega_1,...,\omega_{i-1}$.

and applying (2.4.10a) to $S_m = r_m^{(3)}$ with

$$r_m = 6R^2\sigma_*^2 \sum_{i=0}^{m-1} (\gamma_i - 2\nu\gamma_i^2)^2 \le 6R^2\sigma_*^2 \sum_{i=0}^{m-1} \gamma_i^2$$

we conclude that for some $\Omega_m^{(3)}$ such that $\mathrm{Prob}(\Omega_m^{(3)}) \ge 1 - e^{-t}$ and all $\omega^m \in \Omega_m^{(3)}$

$$r_m^{(3)} \le 2\sqrt{3tR^2\sigma_*^2 \sum_{i=0}^{m-1} \gamma_i^2} \le 3tR^2 + 3\sigma_*^2 \sum_{i=0}^{m-1} \gamma_i^2. \tag{2.4.6}$$

Next, again by (2.2.3), due to the Jensen inequality, $\mathbf{E}_{i-1}\{\tau_i\} \le \sigma_*^2$, and

$$\mathbf{E}_{i-1}\left\{\exp\left(t\|\nabla G(x_*,\omega_i)\|_*\right)\right\} \le \exp\left(t\mathbf{E}_{i-1}\{\|\nabla G(x_*,\omega_i)\|_*\} + \tfrac{3}{4}t^2\sigma_*^2\right) \le \exp\left(t\sigma_* + \tfrac{3}{4}t^2\sigma_*^2\right).$$

Thus, when setting

$$\mu_i = \gamma_{i-1}\sigma_*, \quad s_i^2 = \tfrac{3}{2}\gamma_{i-1}\sigma_*^2, \quad \overline{s} = \max_i \gamma_i s_i,$$

$M_m = r_m^{(1)}$, $v_m + h_m = \frac{21}{4}\sigma_*^4 \sum_{i=0}^{m-1} \gamma_i^4$, and applying the bound (2.4.10b) of Lemma 2.4.1 we obtain

$$r_m^{(1)} \le 3\sigma_*^2 \sum_{i=0}^{m-1} \gamma_i^2 + \underbrace{\sqrt{21t\sigma_*^4 \sum_{i=0}^{m-1} \gamma_i^4}}_{=:\Delta_m^{(1)}} + 3t\overline{\gamma}^2\sigma_*^2$$

for $\overline{\gamma} = \max_i \gamma_i$ and $\omega^m \in \Omega_m^{(1)}$ where $\Omega_m^{(1)}$ is of probability at least $1 - e^{-t}$. Because

$$\overline{\gamma}^2 \sum_{i=0}^{m-1} \gamma_i^2 \ge \sum_{i=0}^{m-1} \gamma_i^4,$$

whenever $\sqrt{21t\sum_{i=0}^{m-1}\gamma_i^4} \ge \sum_{i=0}^{m-1}\gamma_i^2$, one has $21t\overline{\gamma}^2 \ge \sum_{i=0}^{m-1}\gamma_i^2$ and

$$21t \sum_{i=0}^{m-1} \gamma_i^4 \le 21t\overline{\gamma}^2 \sum_{i=0}^{m-1} \gamma_i^2 \le (21t\overline{\gamma}^2)^2$$

Thus,

$$\Delta_m^{(1)} \le \min\left[21t\sigma_*^2\overline{\gamma}^2, \sigma_*^2 \sum_{i=0}^{m-1} \gamma_i^2\right] \le 21t\sigma_*^2\overline{\gamma}^2 + \sigma_*^2 \sum_{i=0}^{m-1} \gamma_i^2,$$

and

$$r_m^{(1)} \le \sigma_*^2\left[4\sum_{i=0}^{m-1} \gamma_i^2 + 24t\overline{\gamma}^2\right] \tag{2.4.7}$$

for $\omega^m \in \Omega_m^{(1)}$.

Finally, by the Lipschitz continuity of ∇G (cf. (2.4.3)), when taking expectation w.r.t. the distribution of ω_i, we get

$$\begin{aligned}\mathbf{E}_{i-1}\{v_i^2\} &\le 4R^2\mathbf{E}_{i-1}\{\|\nabla G(x_{i-1},\omega_i) - \nabla G(x_*,\omega_i)\|_*^2\} \\ &\le 4R^2\nu\mathbf{E}_{i-1}\{\langle \nabla G(x_{i-1},\omega_i) - \nabla G(x_*,\omega_i), x_{i-1} - x_*\rangle\} = 4R^2\nu\langle \nabla g(x_{i-1}), x_{i-1} - x_*\rangle.\end{aligned}$$

On the other hand, one also has $|v_i| \le 2\nu\|x_{i-1}-x_i\|^2 \le 8\nu R^2$. We can now apply Lemma 2.4.2 with $\sigma_i^2 = 4\gamma_{i-1}^2 R^2\nu\langle\nabla g(x_{i-1}), x_{i-1}-x_*\rangle$ to conclude that for $t \ge 2\sqrt{2+\ln m}$

$$r_m^{(2)} \le 4\underbrace{\sqrt{tR^2\nu\sum_{i=0}^{m-1}\gamma_i^2\langle\nabla g(x_i), x_i-x_*\rangle}}_{=:\Delta_m^{(2)}} + 16t\nu R^2\overline{\gamma}$$

for all $\omega^m \in \Omega_m^{(2)}$ such that $\mathrm{Prob}(\Omega_m^{(2)}) \ge 1-2e^{-t}$. Note that

$$\Delta_m^{(2)} \le 2tR^2 + \tfrac{1}{4}\nu\sum_{i=0}^{m-1}\gamma_i^2\langle\nabla g(x_i), x_i-x_*\rangle,$$

and $\gamma_i \le (4\nu)^{-1}$, so that

$$r_m^{(2)} \le \nu\sum_{i=0}^{m-1}\gamma_i^2\langle\nabla g(x_i), x_i-x_*\rangle + 12tR^2 \le \tfrac{1}{4}\sum_{i=0}^{m-1}\gamma_i\langle\nabla g(x_i), x_i-x_*\rangle + 12tR^2 \tag{2.4.8}$$

for $\omega^m \in \Omega_m^{(2)}$.

3^o. When substituting bounds (2.4.6)–(2.4.8) into (2.4.5) we obtain

$$\begin{aligned} R_m &\le \tfrac{1}{4}\sum_{i=0}^{m-1}\gamma_i\langle\nabla g(x_i), x_i-x_*\rangle + 12tR^2 + \sigma_*^2\left[4\sum_{i=0}^{m-1}\gamma_i^2 + 24t\overline{\gamma}^2\right] + 2\sqrt{3tR^2\sigma_*^2\sum_{i=0}^{m-1}\gamma_i^2} \\ &\le \tfrac{1}{4}\sum_{i=0}^{m-1}\gamma_i\langle\nabla g(x_i), x_i-x_*\rangle + 15tR^2 + \sigma_*^2\left[7\sum_{i=0}^{m-1}\gamma_i^2 + 24t\overline{\gamma}^2\right] \end{aligned}$$

for all $\omega^m \in \overline{\Omega}_m = \bigcap_{i=1}^3\Omega_m^{(i)}$ with $\mathrm{Prob}(\overline{\Omega}_m) \ge 1-4e^{-t}$ and $t \ge 2\sqrt{2+\ln m}$.

When substituting the latter bound into (2.4.4) and utilizing the convexity of g and h we arrive at

$$\begin{aligned} &\left(\sum_{i=0}^{m-1}\gamma_i\right)\left(\tfrac{1}{2}[g(\widehat{x}_m)-g(x_*)] + [h(\widehat{x}_m)-h(x_*)]\right) \le \sum_{i=0}^{m-1}\gamma_i\left(\tfrac{1}{2}[g(x_i)-g(x_*)] + [h(x_i)-h(x_*)]\right) \\ &\le \sum_{i=1}^{m}\gamma_{i-1}\left(\tfrac{1}{2}\langle\nabla g(x_{i-1}), x_{i-1}-x_*\rangle + [h(x_{i-1})-h(x_*)]\right) \\ &\le V(x_0,x_*) + 15tR^2 + \sigma_*^2\left[7\sum_{i=0}^{m-1}\gamma_i^2 + 24t\overline{\gamma}^2\right] + \gamma_0[h(x_0)-h(x_*)] - \gamma_m[h(x_m)-h(x_*)]. \end{aligned}$$

In particular, for constant stepsizes $\gamma_i \equiv \gamma$ we get

$$\begin{aligned} &\tfrac{1}{2}[g(\widehat{x}_m)-g(x_*)] + [h(\widehat{x}_m)-h(x_*)] \\ &\quad\le \frac{V(x_0,x_*)+15tR^2}{\gamma m} + \frac{h(x_0)-h(x_m)}{m} + \gamma\sigma_*^2\left(7+\frac{24t}{m}\right). \end{aligned}$$

This implies the first statement of the proposition.

5^o. To prove the bound for the minibatch solution $\widehat{x}_m^{(L)} = \left(\sum_{i=0}^{m-1}\gamma_i\right)^{-1}\sum_{i=0}^{m-1}\gamma_i x_i^{(L)}$, it suffices to note that minibatch gradient observation $H(x,\omega^{(L)})$ is Lipschitz-continuous with Lipschitz constant ν, and that $H(x_*,\omega^{(L)})$ is sub-Gaussian with parameter σ_*^2 replaced with $\overline{\sigma}_{*,L}^2 \lesssim \frac{\Theta\sigma_*^2}{L}$, see Lemma 2.4.3. □

2.4.2 Deviation inequalities

Let us assume that $(\xi_i,\mathcal{F}_i)_{i=1,2,...}$ is a sequence of sub-Gaussian random variables satisfying[5]

$$\mathbf{E}_{i-1}\left\{e^{t\xi_i}\right\} \le e^{t\mu_i+\frac{t^2s_i^2}{2}}, \quad a.s. \tag{2.4.9}$$

for some *nonrandom* μ_i, s_i, $s_i \le \overline{s}$. We denote by $S_n = \sum_{i=1}^n \xi_i - \mu_i$, $r_n = \sum_{i=1}^n s_i^2$, $v_n = \sum_{i=1}^n s_i^4, M_n = \sum_{i=1}^n \xi_i^2 - (s_i^2+\mu_i^2)$, and $h_n = \sum_{i=1}^n 2\mu_i^2 s_i^2$. The following well known result is provided for reader's convenience.

Lemma 2.4.1 *For all $x>0$ one has*

$$\text{Prob}\left\{S_n \ge \sqrt{2xr_n}\right\} \le e^{-x}, \tag{2.4.10a}$$

$$\text{Prob}\left\{M_n \ge 2\sqrt{x(v_n+h_n)} + 2x\overline{s}^2\right\} \le e^{-x}. \tag{2.4.10b}$$

Proof. The inequality (2.4.10a) is straightforward. To prove (2.4.10b), note that for $t < \frac{1}{2}\overline{s}^{-2}$ and $\eta\sim\mathcal{N}(0,1)$ independent of $\xi_0,...,\xi_n$, we have:

$$\begin{aligned}\mathbf{E}_{i-1}\left\{e^{t\xi_i^2}\right\} &= \mathbf{E}_{i-1}\left\{\mathbf{E}_\eta\left\{e^{\sqrt{2t}\xi_i\eta}\right\}\right\} = \mathbf{E}_\eta\left\{\mathbf{E}_{i-1}\left\{e^{\sqrt{2t}\xi_i\eta}\right\}\right\}\\ &\le \mathbf{E}_\eta\left\{\exp\left\{\sqrt{2t}\eta\mu_i + t\eta^2 s_i^2\right\}\right\} = (1-2ts_i^2)^{-1/2}\exp\left\{\frac{t\mu_i^2}{1-2ts_i^2}\right\} \text{ a.s.},\end{aligned}$$

and because, cf [116, Lemma 1],

$$-\tfrac{1}{2}\ln(1-2ts_i^2) + \frac{t\mu_i^2}{1-2ts_i^2} - t(s_i^2+\mu_i^2) \le \frac{t^2s_i^2(s_i^2+2\mu_i^2)}{1-2ts_i^2} \le \frac{t^2s_i^2(s_i^2+2\mu_i^2)}{1-2t\overline{s}^2},$$

one has for $t < \frac{1}{2}\overline{s}^{-2}$

$$\mathbf{E}\left\{e^{tM_n}\right\} \le \exp\left\{\frac{t^2(v_n+h_n)}{1-2t\overline{s}^2}\right\}.$$

By Lemma 8 of [117], this implies that

$$\text{Prob}\left\{M_n \ge 2\sqrt{x(v_n+h_n)} + 2x\overline{s}^2\right\} \le e^{-x}$$

for all $x>0$. □

Now, suppose that ζ_i, $i=1,2,...$ is a sequence of random variables satisfying

$$\mathbf{E}_{i-1}\{\zeta_i\} = \mu_i,\ \mathbf{E}_{i-1}\{\zeta_i^2\} \le \sigma_i^2,\ \ |\zeta_i| \le 1 \text{ a.s.} \tag{2.4.11}$$

Denote $M_n = \sum_{i=1}^n[\zeta_i - \mu_i]$ and $q_n = \sum_{i=1}^n \sigma_i^2$. Note that $q_n \le n$.

[5]Here, same as above, we denote $\mathbf{E}_{i-1}$ the expectation conditional to $\mathcal{F}_{i-1}$.

Lemma 2.4.2 *Let $x \geq 1$; one has*

$$\mathrm{Prob}\left\{M_n \geq \sqrt{2xq_n} + x\right\} \leq \left[e\left(2x\ln\left[\frac{9n}{2x}\right] + 1\right) + 1\right]e^{-x}.$$

In particular, for $x \geq 4\sqrt{2+\ln n}$ one has

$$\mathrm{Prob}\left\{M_n \geq \sqrt{2xq_n} + x\right\} \leq 2e^{-x/2}.$$

Proof. In the premise of the lemma, applying Bernstein's inequality for martingales [113, 118] we obtain for all $x > 0$ and $u > 0$,

$$\mathrm{Prob}\left\{M_n \geq \sqrt{2xu} + \frac{x}{3},\, q_n \leq u\right\} \leq e^{-x}.$$

We conclude that

$$\mathrm{Prob}\left\{M_n \geq x,\, q_n \leq \frac{2x}{9}\right\} \leq e^{-x},$$

and for any $u > 0$

$$\mathrm{Prob}\left\{M_n \geq \sqrt{2(x+1)q_n} + \frac{x}{3},\, u \leq q_n \leq u\big(1+1/x\big)\right\} \leq e^{-x},$$

so that

$$\delta_n(x;u) := \mathrm{Prob}\left\{M_n \geq \sqrt{2xq_n} + \frac{x}{3},\, u \leq q_n \leq u\big(1+1/x\big)\right\} \leq e^{-x+1}.$$

Let now $u_0 = 2x/9$, $u_j = \min\{n, (1+1/x)^j u_0\}$, $j = 0, ..., J$, with

$$J = \left\rfloor \ln\left[n/u_0\right]\ln^{-1}[1+1/x]\right\lfloor.$$

Note that $\ln[1+1/x] \geq 1/(2x)$ for $x \geq 1$, so that

$$J \leq \ln\left[n/u_0\right]\ln^{-1}[1+1/x] + 1 \leq 2x\ln\left[n/u_0\right] + 1.$$

On the other hand,

$$\begin{aligned}\mathrm{Prob}\left\{M_n \geq \sqrt{2xq_n} + x\right\} &\leq e^{-x} + \sum_{j=1}^{J}\delta_n(x;u_j) \leq e^{-x} + Je^{-x+1}\\ &\leq \left[e\Big(2x\ln\left[\frac{9n}{2x}\right] + 1\Big) + 1\right]e^{-x}\end{aligned}$$

Finally, we verify explicitly that for $x \geq 4\sqrt{2+\ln n}$ one has

$$\left[e\Big(2x\ln\left[\frac{9n}{2x}\right] + 1\Big) + 1\right]e^{-x/2} \leq 2,$$

implying that for such x

$$\mathrm{Prob}\left\{M_n \geq \sqrt{2xq_n} + x\right\} \leq 2e^{-x/2}. \qquad \square$$

Let $(\xi_i)_{i=1,...}$ be a sequence of independent random vectors in $\mathbf{R}^n$ such that

$$\mathbf{E}_{i-1}\left\{\exp\left(\frac{\|\xi_i\|_*^2}{s^2}\right)\right\} \leq \exp(1), \quad \mathbf{E}_{i-1}\left\{\xi_i\right\} = 0,$$

and let $\eta = \sum_{i=1}^m \xi_i$, $m \in \mathbf{Z}_+$. We are interested in "sub-Gaussian characteristics" of r.v. $\zeta = \langle u, \eta\rangle$ for some $u \in \mathbf{R}^n$, $\|u\| \le R$, and of $\tau = \|\eta\|_*$.

Because $\mathbf{E}\{\langle u, \xi_i\rangle\} = 0$ and $|\langle u, \xi_i\rangle| \le \|u\|\,\|\xi_i\|_*$, for all t one has (cf.,e.g., Proposition 4.2 of [115])

$$\mathbf{E}\Big\{e^{t\langle u,\eta\rangle}\Big\} = \prod_{i=1}^m \mathbf{E}\Big\{e^{t\langle u,\xi_i\rangle}\Big\} \le \prod_{i=1}^m \exp\left(\tfrac{3}{4}t^2s^2\right) = \exp\left(\tfrac{3}{4}mt^2s^2\right).$$

Let ξ_ℓ, $\ell = 1, 2, ...$ be a sequence of independent random vectors $\xi_\ell \in E$, such that $\mathbf{E}\{\xi_\ell\} = 0$ and $\mathbf{E}\Big\{e^{\|\xi_\ell\|_*^2/s^2}\Big\} \le \exp(1)$. Denote $\eta_j = \sum_{\ell=1}^j \xi_\ell$. We have the following result.

Lemma 2.4.3

$$\forall L \in \mathbf{Z}_+ \quad \mathbf{E}\left\{\exp\left(\frac{\|\eta_L\|_*^2}{18\Theta s^2 L}\right)\right\} \le \exp(1) \tag{2.4.12}$$

where $\Theta = \max_{\|z\|\le 1}\theta(z)$ for the d.-g.f. θ of the unit ball of norm $\|\cdot\|$ in E, as defined in Section 2.2.2.

Proof. Let for $\eta \in E$, $\pi(\eta) = \sup_{\|z\|\le 1}[\langle \eta, z\rangle - \theta(z)]$. Observe that for all $\beta > 0$,

$$\|\eta_L\|_* = \sup_{\|z\|\le 1}\langle \eta_L, z\rangle \le \max_{\|z\|\le 1} \beta\theta(z) + \beta\pi(\eta_L/\beta) \le \beta\Theta + \beta\pi\big(\frac{\eta_L}{\beta}\big). \tag{2.4.13}$$

On the other hand, we know (cf. [119, Lemma 1]) that π is smooth with $\|\nabla\pi\| \le 1$, and $\nabla\pi$ is Lipschitz-continuous w.r.t. to $\|\cdot\|_*$, i.e.,

$$\|\nabla\pi(z) - \nabla\pi(z')\| \le \|z - z'\|_* \quad \forall z, z' \in E.$$

As a consequence of Lipschitz continuity of π, when denoting $\pi_\beta(\eta) = \beta\pi\big(\frac{\eta}{\beta}\big)$, we have

$$\pi_\beta(\eta_{j-1} + \xi_j) - \pi_\beta(\eta_{j-1}) \le \|\xi_j\|_*,$$

so that $\mathbf{E}\left\{\exp\left([\pi_\beta(\eta_j) - \pi_\beta(\eta_{j-1})]^2/s^2\right)\right\} \le \exp(1)$. Furthermore,

$$\pi_\beta(\eta_{j-1} + \xi_j) \le \pi_\beta(\eta_{j-1}) + \langle \nabla\pi_\beta(\eta_{j-1}), \xi_j\rangle + \tfrac{\|\xi_j\|_*^2}{2\beta},$$

and, because η_{j-1} does not depend on ξ_j and $\mathbf{E}\{\|\xi_j\|_*^2\} \le s^2$, we get

$$\mathbf{E}_{j-1}\{\pi_\beta(\eta_j) - \pi_\beta(\eta_{j-1})\} \le \tfrac{s^2}{2\beta}.$$

By [115, Proposition 4.2] we conclude that random variables $\delta_j = \pi_\beta(\eta_j) - \pi_\beta(\eta_{j-1})$ satisfy for all $t \ge 0$,

$$\mathbf{E}_{j-1}\left\{e^{t\delta_j}\right\} \le \exp\left(\tfrac{1}{2}ts^2\beta^{-1} + \tfrac{3}{4}t^2s^2\right).$$

Consequently,

$$\mathbf{E}\left\{e^{t\pi_\beta(\eta_L)}\right\} \le \mathbf{E}\left\{e^{t\pi_\beta(\eta_{L-1})}\right\}\exp\left(\tfrac{1}{2}ts^2\beta^{-1} + \tfrac{3}{4}t^2s^2\right) \le \exp\left(\tfrac{1}{2}ts^2L\beta^{-1} + \tfrac{3}{4}t^2s^2L\right).$$

When substituting the latter bound into (2.4.13), we obtain for $\beta^2 = \frac{s^2L}{2\Theta}$

$$\mathbf{E}\left\{e^{t\|\eta_L\|_*}\right\} \le \exp\left(ts\sqrt{2\Theta L} + \tfrac{3}{4}t^2s^2L\right) \quad \forall t \ge 0. \tag{2.4.14}$$

To complete the proof of the lemma, it remains to show that (2.4.14) implies (2.4.12). This is straightforward. Indeed, for $\chi \sim \mathcal{N}(0,1)$, $\alpha > 0$ and $\zeta = \|\eta_L\|_*$ one has

$$\begin{aligned}\mathbf{E}\left\{e^{\alpha\zeta^2}\right\} &= \mathbf{E}\left\{\mathbf{E}_\eta\left(e^{\sqrt{2\alpha}\zeta\chi}\right)\right\} = \mathbf{E}_\chi\left\{\mathbf{E}\left\{e^{\sqrt{2\alpha}\zeta\chi}\right\}\right\}\\ &\le \mathbf{E}_\chi\left\{\exp\left(2\sqrt{\alpha\Theta L}\, s\chi + \tfrac{3}{2}\alpha L s^2\chi^2\right)\right\} = (1-3\alpha L s^2)^{-1/2}\exp\left\{\frac{4\alpha\Theta L s^2}{1-3\alpha L s^2}\right\}\end{aligned}$$

When setting $\alpha = (18\Theta s^2 L)^{-1}$, we conclude that

$$\mathbf{E}\left\{e^{\alpha\zeta^2}\right\} \le \exp(1)$$

due to $\Theta \ge 1/2$. □

2.4.3 Proof of Theorem 2.2.1

We start with analysing the behaviour of the approximate solution $\widehat{x}^k_{m_0}$ at the stages of the preliminary phase of the procedure.

Lemma 2.4.4 *Let $m_0 = \lceil 64\delta^2\rho\nu s(4\Theta + 60t)\rceil$ (here $\lceil a\rceil$ stands for the smallest integer greater or equal to a), $\gamma = (4\nu)^{-1}$, and let t satisfy $t \ge 4\sqrt{2+\log(m_0)}$.*

Suppose that $R \ge 2\delta\sigma_*\sqrt{6\rho s/\nu}$, that initial condition x_0 of Algorithms 1 and 2 satisfies $\|x_0 - x_*\| \le R$, and that at the stage k of the preliminary phase we choose

$$\kappa_k = R_{k-1}\sqrt{\frac{\nu(4\Theta+60t)}{\rho s m_0}} \tag{2.4.15}$$

where $(R_k)_{k\ge 0}$ is defined recursively:

$$R_{k+1} = \tfrac{1}{2}R_k + \frac{16\sigma_*^2\delta^2\rho s}{\nu R_k}, \quad R_0 = R.$$

Then the approximate solution $\widehat{x}^k_{m_0}$ at the end of the kth stage of the CSMD-SR algorithm satisfies, with probability $\ge 1 - 4ke^{-t}$

$$\|\widehat{x}^k_{m_0} - x_*\| \quad \le R_k \le 2^{-k}R + 4\sigma_*\delta\sqrt{2\rho s/\nu}. \tag{2.4.16}$$

In particular, the estimate $\widehat{x}^{\overline{K}_1}_{m_0}$ after $\overline{K}_1 = \left\lceil \frac{1}{2}\log_2\left(\frac{R^2\nu}{32\sigma_*^2\delta^2\rho s}\right)\right\rceil$ stages satisfies with probaility at least $1 - 4\overline{K}_1 e^{-t}$

$$\|\overline{x}^{\overline{K}_1}_{m_0} - x_*\| \le 8\sigma_*\delta\sqrt{2\rho s/\nu}. \tag{2.4.17}$$

Proof of the lemma.
1^o. Note that initial point x_0 satisfies $x_0 \in X_R(x_*)$. Suppose that the initial point $x^k_0 = \widehat{x}^{k-1}_{m_0}$ of the kth stage of the method satisfy $x^k_0 \in X_{R_{k-1}}(x_*)$ with probability $1-4(k-1)e^{-t}$. In other words, there is a set $\mathcal{B}_{k-1} \subset \Omega$, $\mathrm{Prob}(\mathcal{B}_{k-1}) \ge 1-4(k-1)e^{-t}$, such that for all $\overline{\omega}^{k-1} = [\omega_1; ...; \omega_{m_0(k-1)}] \subset \mathcal{B}_{k-1}$ one has $x^k_0 \in X_{R_{k-1}}(x_*)$. Let us show that upon termination of the $k$the stage $\widehat{x}^k_{m_0}$ satisfy $\|x^k_{m_0} - x_*\| \le R_k$ with probability $1 - 4ke^{-t}$. By Proposition 2.4.1 (with $h(x) = \kappa_k\|x\|$) we conclude that for some $\overline{\Omega}_k \subset \Omega$, $\mathrm{Prob}(\overline{\Omega}_k) \ge 1 - 4e^{-t}$, solution $\widehat{x}^k_{m_0}$ after m_0 iterations of the stage satisfies, for all for all $\omega^k = [\omega_{(k-1)m_0+1}, ..., \omega_{km_0}] \in \overline{\Omega}_k$,

$$F(\widehat{x}^k_{m_0}) - F(x_*) \le \tfrac{1}{m_0}\left(\nu R^2_{k-1}(4\Theta+60t) + \kappa_k R_{k-1}\right) + \frac{\sigma_*^2}{\nu}\left(\tfrac{7}{4} + \tfrac{6t}{m_0}\right).$$

When using the relationship (2.2.11) of Assumption [RSC] we now get

$$\|\widehat{x}^k_{m_0} - x_*\| \le \delta\left[\rho s\kappa_k + \frac{R_{k-1}}{m_0} + \frac{\nu R^2_{k-1}}{\kappa_k m_0}(4\Theta + 60t) + \frac{\sigma^2_*}{\nu\kappa_k}\left(\tfrac{7}{4} + \tfrac{6t}{m_0}\right)\right]. \tag{2.4.18}$$

Note that κ_k as defined in (2.4.15) satisfies $\kappa_k \le R_{k-1}(8\delta\rho s)^{-1}$, while $\kappa_k m_0 \ge 8\delta(4\Theta + 60t)R_{k-1}\nu$. Because $m_0 \ge 3840t$ due to $\rho\nu \ge 1$ and $\delta \ge 1$, one also has $\left(\frac{7}{4} + \frac{6t}{m_0}\right)\kappa_k^{-1} < 16\delta\rho s/R_{k-1}$. When substituting the above bounds into (2.4.18) we obtain

$$\|\widehat{x}^k_{m_0} - x_*\| \le \delta R_{k-1}\left(\tfrac{1}{4\delta} + \tfrac{1}{m_0}\right) + \frac{16\delta^2\rho s\sigma^2_*}{R_{k-1}\nu} \le \tfrac{1}{2}R_{k-1} + \frac{16\delta^2\rho s\sigma^2_*}{R_{k-1}\nu} = R_k. \tag{2.4.19}$$

We conclude that $\widehat{x}^k_{m_0} \in X_{R_k}(x_*)$ for all $\overline{\omega}^k \in \mathcal{B}_k = \mathcal{B}_{k-1} \cap \overline{\Omega}_k$, and

$$\mathrm{Prob}(\mathcal{B}_k) \ge \mathrm{Prob}(\mathcal{B}_{k-1}) - \mathrm{Prob}(\overline{\Omega}^c_k) \ge 1 - 4ke^{-t}.$$

2^o. Let now $a = 16\delta^2\rho s\sigma^2_*/\nu$, and let us study the behaviour of the sequence

$$R_k = \frac{R_{k-1}}{2} + \frac{a}{R_{k-1}} =: f(R_{k-1}), \quad R_0 = R \ge \sqrt{2a}.$$

Function f admits a fixed point at $R = \sqrt{2a}$ which is also the minimum of f, so one has $R_k \ge \sqrt{2a}$ $\forall k$. Thus,

$$d_k := R_k - \sqrt{2a} = \frac{R_{k-1} - \sqrt{2a}}{2} + \frac{2a - \sqrt{2a}R_{k-1}}{2R_{k-1}} \le \tfrac{1}{2}d_{k-1} \le 2^{-k}d_0 \le 2^{-k}(R - \sqrt{2a}).$$

We deduce that $R_k \le 2^{-k}R_0 + \sqrt{2a}$ which is (2.4.16). Finally, after running $\overline{K}_1$ stages of the preliminary phase, the estimate $\widehat{x}^{\overline{K}_1}_{m_0}$ satisfies

$$\|\widehat{x}^{\overline{K}_1}_{m_0} - x_*\| \le 8\delta\sigma_*\sqrt{2\rho s/\nu}. \qquad \square$$

We turn next to the analysis of the asymptotic phase of Algorithm 2. We assume that the preliminary phase of the algorithm has been completed.

Lemma 2.4.5 *Let t be such that $t \ge 4\sqrt{2 + \log(m_1)}$, with $m_1 = \lceil 81\delta^2\rho s\nu(4\Theta + 60t)\rceil$, $\gamma = (4\nu)^{-1}$, and let $\ell_k = \lceil 10 \times 4^{k-1}\Theta\rceil$. We set*

$$\kappa_k = r_{k-1}\sqrt{\frac{\nu(4\Theta + 60t)}{\rho s m_1}}, \quad r_k = 2^{-k}r_0, \quad r_0 = 8\delta\sigma_*\sqrt{2\rho s/\nu}.$$

Then the approximate solution by Algorithm 2 $\widehat{x}^k_{m_1}$ at the end of the kth stage of the asymptotic phase satisfies, with probability $\ge 1 - 4(\overline{K}_1 + k)e^{-t}$, $\|\widehat{x}^k_{m_1} - x_*\| \le r_k$, implying that

$$\|\widehat{x}^k_{m_1} - x_*\| \lesssim \delta^2\sigma_*\rho s\sqrt{\frac{\Theta\,(\Theta + t)}{N_k}}, \tag{2.4.20}$$

where $N_k = m_1\sum_{i=1}^k \ell_i$ is the total count of oracle calls for k asymptotic stages.

Proof of the lemma. Upon terminating the preliminary phase, the initial condition $x_0 = \widehat{x}^{\overline{K}_1}_{m_0}$ of the asymptotic phase satisfies (2.4.17) with probability greater or equal to $1 - 4\overline{K}_1 e^{-t}$. We are about to show that $\forall k \geq 1$, with probability at least $1 - 4(\overline{K}_1 + k)e^{-t}$,

$$\|\widehat{x}^k_{m_1} - x_*\| \leq r_k = 2^{-k} r_0, \quad r_0 = 8\delta\sigma_*\sqrt{2\rho s/\nu}.$$

The claim is obviously true for $k = 0$. Let let us suppose that it holds at stage $k - 1 \geq 0$, and let us prove that it also holds at stage k. To this end, we reproduce the argument used in the proof of Lemma 2.4.4, while taking into account that now ℓ_k observations are averaged at each iteration of the CSMD algorithm. Recall (cf. Lemma 2.4.3) that this amounts to replacing sub-Gaussian parameter σ_*^2 with $\overline{\sigma}_*^2 = 18\Theta\sigma_*^2/\ell_k$. When applying the result of Proposition 2.4.1 and the bound of (2.2.11) we conclude (cf. (2.4.18)) that, with probability $1 - (\overline{K}_1 + k)e^{-t}$,

$$\|\widehat{x}^k_{m_1} - x_*\| \leq \delta\left[\rho s\kappa_k + \frac{r_{k-1}}{m_1} + \frac{\nu r_{k-1}^2}{\kappa_k m_1}(4\Theta + 60t) + \frac{18\Theta\sigma_*^2}{\nu\kappa_k\ell_k}\left(\tfrac{7}{4} + \tfrac{6t}{m_1}\right)\right]$$

By simple algebra, we obtain the following analogue of (2.4.19):

$$\|\widehat{x}^k_{m_1} - x_*\| < \delta r_{k-1}\left(\tfrac{2}{9\delta} + \tfrac{1}{m_1}\right) + 10\frac{4^{-k+1}\delta^2\rho s\sigma_*^2}{r_{k-1}\nu} < \tfrac{r_{k-1}}{4} + \tfrac{r_{k-1}}{4} = r_k.$$

Observe that upon the end of the kth stage we used $N_k = m_1\sum_{i=1}^k \ell_k < 3m_1\Theta\sum_{j=1}^k 4^{j-1} \leq 4^k\Theta m_1$ observations of the asymptotic stage. As a consequence, $4^{-k} < \Theta m_1/N_k$ and

$$r_k = 2^{-k}r_0 \lesssim \delta^2 s\rho\sigma_*\sqrt{\frac{\Theta(\Theta + t)}{N_k}}. \qquad \square$$

Assuming that the preliminary phase of Algorithm 1 was completed, we now consider the asymptotic phase of the algorithm.

Lemma 2.4.6 *Let $t \geq 4\sqrt{2 + \log m_k}$, $m_k = \lceil 4^{k+4}(4\Theta + 60t)\delta^2\rho s\nu\rceil$,*

$$\gamma^k = \frac{r_{k-1}}{2\sigma_*}\sqrt{\frac{(4\Theta + 60t)}{2m_k}}, \quad \kappa_k^2 = \frac{5\sigma_* r_{k-1}}{\rho s}\sqrt{\frac{(4\Theta + 60t)}{m_k}} \tag{2.4.21}$$

where

$$r_k := 2^{-k}r_0, \quad r_0 = 8\delta\sigma_*\sqrt{2\rho s/\nu}.$$

Then the approximate solution $\widehat{x}^k_{m_k}$ upon termination of the kth asymptotic stage satisfies with probability $\geq 1 - 4(\overline{K}_1 + k)e^{-t}$

$$\|\widehat{x}^k_{m_k} - x_*\| \leq 2^{-k}r_0 \lesssim \ 2^{-k}\sigma_*\delta\sqrt{\rho s\nu^{-1}} \lesssim \ \delta^2\sigma_*\rho s\sqrt{\tfrac{\Theta+t}{N_k}} \tag{2.4.22}$$

where $N_k = \sum_{j=1}^k m_j$ is the total iteration count of k stages of the asymptotic phase.

Proof of the lemma.

We are about to show that $\forall k \geq 0$, $\|\widehat{x}^k_{m_k} - x_*\| \leq r_k$ with probability $\geq 1 - 4(\overline{K}_1 + k)e^{-t}$ is true. By Lemma 2.4.4, the claim is true for $k = 0$ (at the start of the asymptotic phase, the initial condition $x_0 = \widehat{x}^{\overline{K}_1}_{m_0}$ satisfies the bound (2.4.17)). We now assume it to hold for $k - 1 \geq 0$, our objective is to implement the recursive step $k - 1 \to k$ of the proof. First, observe that the choice of

γ^k in (2.4.21) satisfies $\gamma^k \le (4\nu)^{-1}$, $k = 1, ...$, so that Proposition 2.4.1 can be applied. From the result of the proposition and bound (2.2.11) we conclude (cf. (2.4.18)) that it holds, with probability $1 - (\overline{K}_1 + k)e^{-t}$,

$$\|\widehat{x}^k_{m_k} - x_*\| \le \delta \left[\rho s \kappa_k + \frac{r_{k-1}}{m_k} + \frac{r^2_{k-1}\,(4\Theta + 60t)}{\gamma^k \kappa_k m_k} + 8\frac{\gamma^k \sigma_*^2}{\kappa_k}\right]$$

When substituting the value of γ^k from (2.4.21) we obtain

$$\|\widehat{x}^k_{m_k} - x_*\| \le \delta \left[\rho s \kappa_k + \frac{r_{k-1}}{m_k} + \frac{4\sigma_* r_{k-1}}{\kappa_k}\sqrt{\frac{2(4\Theta + 60t)}{m_k}}\right],$$

which, by the choice of κ_k in (2.4.21), results in results in

$$\|\widehat{x}^k_{m_k} - x_*\|^2 \le 2\delta^2 \left[10\rho s \sigma_* r_{k-1}\sqrt{\frac{4\Theta + 60t}{m_k}} + \frac{r^2_{k-1}}{m_k^2}\right] \le \frac{r^2_{k-1}}{4} = r_k^2.$$

It remains to note that the total number $N_k = \sum_{j=1}^k m_j$ of iterations during k stages of the asymptotic phase satisfies $N_k \lesssim 4^k(\Theta + t)\delta^2 \rho s \nu$, and $2^{-k} \lesssim \delta\sqrt{\frac{(\Theta+t)\rho s\nu}{N_k}}$, which along with definition of r_0 implies (2.4.22). □

Proof of Theorem 2.2.1. We can now terminate the proof of the theorem. Let us prove the accuracy bound of the theorem for the minibatch variant of the procedure.

Assume that the "total observation budget" N is such that only the preliminary phase of the procedure is implemented. This is the case when either $m_0\overline{K}_1 \ge N$, or $m_0\overline{K}_1 < N$ and $m_0\overline{K}_1 + m_1\ell_1 > N$. The output $\widehat{x}_N$ of the algorithm is then the last update of the preliminary phase, and by Lemma 2.4.4 it satisfies $\|\widehat{x}_N - x_*\| \le R2^{-k}$ where k is the count of completed stages. In the case of $m_0\overline{K}_1 \ge N$ this clearly implies that (recall that $N \ge m_0$) that $k \ge cN/m_0$ and, with probability $\ge 1 - 4ke^{-t}$

$$\|\widehat{x}_N - x_*\| \lesssim R\exp\left\{-\frac{c'N}{\delta^2\rho s\nu(\Theta + t)}\right\}. \tag{2.4.23}$$

On the other hand, when $m_0\overline{K}_1 < N < m_0\overline{K}_1 + m_1\ell_1$, by definition of m_1 and ℓ_1, one has $N \le Cm_0\overline{K}_1$, so that bound (2.4.23) still holds in this case.

Now, consider the case where at least one asymptotic stage has been completed. When $m_0\overline{K}_1 > \frac{N}{2}$ we still have $N \le Cm_0\overline{K}_1$, so that the bound (2.4.23) holds for the approximate solution $\widehat{x}^{(b)}_N$ at the end of the asymptotic stage. Otherwise, the number of oracle calls N_k of asymptotic stages satisfies $N_k \ge N/2$, and by (2.4.20) this implies that with probability $\ge 1 - 4(\overline{K}_1 + \overline{K}_2)e^{-t}$,

$$\|\widehat{x}^{(b)}_N - x_*\| \lesssim \delta^2\sigma_*\rho s\sqrt{\frac{\Theta(\Theta + t)}{N}}.$$

To summarize, in both cases, the bound of Theorem 2.2.1 holds with probability at least $1 - 4(\overline{K}_1 + \overline{K}_2)e^{-t}$.

The proof of the accuracy bound for the "standard" solution $\widehat{x}_N$ is completely analogous, making use of the bound (2.4.22) of Lemma 2.4.6 instead of (2.4.20). □

Remark 2.4.1 *Theorem 2.2.1 as stated in Section 2.2.3 does not say anything about convergence of $g(\widehat{x}_N)$ to $g(x_*)$. Such information can be easily extracted from the proof of the theorem. Indeed, observe that at the end of a stage of the method, one has, with probability $1-Cke^{-t}$,*

$$F_{\kappa_k}(\widehat{x}^k) - F_{\kappa_k}(x_*) \le \upsilon_k,$$

or

$$g(\widehat{x}^k) - g(x_*) \le \upsilon_k + \kappa_k(\|\widehat{x}^k\| - \|x_*\|) \le \upsilon_k + \kappa_k\|\widehat{x}^k - x_*\|$$

where $\widehat{x}^k$ is the approximate solution at the end of the stage k. One the other hand, at the end of the kth stage of the preliminary *phase one has $\|\widehat{x}^k - x_*\| \le R_k \le 2^{-k}R$, with $\kappa_k \lesssim R_k(\delta\rho s)^{-1} \le 2^{-k}R(\delta\rho s)^{-1}$ and $\upsilon_k \lesssim \frac{4^{-k}R^2}{\delta^2\rho s}$ implying that*

$$g(\widehat{x}^k) - g(x_*) \lesssim \upsilon_k + \frac{R_k^2}{\delta^2\rho s} \lesssim (\delta^{-2}+\delta^{-1})\frac{R^2}{\rho s}\exp\left\{-\frac{c}{\delta\rho\nu}\frac{N}{s(\Theta+t)}\right\}$$

where N is the current iteration count. Furthermore, at the end of the kth asymptotic *stage, one has, with probability $1-(\overline{K}_1+k)e^{-t}$, $\|\widehat{x}^k - x_*\| \le R_k \lesssim \delta^2\sigma_*\rho s\sqrt{\frac{\Theta+t}{m_k}}$, while $\kappa_k \asymp 2^{-k}\delta\sigma_*(\rho\nu s)^{-1/2} \lesssim \delta\sigma_*\sqrt{\frac{\Theta+t}{m_k}}$, and $\upsilon_k \lesssim \delta^2\sigma_*^2\rho s(\Theta+t)/m_k$. As a result, the corresponding $\widehat{x}^k$ satisfies*

$$g(\widehat{x}^k) - g(x_*) \le \upsilon_k + \kappa_k\|\widehat{x}^k - x_*\| \lesssim (\delta^2+\delta^3)\rho\sigma_*^2 s\frac{\Theta+t}{m_k}.$$

When putting the above bounds together, assuming that at least 1 stage of the algorithm was completed, we arrive at the bound after N steps:

$$g(\widehat{x}_N) - g(x_*) \lesssim (\delta^{-2}+\delta^{-1})\frac{R^2}{\rho s}\exp\left\{-\frac{c}{\delta^2\rho\nu}\frac{N}{s(\Theta+t)}\right\} + (\delta^2+\delta^3)\rho s\sigma_*^2\frac{\Theta+t}{N} \tag{2.4.24}$$

with probability $1-(\overline{K}_1+\overline{K}_2)e^{-t}$.

2.4.4 Proof of Proposition 2.3.1

1^o. Recall that $\mathfrak{r}$ is $\overline{r}$-Lipschitz continuous, i.e., for all $t,t' \in \mathbf{R}^m$

$$|\mathfrak{r}(t) - \mathfrak{r}(t')| \le \overline{r}|t-t'|.$$

As a result, for all $x, x' \in X$,

$$\|\phi[\mathfrak{r}(\phi_i^T x) - \mathfrak{r}(\phi_i^T x')]\|_\infty \le \overline{r}\|\phi_i\|_\infty|\phi_i^T(x-x')| \le \overline{r}\|\phi_i\|_\infty^2\|x-x'\|_1 \le \overline{r\nu}^2\|x-x'\|_1,$$

so that $\nabla G(x,\omega) = \phi[\mathfrak{r}(\phi^T x) - \eta]$ is Lipschitz continuous w.r.t. ℓ_1-norm with Lipschitz constant $\mathcal{L}(\omega) \le \overline{r\nu}^2$.

2^o. Due to strong monotonicity of $\mathfrak{r}$,

$$\begin{aligned} g(x) - g(x_*) &= \int_0^1 \nabla g(x_* + t(x-x_*))^T(x-x_*)dt \\ &= \int_0^1 \mathbf{E}\Big\{\phi[\mathfrak{r}(\phi^T(x_*+t(x-x_*)) - \mathfrak{r}(\phi^T x_*)]\Big\}^T(x-x_*)dt \\ &\ge \int_0^1 \underline{r}\mathbf{E}\big\{(\phi^T(x-x_*))^2\big\}tdt = \tfrac{1}{2}\underline{r}\|x-x_*\|_\Sigma^2, \end{aligned}$$

what is (2.3.2).

3^o. The sub-Gaussianity in the "batchless" case is readily given by $\nabla G(x_*, \omega_i) = \sigma\phi_i\xi_i$ with $\|\phi_i\xi_i\|_\infty \le \|\phi_i\|_\infty|\xi_i| \le \overline{\nu}\|\xi_i\|_2$ and

$$\mathbf{E}\left\{\exp\left(\frac{\|\nabla G(x_*,\omega_i)\|_\infty^2}{\sigma^2\overline{\nu}^2}\right)\right\} \le e$$

due to $\mathbf{E}\{e^{\xi_i^2}\} \le \exp(1)$. Because Θ variation of the d.-g.f. θ, as defined in (2.3.6), is bounded with $C\ln n$, by Lemma 2.4.3 we conclude that batch observation

$$H\left(x_*, \omega_i^{(L)}\right) = \frac{1}{L}\sum_{\ell=1}^{L}\nabla G(x_*,\omega_i^\ell) = \frac{1}{L}\sum_{\ell=1}^{L}\sigma\phi_i^\ell, \xi_i^\ell$$

is sub-Gaussian with parameter $\lesssim \sigma^2\overline{\nu}^2\ln n$.

4^o. In the situation of Section 2.3.1, Σ is positive definite, $\Sigma \succeq \kappa_\Sigma I$, $\kappa_\Sigma > 0$, and condition $\mathbf{Q}(\lambda,\psi)$ is satisfied with $\lambda = \kappa_\Sigma$ and $\psi = 1$. Because quadratic minoration condition (2.3.3) for g is verified with $\mu \ge \underline{r}$ due to (2.3.2), when applying the result of Lemma 2.3.1, we conclude that Assumption [RSC] holds with $\delta = 1$ and $\rho = (\kappa_\Sigma\underline{r})^{-1}$.[6] □

2.4.5 Properties of sparsity structures

Sparsity structures

The scope of results of Section 2.2 is much broader than "vanilla" sparsity optimization. We discuss here general notion of *sparsity structure* which provides a proper application framework for these results.

In what follows we assume to be given a *sparsity structure* [120] on E—a family $\mathcal{P}$ of projector mappings $P = P^2$ on E such that

A1.1 every $P \in \mathcal{P}$ is assigned a linear map $\overline{P}$ on E such that $P\overline{P} = 0$ and a nonnegative weight $\pi(P)$;

A1.2 whenever $P \in \mathcal{P}$ and $f, g \in E$ such that $\|f\|_* \le 1$, $\|g\|_* \le 1$,

$$\|P^*f + \overline{P}^*g\|_* \le 1$$

where for a linear map $Q : E \to F$, $Q^* : F \to E$ is the conjugate mapping.

Following [120], we refer to a collection of the just introduced entities and *sparsity structure on* E. For a nonnegative real s we set

$$\mathcal{P}_s = \{P \in \mathcal{P} : \pi(P) \le s\}.$$

Given $s \ge 0$ we call $x \in E$ *s-sparse* if there exists $P \in \mathcal{P}_s$ such that $Px = x$.

Typically, one is interested in the following "standard examples":

1. "Vanilla (usual)" sparsity: in this case $E = \mathbf{R}^n$ with the standard inner product, $\mathcal{P}$ is comprised of projectors on all coordinate subspaces of $\mathbf{R}^n$, $\pi(P) = \mathrm{rank}(P)$, and $\|\cdot\| = \|\cdot\|_1$.

[6] We refer to Section 5.7.10 and Lemma 2.4.7 for the proof of Lemma 2.3.1.

2. Group sparsity: $E = \mathbf{R}^n$, and we partition the set $\{1, ..., n\}$ of indices into K nonoverlapping subsets $I_1, ..., I_K$, so that to every $x \in \mathbf{R}^n$ we associate blocks x^k with corresponding indices in I_k, $k = 1, ..., K$. Now $\mathcal{P}$ is comprised of projectors $P = P_I$ onto subspaces $E_I = \{[x^1, ..., x^K] \in \mathbf{R}^n : x^k = 0 \, \forall k \notin I\}$ 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.*

3. Low rank structure: in this example $E = \mathbf{R}^{p\times q}$ with, for the sake of definiteness, $p \geq q$, 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, $\overline{P}(x) = (I - P_\ell)x(I - P_r)$, and $\|\cdot\|$ is the nuclear norm $\|x\| = \sum_{i=1}^q \sigma_i(x)$ where $\sigma_1(x) \geq \sigma_2(x) \geq ... \geq \sigma_q(x)$ are singular values of x, $\|\cdot\|_*$ is the spectral norm, so that $\|x\|_* = \sigma_1(x)$, and $\pi(P) = \max[\mathrm{rank}(P_\ell), \mathrm{rank}(P_r)]$.

   In this case, for $\|f\|_* \leq 1$ and $\|g\|_* \leq 1$ one has

$$\|P^*(f)\|_* = \|P_\ell f P_r\|_* \leq 1, \quad \|\overline{P}^*(g)\|_* = \|(I - P_\ell)g(I - P_r)\|_* \leq 1,$$

   and because the images and orthogonal complements to the kernels of P and $\overline{P}$ are orthogonal to each other, $\|P^*(f) + \overline{P}^*(g)\|_* \leq 1$.

Condition $\mathbf{Q}(\lambda, \psi)$

We say that a positive semidefinite mapping $\Sigma : E \to E$ satisfies condition $\mathbf{Q}(\lambda, \psi)$ for given $s \in \mathbf{Z}_+$ if for some $\psi, \lambda > 0$ and all $P \in \mathcal{P}_s$ and $z \in E$

$$\|Pz\| \leq \sqrt{s/\lambda}\|z\|_\Sigma + \|\overline{P}z\| - \psi\|z\|. \tag{2.4.25}$$

Lemma 2.4.7 *Suppose that x_* is an optimal solution to (2.2.2) such that for some $P \in \mathcal{P}_s$, $\|(I - P)x_*\| \leq \Delta$, and that condition $\mathbf{Q}(\lambda, \psi)$ is satisfied. Furthermore, assume that objective g of (2.2.2) satisfies the following minoration condition*

$$g(x) - g(x_*) \geq \mu\big(\|x - x_*\|_\Sigma\big)$$

where $\mu(\cdot)$ is monotone increasing and convex. Then a feasible solution $\widehat{x} \in X$ to (2.2.4) such that

$$\mathrm{Prob}\left\{F_\kappa(\widehat{x}) - F_k(x_*) \leq \upsilon\right\} \geq 1 - \epsilon.$$

satisfies, with probability at least $1 - \epsilon$,

$$\|\widehat{x} - x_*\| \leq \frac{\mu^*\left(\kappa\sqrt{s/\lambda}\right) + \upsilon}{\kappa\psi} + \frac{2\Delta}{\psi} \tag{2.4.26}$$

where $\mu^* : \mathbf{R}_+ \to \mathbf{R}_+$ is conjugate to $\mu(\cdot)$, $\mu^*(t) = \sup_{u\geq 0}[tu - \mu(u)]$.

Proof. When setting $z = \widehat{x} - x_*$ one has

$$\begin{aligned}\|\widehat{x}\| = \|x_* + z\| &= \|Px_* + (I - P)x_* + z\| \geq \|Px_* + z\| - \|(I - P)x_*\| \\ &\geq \|Px_*\| + \|\overline{P}z\| - \|Pz\| - \Delta\end{aligned}$$

where we used the relation

$$\|Px_* + z\| \geq \|Px_*\| - \|Pz\| + \|\overline{P}z\|$$

(cf. Lemma 3.1 of [120] applied to $w = Px_*$). When using condition $\mathbf{Q}(\lambda, \psi)$ we obtain

$$\|\widehat{x}\| \geq \|Px_*\| - \sqrt{s/\lambda}\|z\|_\Sigma + \psi\|z\| - \Delta,$$

so that $F_k(\widehat{x}) \leq F_k(x_*) + \upsilon$ implies

$$\begin{aligned}\kappa\left(\|Px_*\| + \psi\|z\| - \Delta\right) &\leq \tfrac{1}{2}[g(x_*) - g(\widehat{x})] + \kappa\sqrt{s/\lambda}\|z\|_\Sigma + \kappa\|x_*\| + \upsilon \\ &\leq -\tfrac{1}{2}\mu(\|z\|_\Sigma) + \kappa\sqrt{s/\lambda}\|z\|_\Sigma + \kappa\|x_*\| + \upsilon \\ &\leq \tfrac{1}{2}\mu^*(2\kappa\sqrt{s/\lambda}) + \kappa\|x_*\| + \upsilon,\end{aligned}$$

and we conclude that

$$\kappa\psi\|z\| \leq \tfrac{1}{2}\mu^*(2\kappa\sqrt{s/\lambda}) + 2\kappa\Delta + \upsilon$$

due to $\|x_*\| - \|Px_*\| \leq \|(I-P)x_*\| \leq \Delta$. □

Note that when $\mu(u) = \frac{\mu}{2}u^2$, one has $\mu^*(t) = \frac{1}{2\mu}t^2$, and in the case of $\|\cdot\| = \|\cdot\|_1$, with probability $1-\epsilon$,

$$\|\widehat{x} - x_*\|_1 \leq \frac{s\kappa}{\mu\lambda\psi} + \frac{\upsilon}{\kappa\psi} + \frac{2\Delta}{\psi}.$$

This, in particular, implies bound (2.3.4) of Lemma 2.3.1.

Remark 2.4.2 *We discuss implications of condition $\mathbf{Q}(\lambda, \psi)$ and result of Lemma 2.4.7 for "usual" sparsity in Section 2.3 of the paper. Now, let us consider the case of the low rank sparsity. Let $z \in \mathbf{R}^{p\times q}$ with $p \geq q$ for the sake of definiteness. In this case, $\|\cdot\|$ is the nuclear norm, and we put $P(z) = P_\ell z P_r$ where P_ℓ and P_r are orthoprojectors of rank $s \leq q$ such that $\|(I-P)(x)\| = \|x_* - P_\ell x_* P_r\| \leq \Delta$.*[7]

Furthermore, for a $p \times q$ matrix z let us put

$$\sigma^{(k)}(z) = \sum_{i=1}^{k}\sigma_i(z),\ 1 \leq k \leq q.$$

With the sparsity parameter s being a nonnegative integer,

$$\forall(z \in \mathbf{R}^{p\times q}, P \in \mathcal{P}_s): \quad \|P(z)\| \leq \sigma^{(s)}(z), \quad \|\overline{P}(z)\| \geq \|z\| - \sigma^{(2s)}(z).$$[8]

and we conclude that in the present situation condition

$$\sigma^{(s)}(z) + \sigma^{(2s)}(z) \leq \sqrt{s/\lambda}\|z\|_\Sigma + (1-\psi)\|z\| \tag{2.4.27}$$

is sufficient for the validity of $\mathbf{Q}(\lambda, \psi)$. As a result, condition (2.4.27) with $\psi > 0$ is sufficient for applicability of the bound of Lemma 2.4.7. It may also be compared to the necessary and sufficient condition of "s-goodness of Σ" in [121]:

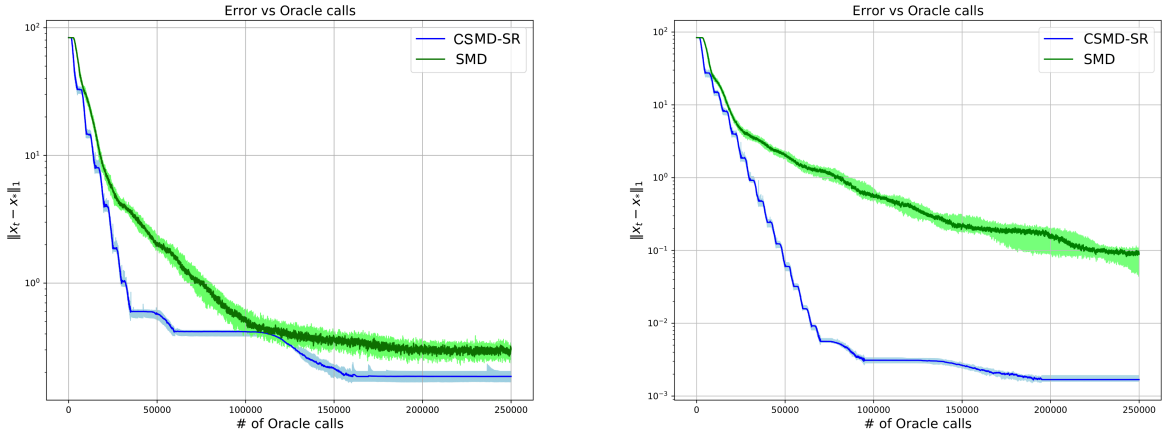
$$\exists\psi > 0:\ 2\sigma^{(s)}(z) \leq (1-\psi)\|z\| \ \ \forall z \in \mathrm{Ker}(\Sigma).$$

[7] E.g., choose P_ℓ and P_r as left and right projectors on the space generated by s principal left and right singular vectors of x_*, so that $\|x_* - P_\ell x_* P_r\| = \|(I-P_\ell)x_*(I-P_r)\| = \sum_{i=s+1}^{q}\sigma_i \leq \Delta$.

[8] Indeed, let $P \in \mathcal{P}_s$, so that $\mathrm{rank}(P_\ell) \leq s$ and $\mathrm{rank}(P_r) \leq s$, and $\|P(z)\| = \|P_\ell z P_r\| \leq \sigma^{(s)}(z)$. Since the matrix $\overline{P}(z)$ differs from z by a matrix of rank at most $2s$, by the Singular Value Interlacing theorem we have $\sigma_i(\overline{P}(z)) \geq \sigma_{i+2s}(z)$, whence $\|\overline{P}(z)\| \geq \|z\| - \sigma^{(2s)}(z)$.

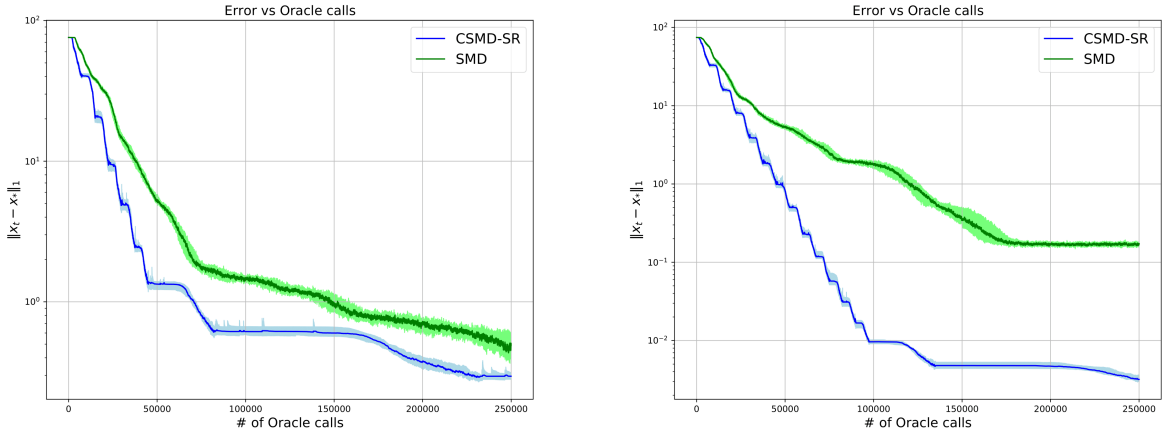
2.4.6 Supplementary numerical experiments

This section complements the numerical results appearing on the main body of the paper. We consider the setting in Section 2.3.3 of sparse recovery problem from GLR model observations (2.3.1). In the experiments below, we consider the choice (2.3.5) of activation function $\mathfrak{r}_\alpha(t)$ with values $\alpha = 1$ and $\alpha = 1/10$; value $\alpha = 1$ corresponds to linear regression with $\mathfrak{r}(t) = t$, whereas when $\alpha = 0.1$ activation have a flatter curve with rapidly decreasing with r modulus of strong convexity for $|t| \le r$. Same as before, in our experiments, the dimension of the parameter space is $n = 500\,000$, the sparsity level of the optimal point x_* is $s = 100$; we use the minibatch Algorithm 2 with the maximal number of oracle calls is $N = 250\,000$. In Figures 2.4 and 2.5 we report results for $\kappa_\Sigma \in \{0.1, 1\}$ and $\sigma \in \{0.001, 0.1\}$; the simulations are repeated 10 times, we trace the median of the estimation error $\|\widehat{x}_i - x_*\|_1$ along with its first and the last deciles against the number of oracle calls.



(a) $\kappa_\Sigma = 1, \sigma = 0.1, m_0 = 5000$

(b) $\kappa_\Sigma = 1, \sigma = 0.001, m_0 = 5000$

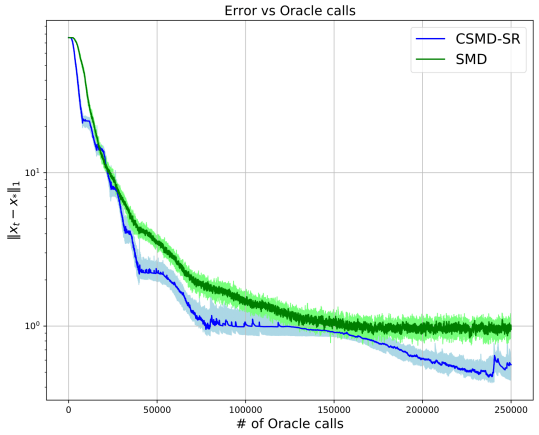


(c) $\kappa_\Sigma = 0.1, \sigma = 0.1, m_0 = 7500$

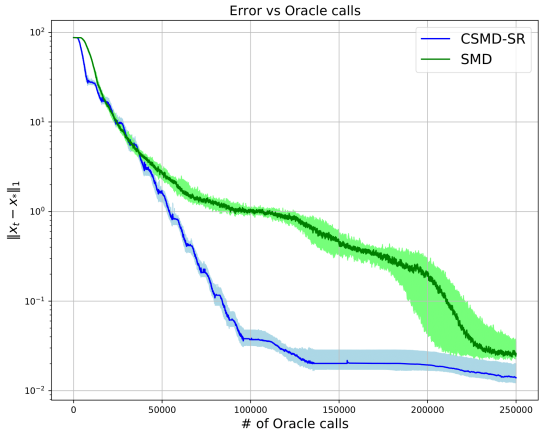
(d) $\kappa_\Sigma = 0.1, \sigma = 0.001, m_0 = 7500$

Figure 2.4: CSMD-SR and "vanilla" SMD in Linear Regression problem (activation function $\mathfrak{r}(t) = t$); ℓ_1 error as a function of the number of oracle calls

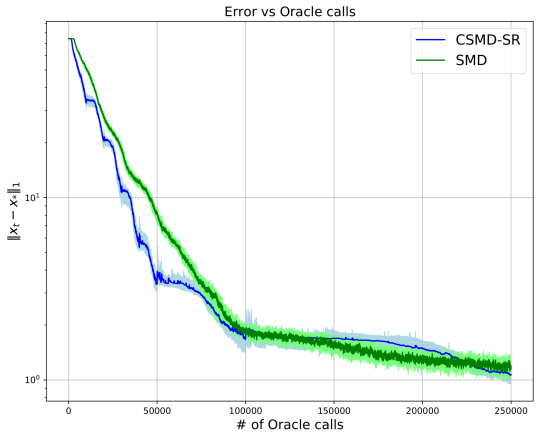
In our experiments, multistage algorithms exhibit linear convergence on initial iterations. Surprisingly, "standard" (non-Euclidean) SMD also converges fast in the "preliminary" regime. This may be explained by the fact that iteration x_i of the SMD obtained by the "usual" proximal mapping $\mathrm{Prox}(\gamma_{i-1}\nabla G(x_{i-1},\omega_i),x_{i-1})$ is computed as a solution to the optimization problem with "penalty" $\theta(x)=c\|x\|_p^p$, $p=1+1/\ln n$ which results in a "natural" sparsification of x_i. As iterations progress, such "sparsification" becomes insufficient, and the multistage routine eventually outperforms the SMD. Implementing the method for "flatter" nonlinear activation $\mathfrak{r}(t)$ or increased condition number of the regressor covariance matrix Σ requires increasing the length m_0 of the stage of the algorithm.



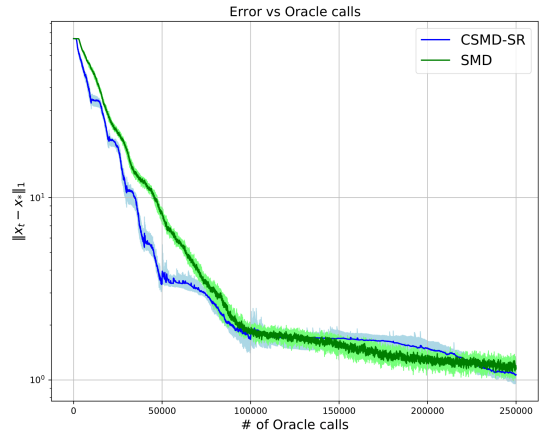
(a) $\kappa_\Sigma = 1, \sigma = 0.1, m_0 = 8000$



(b) $\kappa_\Sigma = 1, \sigma = 0.001, m_0 = 8000$



(c) $\kappa_\Sigma = 0.1, \sigma = 0.1, m_0 = 10\,000$



(d) $\kappa_\Sigma = 0.1, \sigma = 0.001, , m_0 = 10\,000$

Figure 2.5: CSMD-SR and "vanilla" SMD in Generalized Linear Regression problem: activation function $\mathfrak{r}_{1/10}(t)$; ℓ_1 error as a function of the number of oracle calls

Chapter 3

Extensions

3.1 Adaptive CSMD-SR via Lepski's Procedure

We present in this section an algorithm inspired by the CSMD-SR with hyper-parameters independent of problem parameters ρ and s, and thus adaptive to the latter. We call this new algorithm Ada-CSMD-SR.
More precisely, we are given N samples, a desired precision level $\epsilon \in (0,1)$ and a starting point x_0 with an initial prior R such that $\|x_0 - x_*\| \leq R$. We also assume that parameters $\sigma_*, \nu, \delta, \Theta$ are known, and that we have access to a stochastic approximation of objective function g's gradient, as in the previous chapter. Here, objective g satisfies the (RSC) assumption with *unknown* parameter ρ, and is minimized by x_* that is a sparse vector with *unknown* level of sparsity s. With this setting in place, we aim to produce an estimate $\widehat{x}^{(a)}$ with guarantees on the $1-\epsilon$ quantile of

$$\|\widehat{x}^{(a)} - x_*\|$$

that are almost the same as those we provide for the CSMD-SR estimate computable when knowing ρ and s.
In deterministic optimization, [122] proposes a first order algorithm adaptive to the smoothness of the objective, whereas in [48], the proposed multistage method is respectively adaptive to the uniform convexity-parameter. Authors in [48] also provide a stochastic variant of their algorithm, whereas [123] proposes a version of SGD adaptive to local strong convexity of the objective.
For the rest of this section, we will say that the pair (g, x_*) satisfies the (RSC) assumption with parameter ρs if (2.2.11) holds with parameters ρ and s.

3.1.1 Motivation

Observe that in our previous developments, one of the crucial hyper-parameter choice is the constant stage-length of the preliminary phase. In the situation where one knows $\nu, \Theta, \delta, \rho, s$, we advocate the choice

$$m = \lceil 64\delta^2 \nu \rho s(4\Theta + 60t) \rceil, \quad t \geq 4\sqrt{2 + \ln(m)}.$$

Observe that this hyper-parameter can replace the term ρs in other ones. Indeed, at the kth stage of the preliminary phase, recall that

$$\kappa_k = R_{k-1}\sqrt{\frac{\nu(4\Theta + 60t)}{\rho s m}} \asymp R_{k-1}\frac{\delta\nu(\Theta + t)}{m},$$

$$R_k = \tfrac{1}{2}R_{k-1} + \frac{16\sigma_*^2\delta^2\rho s}{\nu R_{k-1}} \asymp \tfrac{1}{2}R_{k-1} + \left(\frac{\sigma_*}{\nu}\right)^2 \frac{m}{(\Theta+t)R_{k-1}}.$$

This illustrates that our algorithm's hyper-parameters can be formulated based on the stage-length m, rather than the parameters ρ and s. Hence, for any objective g and its minimizer x_*, an estimate that adapts to the former is also adaptive to both ρ and s. This leads us to focus on generating estimates adaptive to the stage-length m. It is worth mentioning that previous work in [124] introduces a deterministic multistage method that adjusts the length of each stage without knowing the sharpness parameters of the objective function. We begin by briefly analyzing a version of the CSMD-SR algorithm where hyper-parameters are modified to depend on m.

The stage-length dependent CSMD estimate.

Let integer $m \in [\lceil 4\delta \rceil, N]$, and define $\beta(m) := \frac{m}{(64\delta)^2\nu t}$, with

$$t := \max\left\{\Theta;\ 4\sqrt{2+\ln(m)};\ t_\epsilon(m)\right\}, \tag{3.1.1}$$
$$t_\epsilon(m) := \ln\left(\frac{4}{\epsilon}\log_2\left(\left(\frac{2\nu R}{\sigma_*\sqrt{m}} \bigvee 1\right)\sqrt{1+\frac{8N}{3\Theta m}}\right)\right).$$

Observe that if (g, x_*) satisfies the RSC with $\rho s \le \beta(m)$, then

$$m \ge \left\lceil (64\delta)^2\nu t\rho s\right\rceil \ge \left\lceil 64\delta^2\nu\rho s(4\Theta + 60t)\right\rceil.$$

Thus, we say that a specific stage-length m is adapted for (g, x_*) if the latter satisfies RSC with parameter $\beta(m)$. We now present the notations used to define a version of the CSMD-SR algorithm that depends on m. For $k \ge 0$, define the sequence of preliminary rates

$$R_k(m) := \frac{R_{k-1}(m)}{2} + \left(\frac{\sigma^*}{\nu}\right)^2 \frac{m}{128R_{k-1}(m)},\ \ R_0(m) = R, \tag{3.1.2}$$

and asymptotic rates

$$r_k(m) := \frac{\sigma^*}{\nu}\frac{\sqrt{m}}{2}2^{-k}. \tag{3.1.3}$$

Observe that for all $k \ge 1$, $R_k(m) \ge \frac{\sigma_*\sqrt{m}}{16\nu}$, implying in turn that $R_k(m) \le 2^{-k}R + \frac{\sigma_*\sqrt{m}}{4\nu}$. In particular,

$$R_{K_0(m)}(m) \le r_0(m),\ \text{with } K_0(m) := \left(\left\lceil \log_2\left(\frac{4\nu R}{\sigma_*\sqrt{m}}\right)\right\rceil\right)_+.$$

With this definition, the total number of stages in the preliminary phase is

$$K_P(m) := \left\lfloor \frac{N}{m}\right\rfloor \bigwedge K_0(m).$$

For the asymptotic phase, we consider the mini-batch version of the CSMD-SR algorithm with batches of size $L_l = 4^{l-1}\lceil 9\Theta/8\rceil$ at its lth stage. Therefore, in the asymptotic phase, one can compute a maximum of $K_A(m)$ stages, with the latter being defined as

$$K_A(m) := \left(\left\lfloor \frac{1}{2}\log_2\left(1+\frac{3}{\lceil 9\Theta/8\rceil}\left(\left\lfloor\frac{N}{m}\right\rfloor - K_P(m)\right)\right)\right\rfloor\right)_+ \tag{3.1.4}$$
$$:= \max\left\{k:\ k\ge 0,\ \sum_{l=0}^{k-1}\lceil 9\Theta/8\rceil 4^l \le \left(\left\lfloor\frac{N}{m}\right\rfloor - K_P(m)\right)\right\}.$$

Given these definitions, we define the total number of stages that can be completed with N samples as $K(m) := K_P(m) + K_A(m)$.
We now introduce the adaptive version of the CSMD-SR algorithm, beginning with the initial starting point $x_0 = \widehat{x}_0(m)$. For $k \in [1 : K(m)]$, if the algorithm starts

- the preliminary phase, i.e., $1 \leq k \leq K_0(m)$, we define the kth stage's output be

$$\widehat{x}_k(m) = \text{CSMD}\left(\widehat{x}_{k-1}(m), \frac{1}{4\nu}, \kappa_k, R_{k-1}(m), m, 1\right), \tag{3.1.5}$$

  with $\kappa_k := 512\delta\nu t\frac{R_{k-1}(m)}{m}$.

- the asymptotic phase, i.e. $k > K_0(m)$, we define the lth stage's output

$$\widehat{x}_k(m) = \text{CSMD}\left(\widehat{x}_{k-1}(m), \frac{1}{4\nu}, \kappa_k, r_{l-1}(m), m, L_l\right), \tag{3.1.6}$$

  with $l = k - K_0(m)$ and $\kappa_k := 512\delta\nu t\frac{r_{l-1}(m)}{m}$.

Finally, we will denote the output of the final stage

$$\widehat{x}(m) := \widehat{x}_{K(m)}(m) \tag{3.1.7}$$

and refer to it as the CSMD-SR estimate associated with stage-length m. We also use the notation

$$R(m) := R_{K_P(m)}(m)\mathbf{1}_{\{K_A(m)=0\}} + r_{K_A(m)}(m)\mathbf{1}_{\{K_A(m)>0\}}, \tag{3.1.8}$$

for the high-probability upper bound on its error, as shown in the following result.

Proposition 3.1.1 *Let $m \in [\lceil 4\delta \rceil : N]$. Provided that (g, x_*) is such that $\rho s \leq \beta(m)$, the inequality*

$$\|\widehat{x}(m) - x_*\| \leq R(m) \tag{3.1.9}$$

holds with probability greater than $1 - \epsilon$.

Proof of Proposition 3.1.1:
We use the same arguments as for the proof of 2.4.4 and 2.4.5, adapted to the outlined choices of hyperparameters.
1°: Assume that we are at the $k+1$-th stage of the preliminary phase, starting with point $\widehat{x}_k(m)$ such that with probability greater than $1 - 4ke^{-t}$,

$$\|\widehat{x}_k(m) - x_*\| \leq R_k(m).$$

Inserting the values $\gamma = (4\nu)^{-1}$, $\kappa \equiv \kappa_{k+1}(m)$ into the equation (2.4.2) results in the upper bound

$$F_\kappa(\widehat{x}_{k+1}(m)) - F_\kappa(x_*) \leq \frac{64\nu t\left(R_k(m)\right)^2}{m}\left(1 + \frac{8\delta}{m}\right) + \frac{\sigma_*^2}{\nu}\left(\frac{7}{4} + \frac{6t}{m}\right).$$

which occurs on an event of probability greater than $1 - 4(k+1)e^{-t}$. Combining the latter with the RSC assumption being satisfied with $\rho s \leq \beta(m)$ yields

$$\|\widehat{x}_{k+1}(m) - x_*\| \leq R_k(m)\left(\frac{1}{4} + \frac{\delta}{m}\right) + \left(\frac{\sigma_*}{\nu}\right)^2\left(\frac{7}{4} + \frac{6t}{m}\right)\frac{m}{64t}\frac{1}{R_k(m)}.$$

Since $m \geq 4\delta \geq 4$ and (3.1.14) implies $t \geq 1$, one has that on the same event,

$$\|\widehat{x}_{k+1}(m) - x_*\| \leq \frac{R_{k+1}(m)}{2} + \Big(\frac{\sigma_*}{\nu}\Big)^2 \frac{m}{128 R_k(m)} = R_{k+1}(m).$$

From the results we have just established, it follows through induction that at the conclusion of the preliminary phase, with a probability exceeding $1 - 4K_P(m)e^{-t}$

$$\|\widehat{x}_{K_P(m)}(m) - x_*\| \leq R_{K_P(m)}(m) \leq R2^{-K_P(m)} + \frac{\sigma_*\sqrt{m}}{4\nu}, \tag{3.1.10}$$

holds true. Moreover, if $K_P(m) = K_0(m)$, the last inequality can be upper bounded by $\frac{\sigma_*\sqrt{m}}{2\nu} = r_0(m)$.

2°: Assume that $K_A(m) > 0$ and that we have already completed k stages, with $k \in [K_P(m); K(m) - 1]$. It follows that we are in the $l+1$-th stage of the asymptotic phase, with $l = k - K_0(m)$, with $\widehat{x}_k(m)$ as our starting point, such that

$$\text{Prob}\left[\|\widehat{x}_k(m) - x_*\| \leq r_l(m)\right] \geq 1 - 4ke^{-t}.$$

We use similar calculations and arguments as for the proof of 2.4.5. Recall that when using batches of size L, one can replace the sub-gaussianity parameter σ_*^2 by $18\Theta\sigma_*^2/L$. This yields that with probability greater than $1 - (k+1)e^{-t}$, one has that

$$\begin{aligned}\|\widehat{x}_{k+1}(m) - x_*\| &\leq \frac{r_l(m)}{2} + \Big(\frac{\sigma_*}{\nu}\Big)^2 \frac{m}{128 r_l(m)} \frac{18\Theta}{\lceil 9\Theta/8\rceil 4^{l-1}} \\ &\leq r_l(m)\left(\frac{1}{2} + \frac{18\Theta}{36\Theta}\right) = r_{l+1}(m).\end{aligned} \tag{3.1.11}$$

3°: Setting $k = K(m)$ leads to

$$\text{Prob}\left[\|\widehat{x}_{K(m)}(m) - x_*\| \leq R(m)\right] \geq 1 - 4K(m)e^{-t}.$$

Noting that $K_0(m) \leq \log_2(\frac{2\nu R}{\sigma_*\sqrt{m}} \bigvee 1)$ and $K_A(m) \leq \log_2\left(\sqrt{1 + \frac{8N}{3\Theta m}}\right)$, we can derive

$$\begin{aligned}4K(m)e^{-t} &\leq 4e^{-t}(K_0(m) + K_A(m)) \\ &\leq 4e^{-t}\log_2\left(\left(\frac{2\nu R}{\sigma_*\sqrt{m}} \bigvee 1\right)\sqrt{1 + \frac{8N}{3\Theta m}}\right) \leq \epsilon,\end{aligned} \tag{3.1.12}$$

where the final inequality is justified by definition (3.1.1). □

3.1.2 The adaptive CSMD-SR estimate

The adaptive estimate we propose is based on Lepski's [91] adaptive procedure. In our setting, the latter is applied to a collection of estimates $(\widehat{x}^{(i)})_{i=1}^I$ to select the best estimate in the context of (g, x_*). More formally, for an integer I, we assume that we are given the following grid of stage-lengths

$$\lceil 4\delta \rceil = m_1 < ... < m_I \leq N. \tag{3.1.13}$$

For all i, we define estimates $\widehat{x}^{(i)} := \widehat{x}(m_i)$, generated by the CSMD-SR algorithm presented above, with parameter

$$t_i := \max\left\{\Theta;\ 4\sqrt{2+\ln(m_i)};\ t_\epsilon(m_i)\right\} + \ln(I) \tag{3.1.14}$$

instead of t. We also define the associated quantities $\beta_i := \frac{m_i}{(64\delta)^2\nu t_i}$, $K^{(i)} := K(m_i)$, and $R^{(i)} := R(m_i)$.

Proposition 3.1.2 *Let collection of estimates $(\widehat{x}^{(i)})_{i=1}^I$ be as previously stated. The event "For all $i \in [1:I]$ such that (g, x_*) satisfies the RSC assumption with $\rho s \leq \beta_i$,*

$$\|\widehat{x}^{(i)} - x_*\| \leq R^{(i)}." \tag{3.1.15}$$

holds with probability greater than $1-\epsilon$.

Proof of Proposition 3.1.2:

1°: Using similar arguments as for the proof of (3.1.12), one has that

$$\begin{aligned} 4\sum_{i=1}^I K^{(i)} e^{-t_i} &\leq 4\sum_{i=1}^I \log_2\left(\left(\frac{2\nu R}{\sigma_*\sqrt{m_i}} \bigvee 1\right)\sqrt{1+\frac{8N}{3\Theta m_i}}\right) e^{-t_i} \\ &\leq \sum_{i=1}^I \frac{\epsilon}{I} = \epsilon. \end{aligned} \tag{3.1.16}$$

2°: We are now ready to prove (3.1.15). Let us call $\mathcal{E}_i$ the event
"If the RSC property is satisfied with parameters smaller than β_i, one has $\|\widehat{x}^{(i)} - x_*\| \leq R^{(i)}$."
Proposition 3.1.1 states that $\text{Prob}\left[\mathcal{E}_i\right] \geq 1 - 4K^{(i)}e^{-t_i}$. As the event we are interested in is $\mathcal{E} = \bigcap_{i=1}^I \mathcal{E}_i$, the result follows directly from the fact that

$$\text{Prob}\left[\mathcal{E}\right] \geq 1 - \text{Prob}\left[\bigcup_{i=1}^I \overline{\mathcal{E}_i}\right] \geq 1 - 4\sum_{i=1}^I K^{(i)} e^{-t_i} \geq 1-\epsilon,$$

where we have used (3.1.16) to prove the last inequality. □

Given the grid $(m_i)_{i=1}^I$, we now propose the following construction of our adaptive estimate

- For all $i \in [1:I]$, compute $\widehat{x}^{(i)}$, and set $\widehat{x}^{(I+1)} = x_0$, $R^{(I+1)} = R$.
- Define the set of admissible indexes

$$\mathcal{A} := \left\{i \in [1:I]:\ \forall j,\ i < j \leq 1+I,\ \|\widehat{x}^{(i)} - \widehat{x}^{(j)}\| \leq R^{(i)} + R^{(j)}\right\}, \tag{3.1.17}$$

  and let $\widehat{i} := \min \mathcal{A}$.
- Select $\widehat{x}^{(a)} := \widehat{x}^{(\widehat{i})}$ if $\mathcal{A}$ is non empty, and x_0 otherwise.

Prior to establishing guarantees on the error quantile of its estimation, we define the error of a CSMD-SR estimate that knows the value of ρs. With notation

$$m(t) := \lceil (64\delta)^2 \rho s \nu t \rceil, \tag{3.1.18}$$

we define parameters $t_* := \max\left\{\Theta; \overline{t}_*\right\}$ and $t_*^I := t_* + \ln(I)$, where

$$\overline{t}_* := \min\left\{t : t \geq \max\left\{4\sqrt{2+\ln(m(t))}; t_\epsilon(m(t))\right\}\right\}. \tag{3.1.19}$$

The next proposition requires that $R, \epsilon, \nu, \sigma_*, N, \delta$ are such that they verify the following assumptions.

Assumption 1:
For all m such that $\lceil 4\delta \rceil \leq m \leq N$, one has $R(m) < R(m+1) \leq 2R(m)$.

Assumption 2:
One can compute I and integers $(m_i)_{i=1}^{I+1}$ such that for all $i \in [1:I]$, $\lceil 4\delta \rceil \leq m_i < m_{i+1} \leq N$ and

$$2R^{(i)} \leq R^{(i+1)} \leq 4R^{(i)}, \tag{3.1.20}$$

and in particular

$$2R^{(I)} + R^{(I-1)} \leq R \leq 2R^{(I+1)} + R^{(I)}. \tag{3.1.21}$$

Assumption 3:
ρ is large enough so that $m(t_*) \geq \lceil 4\delta \rceil$.

Proposition 3.1.3 *With probability greater than $1-\epsilon$, one has*

$$\|\widehat{x}^{(a)} - x_*\| \leq 9R\left(m(t_*^I)\right) \bigwedge 3R \tag{3.1.22}$$

$$\leq 27R\left(m_*^I\right). \tag{3.1.23}$$

Remark 3.1.1 *Observe that assumptions 1 is not overly restrictive. Indeed, it can be computationally verified in a number of operation linear in N by simply calculating all the $R(m)$ for $m \in [\lceil 4\delta \rceil, N]$. Moreover, the second inequality in (3.1.20) being true essentially depends on the ratio $\sigma_*/(\nu\sqrt{R})$. For instance, if one has*

$$\frac{\sigma_*^2 \lceil 4\delta \rceil}{\nu^2 R} \geq 1,$$

for all considered m, the resulting CSMD-SR algorithm will always be in the asymptotic phase, and one essentially needs to multiply m by 4 to perform $K_A(m) - 1$ stages, which will result in multiplying the rate by two. On the other hand, when $\sigma_*/\nu \to 0$, our algorithm always stays in the preliminary phase, and in that case, we can not ensure the upper bound on $R(m+1)/R(m) \leq 2$ as $\lfloor N/(m+1) \rfloor - \lfloor N/m \rfloor$ can be greater than 1, and $R(m) \asymp R2^{-\lfloor N/m \rfloor}$.

Remark 3.1.2 *Note that if assumption 1 holds, assumption 2's fulfillement only depends on N being large enough. Indeed, starting with $m_1 = \lceil 4\delta \rceil$, one can sequentially choose m_{i+1} from the interval $[m_i + 1 : N]$ as the largest m satisfying (3.1.20), until condition (3.1.21) is achieved.*

Remark 3.1.3 *Note that if the last assumption does not hold, our procedure still yields the upper bound*

$$\|\widehat{x}^{(a)} - x_*\| \leq R(\lceil 4\delta \rceil),$$

corresponding to the smallest stage-length ensuring division of the error by 2 in the preliminary phase. Moreover, it is guaranteed to hold when $\rho \geq \frac{1}{4096\nu}$. In the situation of Section 2.3, the latter is true when

$$\kappa_\Sigma \leq \frac{4096\overline{r}}{\underline{r}}\overline{\nu}^2,$$

where κ_Σ is such that $\mathbf{E}\left[\phi\phi^T\right] \succeq \kappa_\Sigma I$.

Proof of Proposition 3.1.3:

1°: We first treat the case where it exists $\bar{i} \in [1:I]$ such that $\beta_{\bar{i}-1} < \rho s \leq \beta_{\bar{i}}$.
Using proposition 3.1.2, we have that on an event $\mathcal{E}$ of prability greater than $1-\epsilon$, the following is true:
For all $i \geq \bar{i}$ and $j > i$, one has that

$$\|\widehat{x}^{(i)} - \widehat{x}^{(j)}\| \leq \|\widehat{x}^{(i)} - x_*\| + \|\widehat{x}^{(j)} - x_*\| \leq R^{(i)} + R^{(j)}.$$

In particluar, this proves that $\bar{i} \in \mathcal{A}$, implying that on event $\mathcal{E}$, $\widehat{i} \leq \bar{i}$. We first treat the case where $\widehat{i} < \bar{i}$, where one has

$$\begin{aligned}
\|\widehat{x}^{(\widehat{i})} - x_*\| &\leq \|\widehat{x}^{(\widehat{i})} - \widehat{x}^{(\bar{i})}\| + \|\widehat{x}^{(\bar{i})} - x_*\| \\
&\leq 2R^{(\bar{i})} + R^{(\widehat{i})} \\
&\leq 2R^{(\bar{i})}(N) + \max_{l<\bar{i}} R^{(l)}(N) \qquad (3.1.24) \\
&\leq 2R^{(\bar{i})}(N) + R^{(\bar{i}-1)}(N), \qquad (3.1.25)
\end{aligned}$$

where last inequality is a direct consequence of sequence $\left(R^{(i)}\right)_{i=1}^{I}$ being increasing. Observe that in the case where $\widehat{i} = \bar{i}$, inequality (3.1.25) is still valid. By definition, one has that $m_{\bar{i}-1} < \rho s(64\delta)^2 \nu t_I \leq m_{\bar{i}}$. Using assumption 1, one also has

$$R^{(\bar{i}-1)} < R(m(t_*^I)) \leq R^{(\bar{i})},$$

which yields the desired result when combined with first part of assumption 2.

2°: If $\rho s \geq \beta_I$, then either x_0 is selected, or an index $j \in [1:I]$ is. If index j is selected, then

$$\|\widehat{x}^{(a)} - x_*\| \leq \|\widehat{x}^{(a)} - x_0\| + \|x_0 - x_*\| \leq R^{(j)} + 2R \leq 3R.$$

Observing that one has $2R^{(I+1)} + R^{(I)} \leq 9R^{(I)}$ implies that $R^{(I)} \geq R/9$, and that $\beta_I < \rho s$, one has

$$3R \leq 27R^{(I)} \leq 27R(m(t_*^I)),$$

which yields (3.1.23). □

Under the same assumptions, we state the main result of this section, an upper bound on the precision of our adaptive estimate.

Theorem 3.1.1 *One has with probability greater than $1-\epsilon$ that*

$$\|\widehat{x}^{(a)} - x_*\| \lesssim R \exp\left\{-\frac{cN}{\delta^2 \rho s \nu (t_* + \ln(I))}\right\} + \sigma_* \delta^2 \rho s \sqrt{\frac{\Theta(t_* + \ln(I))}{\nu N}}. \qquad (3.1.26)$$

Moreover, under assumption 2, the grid-size I is such that

$$I \leq \log_2\left(\frac{R}{R(\lceil 4\delta \rceil)}\right). \qquad (3.1.27)$$

Proof of Theorem 3.1.1: Using the same arguments as for the proof of 2.2.1, one has that (3.1.23) implies that with probability greater than $1-\epsilon$,

$$\|\widehat{x}^{(a)} - x_*\| \lesssim R \exp\left\{-\frac{cN}{\delta^2 \rho s \nu t_*^I}\right\} + \sigma_* \delta^2 \rho s \sqrt{\frac{\Theta t_*^I}{\nu N}}.$$

Moreover, observe that assumptions (3.1.21) and (3.1.20) implies that

$$R \geq 5R^{I-1} \geq \frac{5R_1}{4} 2^I \geq 2^I R_1,$$

which in turn implies (3.1.27) when noticing that $R_1 = R(\lceil 4\delta \rceil)$.

3.2 Analysis under Reduced Uniform Convexity hypothesis (RUC)

We present in this section an extension for the analysis of the *CSMD-SR* algorithm, introduced in the last chapter, when applying it to solve sparse recovery problem of the form (2.1.2). This section is justified by the introduction of a new condition on the regularity of the objective function g, that is linked to the notion of uniform convexity [48, 124–127]. We replace the quadratic lower bound assumption verified by the objective function g by a higher order polynomial lower bound. As discussed in the previous chapter, we will see that the $\mathbf{Q}(\lambda, \psi)$ condition and the new lower bound assumption can be integrated to establish a new assumption, similar to our approach with the RSC assumption. Before introducing this general assumption, we present for the sake of clarity a definition of uniform convexity for a differentiable function h.

Definition 3.2.1 *Let X be a convex closed subset of an Euclidean space E. A differentiable function $h : X \to \mathbf{R}$ is said to be (μ, p)-uniformly convex on X if there exists $p \geq 2$ and $\mu > 0$ such that $\forall x, y \in X$,*

$$h(x) - h(y) - \langle \nabla h(y), x - y \rangle \geq \frac{\mu}{p} \|x - y\|^p. \tag{3.2.1}$$

Note that this definition simply reduces to μ-strong convexity when h is $(\mu, 2)$-uniformly convex.
Recall that in the previous chapter, the Reduced Strong Convexity (RSC) assumption was introduced to offer a comprehensive framework for analyzing sparse problems across various setups of sparsity structures. The RSC assumption takes it origins from Lemma 2.4.7 where the function $\mu(x) = \frac{\mu}{2}x^2$ is used to provide the quadratic lower bound on the suboptimality. This section is devoted to give the analysis of the CSMD-SR algorithm when the objective function g satisfies the minoration condition presented in Lemma 2.4.7 using $\mu(x) = \frac{\mu}{p}x^p$ with $p > 2$. This leads us to introduce the Reduced Uniform Convexity assumption (RUC).

Assumption [RUC] For a general norm $\|\cdot\|$ and two constants $p > 2$ and $q \in [1, 2)$ such that $1/p + 1/q = 1$, there exist $\delta \geq 1$, and problem dependent positive constants ν, ρ, α such that as long as feasible solution $\widehat{x} \in X$ to the composite problem (2.2.4) satisfies

$$\|\widehat{x} - x_*\| \leq R_{\mathrm{RUC}} := \Gamma_{p,n,\|\cdot\|} \cdot \left(\frac{\nu\alpha^p}{\rho^{p-1}}\right)^{\frac{1}{p-2}} \text{ and } F_\kappa(\widehat{x}) - F_\kappa(x_*) \leq \upsilon,$$

it holds that

$$\|\widehat{x} - x_*\| \leq \delta\kappa^{-1}\left[\rho s^{\frac{q}{2}}\kappa^q + \upsilon\right], \tag{3.2.2}$$

where $\Gamma_{p,n,\|\cdot\|}$ is a problem dependent constant.
In line with our approach in the previous chapter, where the CSMD-SR algorithm was formulated based on the RSC assumption, we will now leverage the RUC assumption in the following sections. This will allow us to explore new parameter choices aimed at adapting the multistage algorithm for this updated framework.

Remark 3.2.1 *Notice that when taking $p = 2$, we retrieve assumption RSC. The maximal radius R_{RUC} on which the assumption can hold goes to infinity, meaning that the condition holds on $\mathbf{R}^n$, while the final bound on $\|\widehat{x} - x_*\|$ remains unchanged. This is not surprising as assuming uniform convexity around the optimum tends to assuming strong convexity around the optimum when p goes to 2.*

3.2.1 An example motivating the RUC assumption

In this section we motivate the use of the RUC assumption by studying a theoretical example. We place ourselves in the "vanilla" sparsity setting that has been thoroughly studied in Section 2.4.5 of the previous chapter. Consider X a convex set such that $X \subset E = \mathbf{R}^n$, and we have $\|\cdot\| = \|\cdot\|_1$. Our objective is to accurately recover an s-sparse unconditional ground truth x_*. Recall that these assumptions are made in our analysis :

- the stochastic gradients $\nabla G(\cdot, \omega)$ are assumed to be ν-Lipschitz almost surely, i.e.,

$$\forall x, x' \in X, \quad \|\nabla G(x,\omega) - \nabla G(x',\omega)\|_\infty \leq \mathcal{L}(\omega)\|x - x'\|_1, \quad \mathcal{L}(\omega) \leq \nu, \ a.s.. \tag{3.2.3}$$

- for some $\Sigma \in \mathbf{S}^n_+$, g is lower bounded around x_* such that

$$\forall x \in X, \quad g(x) - g(x_*) \geq \frac{\mu}{p}\|x - x_*\|^p_\Sigma. \tag{3.2.4}$$

- There exists two positive constants λ, ψ, such that condition $\mathbf{Q}(\lambda, \psi)$ holds for the positive definite matrix Σ.

Let us note initially that the first assumption inherently suggests that the objective function g exhibits ν-smoothness. When combined with the fact that the optimal point x_* is unconditional, it follows that the subsequent inequality holds true

$$\forall x \in X, \quad g(x) - g(x_*) \leq \frac{\nu}{2}\|x - x_*\|^2_1. \tag{3.2.5}$$

Additionally, recall that both of the following results are valid : $\forall x \in \mathbf{R}^n$, $\|x\|_1^2 \leq n\|x\|_2^2$ and $\|x\|_\Sigma = \sqrt{x^T\Sigma x} = \|\Sigma^{1/2}x\|_2$. These relations enable us to recast the second assumption in the following manner:

$$\forall x \in X, \quad g(x) - g(x_*) \geq \frac{\mu \varsigma_{\min}(\Sigma)^{p/2}}{p n^{p/2}}\|x - x_*\|^p_1, \tag{3.2.6}$$

where $\varsigma_{\min}(\Sigma)$ represents the smallest eigenvalue of the matrix Σ.
Simple calculations reveal that (3.2.5) and (3.2.6) cannot be both true for all elements of X. The first and second assumptions are, however, valid whenever $x \in X$ statisfies the following condition

$$\forall x \in X, \quad \|x - x_*\|_1 \leq \left(\frac{p\nu}{2\mu}\left(\frac{n}{\varsigma_{\min}(\Sigma)}\right)^{p/2}\right)^{\frac{1}{p-2}} =: \overline{R}. \tag{3.2.7}$$

This indicates that the compatibility of the first and second assumptions is confined to a particular region within X of radius $\overline{R}$.
Almost analogously, we can draw few consequences from the last assumption leading to assumption RUC being true. The first consequence being that s-sparsity of x_* and condition $\mathbf{Q}(\lambda, \psi)$ being true for some $\Sigma \succcurlyeq 0$ ensures that for any point $\widehat{x} \in X$ the following inequality holds true:

$$\|\widehat{x}\|_1 - \|x_*\|_1 \geq \psi\|\widehat{x} - x_*\|_1 - \sqrt{\frac{s}{\lambda}}\|\widehat{x} - x_*\|_\Sigma.$$

The precedent inequality is obtained by a direct application of the inequality (2.4.25) presented in the last chapter and adapted to our "vanilla" sparsity setting. The second one is that lower bound (3.2.4) being true implies that as soon as $\widehat{x}$ is such that $F_\kappa(\widehat{x}) - F_\kappa(x_*) \leq \upsilon$, one has

$$\begin{aligned}\upsilon &\geq \frac{\mu}{2p}\|\widehat{x} - x_*\|_\Sigma^p + \kappa\left(\psi\|\widehat{x} - x_*\|_1 - \sqrt{\frac{s}{\lambda}}\|\widehat{x} - x_*\|_\Sigma\right)\\ &\geq -\frac{1}{2}\max_{t\geq 0}\left\{2\kappa\sqrt{\frac{s}{\lambda}}t - \frac{\mu}{p}t^p\right\} + \kappa\psi\|\widehat{x} - x_*\|_1\\ &= -\frac{(2\kappa\sqrt{s/\lambda})^q}{2q\mu^{q-1}} + \kappa\psi\|\widehat{x} - x_*\|_1,\end{aligned}$$

since the Fenchel-Legendre transform of function $t \mapsto \frac{\mu}{p}t^p$ is $t \mapsto \frac{1}{q\mu^{q-1}}t^q$ where q is such that $p^{-1} + q^{-1} = 1$. The last inequality can then be rewritten as

$$\|\widehat{x} - x_*\|_1 \leq \frac{1}{\psi}\left[\frac{\upsilon}{\kappa} + \frac{1}{q}\Big(\frac{2\kappa}{\mu}\Big)^{q-1}\Big(\frac{s}{\lambda}\Big)^{\frac{q}{2}}\right],$$

as long as $\widehat{x}$ is at a $\|\cdot\|_1$-radius of at most

$$\overline{R} = \left(\frac{p\nu}{2\mu\lambda^{p/2}}\left(\frac{n\sqrt{\lambda}}{\sqrt{\varsigma_{\min}(\Sigma)}}\right)^p\right)^{\frac{1}{p-2}}.$$

Remark 3.2.2 *In the proposed motivating example, assumption RUC holds for* $\|\cdot\| = \|\cdot\|_1$ *and* $\delta = \frac{1}{\psi}$, $\rho = \frac{1}{q}\big(\frac{2}{\mu\lambda^{p/2}}\big)^{q-1}$, $\alpha = \sqrt{\lambda/\varsigma_{min}(\Sigma)}$, *and* $\Gamma_{p,n,\|\cdot\|} = pn^{p/(p-2)}/4$.

3.2.2 Prescribed choice of parameter and convergence results

In this section, we present the analysis of the CSMD-SR algorithm, focusing on its adaptation to the new setting where the RUC Assumption is valid. Similar to the analysis provided in the last chapter, the CSMD-SR algorithm is characterized by two distinct phases: a Preliminary phase and an Asymptotic phase. Accordingly, we will outline a convergence analysis and will prescribe a choice of parameter that is associated to each phase. For any stage $k \geq 1$, we introduce the notations $\kappa_k^{(P)}$ and $m_k^{(P)}$ to denote the penalization and length of the k-th stage in the CSMD-SR algorithm's Preliminary phase, respectively. Similarly, $\kappa_k^{(A)}$ and $m_k^{(A)}$ refer to the penalization and length of the k-th stage in its Asymptotic phase.
Next lemma provides the theoretical guarantees for the multistage method when applied in solving the sparse recovery problem.

Lemma 3.2.1 *Assume that problem's parameters and the algorithm's initialization point* x_0 *are such that* $R := \|x_0 - x^*\|$ *verifies*

$$r_0 \leq R \leq R_{RUC} \wedge R^{(P)}, \tag{3.2.8}$$

where

$$r_0 := C_1(p)\delta\sqrt{s}\rho^{\frac{p-1}{p}}\nu^{-1/p}\sigma_*^{2/p} \quad \textit{and} \quad R^{(P)} := \frac{1}{4(p-1)^{\frac{p-1}{p-2}}}(\rho^{p-1}\nu)^{\frac{1}{p-2}}.$$

This ensures that both the RUC Assumption and some other condition that will be discussed in the proof of the Lemma are verified.
Consider the size of the Preliminary phase defined as $\overline{K}_1 := \left\lceil \log_2\left(\frac{2R}{r_0}\right)\right\rceil$. *Now for* $k \in \{1,\dots,\overline{K}_1\}$ *and* $t \geq 4\sqrt{2+\ln\left(m_{\overline{K}_1}^{(P)}\right)}$ *we define the value of the penalization parameter chosen at the k-th stage and the length of the k-th stage of the Preliminary phase such as*

$$\kappa_k^{(P)} := \frac{1}{\sqrt{s}}\left(\frac{R_{k-1}^2\nu\,(4\Theta+60t)}{m_k^{(P)}\rho(q-1)}\right)^{\frac{p-1}{p}} \quad \text{and} \quad m_k^{(P)} := \left\lceil \frac{4^p p^p}{(p-1)^{p-1}}\delta^p s^{\frac{p}{2}}\rho^{p-1}\nu\,(4\Theta+60t)\,R_{k-1}^{2-p}\right\rceil,$$

where the sequence $(R_k)_{k\geq 0}$ *verifies the following recursion*

$$R_{k+1} = \frac{1}{2}R_k + \frac{C_0(p)\sigma_*^2}{\nu}\left(\frac{\delta\sqrt{s}\rho^{\frac{p-1}{p}}}{R_k^{\frac{p-1}{p}}}\right)^p, \text{ and } R_0 = R.$$

Given this setup, for any $k \in \{1,\dots,\overline{K}_1\}$, *the approximate solution* $\widehat{x}^k_{m_k^{(P)}}$ *computed at the end of the k-th stage of the CSMD-SR algorithm satisfies with probability* $\geq 1-4ke^{-t}$

$$\|\widehat{x}^k_{m_k^{(P)}} - x_*\| \quad \leq R_k \leq 2^{-k}R + \tfrac{1}{2}C_1(p)\delta\nu^{-\frac{1}{p}}\sigma_*^{\frac{2}{p}}\sqrt{s}\rho^{\frac{p-1}{p}}. \tag{3.2.9}$$

In particular, the estimate $\widehat{x}^{\overline{K}_1}_{m_{\overline{K}_1}^{(P)}}$ *computed after* $\overline{K}_1$ *stages of the preliminary phase, satisfies with probability at least* $1-4\overline{K}_1e^{-t}$

$$\|\widehat{x}^{\overline{K}_1}_{m_{\overline{K}_1}^{(P)}} - x_*\| \leq C_1(p)\sigma_*^{\frac{2}{p}}\delta\sqrt{s}\rho^{\frac{p-1}{p}}\nu^{-\frac{1}{p}} = r_0. \tag{3.2.10}$$

The values of the 'constants' $C_0(p)$ *and* $C_1(p)$ *can be found in the proof of the lemma.*

The analysis of the Preliminary phase presented in Lemma 3.2.1 demonstrates that under the RUC Assumption, the stage lengths exhibit exponential growth with the stage count. This contrasts with the behavior appearing under the RSC Assumption, where stage lengths remain constant. As a result, the linear decay observed in the Preliminary phase under the RSC Assumption does not occur in analyses based on the RUC Assumption. This is presented later in the manuscript in Theorem 3.2.1.

Now we assume that the Preliminary phase of the algorithm is terminated, in other words that we have completed $\overline{K}_1$ stages of the Preliminary phase, we transition to analysis of the Asymptotic phase. For the sake of simplicity, the analysis is provided in the mini-batch setting.

Lemma 3.2.2 *Recall that the initial radius verifies* $r_0 \leq R \leq R_{RUC} \wedge R^{(P)}$ *as defined in Lemma 3.2.1. Consider the size of the Asymptotic phase defined as*

$$\overline{K}_2 := \max\left\{ k \,\middle|\, \sum_{i=1}^{k} m_i^{(A)}\ell_i \leq N - \sum_{i=1}^{\overline{K}_1} m_i^{(P)} \right\}.$$

Now for $k \in \{1, \dots, \overline{K}_2\}$ *and* $t \geq 4\sqrt{2 + \ln\left(m_{\overline{K}_2}^{(P)}\right)}$ *we define the value of the penalization parameter, the length of the stages and the batch-size chosen at the* k*-th stage of the Asymptotic phase such as*

$$\kappa_k^{(A)} = 2^{-k(p-1)} C_2(p) \left(\frac{\sigma_*^2}{s^{\frac{p}{2(p-1)}} \rho \nu} \right)^{\frac{p-1}{p}}, \qquad m_k^{(A)} = \left\lceil 2^{k(p-2)} C_3(p) \delta^2 (\nu\rho)^{\frac{2(p-1)}{p}} s(4\Theta + 60t) \sigma_*^{\frac{2(2-p)}{p}} \right\rceil, \tag{3.2.11}$$

$$\ell_k = \left\lceil 2^{kp} C_4(p) \Theta \right\rceil.$$

We set the sequence $(r_k)_{k \geq 0}$ *such that it verifies the following recursion*

$$r_k = 2^{-k} r_0, \text{ and } r_0 = C_1(p) \sigma_*^{\frac{2}{p}} \delta \sqrt{s} \rho^{\frac{p-1}{p}} \nu^{-\frac{1}{p}}.$$

Given this setting, the approximate solution produced by the CSMD-SR Algorithm denoted $\widehat{x}^k_{m_k^{(A)}}$*, satisfies at the end of the* k*-th stage of the Asymptotic phase, for* $k \in \{1, \dots, \overline{K}_2\}$*, with probability* $\geq 1 - 4(\overline{K}_1 + k)e^{-t}$*,* $\|\widehat{x}^k_{m_k^{(A)}} - x_*\| \leq r_k$*, implying that*

$$\|\widehat{x}^k_{m_k^{(A)}} - x_*\| \leq \left(\frac{C_5(p) \sigma_*^2 \delta^{2p} s^p \rho^{2(p-1)} \Theta \left(4\Theta + 60t\right)}{N_k} \right)^{\frac{1}{2(p-1)}}, \tag{3.2.12}$$

where $N_k = \sum_{i=1}^k m_i^{(A)} \ell_i$ *is the total count of oracle calls for* k *asymptotic stages. The values of the constants* $C_2(p), C_3(p), C_4(p), C_5(p)$ *are provided in the proof of the lemma.*

Similar to the result appearing in the first lemma, Lemma 3.2.2 shows that during the Asymptotic phase both the length of the stages and the minibatch size grow exponentially with the stage count.

Now we present the main result of the current analysis. In the following theorem, we present the rate of recovery achieved by the CSMD-SR algorithm under the RUC assumption when the sample size N is fixed in advance.

Theorem 3.2.1 *Assume that the total sample budget satisfies* $N \geq m_1^{(P)}$*, so that at least one stage of the Preliminary phase of the CSMD-SR Algorithm is completed, then for* $t \geq 4\sqrt{2 + \ln(m_{\overline{K}_2}^{(A)})}$*, the corresponding solution* $\widehat{x}_N^{(b)}$ *of the CSMDR-SR algorithm satisfies with probability at least* $1 - 4(\overline{K}_1 + \overline{K}_2)e^{-t}$

$$\|\widehat{x}_N^{(b)} - x_*\| \leq \left(\frac{C_6(p) \delta^p s^{\frac{p}{2}} \rho^{p-1} \nu (\Theta + t)}{N} \right)^{\frac{1}{p-2}} + \left(\frac{C_7(p) \sigma_*^2 \delta^{2p} s^p \rho^{2(p-1)} \Theta \left(\Theta + t\right)}{N} \right)^{\frac{1}{2(p-1)}}.$$

where $\overline{K}_2 := \max\left\{ k \,\middle|\, \sum_{i=1}^k m_i^{(A)} \ell_i \leq N - \sum_{i=1}^{\overline{K}_1} m_i^{(P)} \right\}$ *is the count for the number of stages of the Asymptotic phase of the algorithm. Value of* $C_6(p)$ *and* $C_7(p)$ *can be found in the proof of the theorem.*

Remark 3.2.3 *At first glance there seems to be a discrepancy between the settings where* $p = 2$ *and* $p > 2$ *since the term related to the Preliminary phase in the bound of Theorem 3.2.1 exhibits a sublinear decay whereas the same term of Theorem 2.2.1 exhibits a linear decay. This difference arises due to the majoration made in the proof of the theorem in equation (3.3.10) to obtain equation*

(3.3.11). Indeed, if we look back at (3.3.10) when p is close to 2 we have $2^{k(p-2)} - 1 \underset{p\to 2}{\sim} k(p-2)$ and similarly $2^{p-1} - 1 \underset{p\to 2}{\sim} p - 2$. Plugging everything together, we have $N \lesssim m_0 k$, which ultimately leads to the bound (2.4.23) represented in the preceding chapter.

Remark 3.2.4 *In the same way as the CSMD-SR algorithm can be made adaptive to the quantity ρs under the RSC assumption, it can also made adaptive to the quantity $\rho s^{\frac{q}{2}}$ by using Lepski's adaptation protocole under the RUC assumption. Note that the algorithm can also be made adaptive to the uniform convexity parameter p since the convergence bounds are monotone in p.*

3.3 Appendix: proofs.

3.3.1 Proof of Lemma 3.2.1

Proof.

1^o. First let us start by proving that $\forall k \geq 0$, we have $R_k \leq 2^{-k}R + r_0/2$.

Set $\alpha = C_0(p)\sigma_*^2\delta^p\sqrt{s}^p\rho^{p-1}\nu^{-1}$, and let us study the behaviour of the sequence

$$R_k = \frac{R_{k-1}}{2} + \frac{\alpha}{R_{k-1}^{p-1}} =: f(R_{k-1}) \quad \forall k \geq 1, \quad R_0 = R.$$

One can easily check that function f is convex and admits a minimum at $\overline{R} := (2(p-1)\alpha)^{\frac{1}{p}}$. For any initial radius $R_0 > 0$, we have $\forall k \geq 0,\ R_{k+1} = f(R_k) \geq f(\overline{R})$, where $f(\overline{R}) = \frac{q}{2}(2(p-1)\alpha)^{\frac{1}{p}}$, we thus have $\forall k \geq 1$, $R_k \geq \frac{q}{2}(2(p-1)\alpha)^{\frac{1}{p}}$. Then, by using the precedent result, we can upper bound $R_{k-1}^{-(p-1)}$ as follows, $R_{k-1}^{-(p-1)} \leq \left(\frac{2}{q}\right)^{p-1}(2(p-1)\alpha)^{-\frac{1}{q}}$. We have then shown that

$$\forall k \geq 1,\ R_k = f(R_{k-1}) \leq \frac{1}{2}R_{k-1} + \left(\frac{2}{q}\right)^{p-1}(2(p-1))^{-\frac{1}{q}}\alpha^{\frac{1}{p}}$$

By plugging the value of α into the last inequality and by setting

$$C_1(p)/4 := C_0(p)^{1/p}\left(\frac{2}{q}\right)^{p-1}(2(p-1))^{-\frac{1}{q}}$$

we obtain that

$$\forall k \geq 1,\ R_k \leq \frac{1}{2}R_{k-1} + \frac{r_0}{4}. \tag{3.3.1}$$

Now by invoking the recursive relationship of the sequence $(R_k)_{k\geq 0}$ and last inequality, we immediately have that

$$\forall k \geq 1,\ R_k \leq 2^{-k}R + \frac{r_0}{4}\sum_{i=0}^{k-1}2^{-i} \leq 2^{-k}R + \frac{r_0}{2}. \tag{3.3.2}$$

2^o. We provide a brief explanation of the idea of the proof. Observe that under the hypothesis validating Proposition 2.2.1, for $t \geq 4\sqrt{2+\ln(m)}$, we have with probability at least $1-4e^{-t}$, for $\widehat{x}_m$ an approximate solution obtained after having applied m-step of the CSMD algorithm, that the following inequality holds true

$$F_\kappa(\widehat{x}_m) - F_\kappa(x_*) \leq \frac{R^2}{m\gamma}(\Theta + 15t) + \frac{\kappa R}{m} + \sigma_*^2\gamma\left(7 + \frac{24t}{m}\right) := \upsilon.$$

Recall that with the choice $\gamma = (4\nu)^{-1}$, the quantity υ becomes

$$\upsilon = \frac{R^2\nu}{m}(4\Theta + 60t) + \frac{\kappa R}{m} + \frac{\sigma_*^2}{\nu}\left(\frac{7}{4} + \frac{6t}{m}\right).$$

We can now use the *Reduced Uniform Convexity* assumption since the radius R is such that $R \leq R_{\mathrm{RUC}}$. Therefore the above value of υ together with result (3.2.2) results in

$$\|\widehat{x}_m - x_*\| \leq \delta\left[\kappa^{q-1}s^{\frac{q}{2}}\rho + \frac{R^2\nu}{\kappa m}(4\Theta + 60t) + \frac{R}{m} + \frac{\sigma_*^2}{\kappa\nu}\left(\frac{7}{4} + \frac{6t}{m}\right)\right]. \tag{3.3.3}$$

The rest of the proof is carried out by induction. It consists in applying result (3.3.3) for each stage of the algorithm and choosing its parameters accordingly to the statement of Lemma 3.2.1.
First note that initial point x_0 satisfies $x_0 \in X_R(x_*)$ with probability 1 by definition. Now suppose that the initial point $x_0^k = \widehat{x}_{m_{k-1}}^{k-1}$ of the kth stage of the method satisfy $x_0^k \in X_{R_{k-1}}(x_*)$ with probability $1-4(k-1)e^{-t}$. In other words, there is a set $\mathcal{B}_{k-1} \subset \Omega$, with $\mathrm{Prob}(\mathcal{B}_{k-1}) \geq 1-4(k-1)e^{-t}$, such that for all $\overline{\omega}^{k-1} = [\omega_1; ...; \omega_{m_{k-1}^{(P)}}] \subset \mathcal{B}_{k-1}$ one has $x_0^k \in X_{R_{k-1}}(x_*)$. Let us show that upon termination of the kth stage $\widehat{x}^k_{m_k^{(P)}}$ satisfy $\|\widehat{x}^k_{m_k^{(P)}} - x_*\| \leq R_k$ with probability $1-4ke^{-t}$. By result (3.3.3) we conclude that for some $\overline{\Omega}_k \subset \Omega$, $\mathrm{Prob}(\overline{\Omega}_k) \geq 1-4e^{-t}$, solution $\widehat{x}^k_{m_k^{(P)}}$ after $m_k^{(P)}$ iterations of the stage satisfies, for all for all $\omega^k = [\omega_{m_{k-1}^{(P)}+1}, ..., \omega_{m_k^{(P)}}] \in \overline{\Omega}_k$,

$$\|\widehat{x}^k_{m_k^{(P)}} - x_*\| \leq \delta\left[{\kappa_k^{(P)}}^{q-1}s^{\frac{q}{2}}\rho + \frac{R_{k-1}^2\nu}{\kappa_k^{(P)}m_k^{(P)}}(4\Theta + 60t) + \frac{R_{k-1}}{m_k^{(P)}} + \frac{\sigma_*^2}{\kappa_k^{(P)}\nu}\left(\frac{7}{4} + \frac{6t}{m_k^{(P)}}\right)\right]. \tag{3.3.4}$$

We now choose $\kappa_k^{(P)}$ in order to minimize the first two terms of equation (3.3.4), i.e.,

$$\kappa_k^{(P)} = \frac{1}{\sqrt{s}}\left(\frac{R_{k-1}^2\nu\,(4\Theta + 60t)}{m_k^{(P)}(q-1)\rho}\right)^{\frac{1}{q}}$$

Plugging the parameter value into (3.3.4) gives

$$\begin{aligned}\|\widehat{x}^k_{m_k^{(P)}} - x_*\| \leq \delta\Big[&qR_{k-1}^{\frac{2}{p}}\sqrt{s}\rho^{\frac{1}{q}}\left(\frac{\nu\,(4\Theta + 60t)}{(q-1)m_k^{(P)}}\right)^{\frac{1}{p}} + \frac{R_{k-1}}{m_k^{(P)}} \\ &+ \frac{\sigma_*^2\sqrt{s}}{\nu R_{k-1}^{\frac{2}{q}}}\left(\frac{7}{4} + \frac{6t}{m_k^{(P)}}\right)\left(\frac{(q-1)\rho m_k^{(P)}}{\nu(4\Theta + 60t)}\right)^{\frac{1}{q}}\Big].\end{aligned} \tag{3.3.5}$$

Now observe that by choosing the length of each stage such that $\forall k \in \{1, \ldots, \overline{K}_1\}$

$$m_k^{(P)} = \left\lceil\left(4p\delta\sqrt{s}\rho^{\frac{p-1}{p}}\right)^p \nu(p-1)^{1-p}\,(4\Theta + 60t)\,R_{k-1}^{2-p}\right\rceil,$$

we have

$$m_k^{(P)} \geq \left(4p\sqrt{s}\rho^{\frac{p-1}{p}}\delta\right)^p \nu(p-1)^{1-p}\left(4\Theta + 60t\right)R_{k-1}^{2-p},$$

the value is specifically chosen to bound the first term of (3.3.5) by $\frac{R_{k-1}}{4}$. We will now establish that the second term is likewise bounded by $\frac{R_{k-1}}{4}$. Let us introduce $K_1 := \log_2\left(\frac{2R}{r_0}\right)$ such that $\overline{K}_1 = \lceil K_1 \rceil$. Recall from the initial part of the proof that we proved

$$\forall k \in \{1, \dots, \overline{K}_1\},\ R_{k-1} \leq 2^{-k+1}R + \frac{r_0}{2}.$$

Coupling the latter with the choice of the initial point x_0 such that $R \geq r_0$ and

$$2^{-\overline{K}_1+1}R \geq 2^{-K_1}R = \frac{r_0}{2} \geq 2^{-\overline{K}_1}R$$

boils down to state that

$$\forall k \in \{1, \dots, \overline{K}_1\},\ R_{k-1} \leq 2^{-k+2}R. \tag{3.3.6}$$

By using (3.3.6), the assumption $R \leq R_{\mathrm{RUC}} \wedge R^{(P)}$ and $p \geq 2$, we immediately obtain that the following inequalities holds true

$$\begin{aligned} R_{k-1}^{p-2} &\leq \left(2^{(-k+2)}R^{(P)}\right)^{p-2} \\ &\leq 2^{-k(p-2)}(p-1)^{1-p}(\rho^{p-1}\nu)^{-1}. \end{aligned}$$

The last inequality gives us that

$$(4p\sqrt{s}\delta)^p\rho^{p-1}\nu(p-1)^{1-p}R_{k-1}^{2-p} \geq 2^{k(p-2)}(4p\sqrt{s}\delta)^p \geq 1,$$

since $s, \delta \geq 1$, $p \geq 2$ and $k \in \{1, \dots, \overline{K}_1\}$. We have just proved that from the latter choice of the length of each stages of the Preliminary phase $m_k^{(P)}$ we have

$$m_k^{(P)} \geq (4\Theta + 60t) > 60t > 339,$$

this implies that at the same time the following inequalities holds true

$$\frac{R_{k-1}}{m_k^{(P)}} \leq \frac{R_{k-1}}{4}, \quad \text{and} \quad \frac{6t}{m_k^{(P)}} \leq \frac{1}{4}.$$

By bringing all these results together we can show that

$$\|\widehat{x}^k_{m_k^{(P)}} - x_*\| \leq \frac{R_{k-1}}{2} + \frac{2\sigma_*^2\delta\sqrt{s}}{\nu R_{k-1}^{\frac{2}{q}}}\left(\frac{(q-1)\rho m_k^{(P)}}{\nu(4\Theta + 60t)}\right)^{\frac{1}{q}} \tag{3.3.7}$$

By using the fact that for all $x \geq 1$, $\lceil x \rceil \leq 2x$, we can bound $m_k^{(P)}$ and obtain

$$\|\widehat{x}^k_{m_k^{(P)}} - x_*\| \leq \frac{R_{k-1}}{2} + \frac{\sigma_*^2}{\nu}\frac{\delta^p\sqrt{s}^p\rho^{p-1}}{R_{k-1}^{p-1}}\underbrace{q^{p-1}2^{\frac{1}{q}+2p-1}}_{=:C_0(p)} = R_k. \tag{3.3.8}$$

We conclude that $\widehat{x}^k_{m_k^{(P)}} \in X_{R_k}(x_*)$ for all $\overline{\omega}^k \in \mathcal{B}_k = \mathcal{B}_{k-1} \cap \overline{\Omega}_k$, and by application of the union bound we obtain

$$\mathrm{Prob}(\mathcal{B}_k) \geq \mathrm{Prob}(\mathcal{B}_{k-1}) - \mathrm{Prob}(\overline{\Omega}_k^c) \geq 1 - 4ke^{-t}.$$

Proof of results (3.2.9) and (3.2.10) follows immediately by bounding the sequence $(R_k)_{k\geq 0}$ with (3.3.2) and plugging the value of $\overline{K}_1$. This concludes the proof of the lemma. □

3.3.2 Proof of Lemma 3.2.2

Proof. Note that proof of Lemma 3.2.2 is very similar to the proof of Lemma 3.2.1. It starts with the exact same arguments as invoked in the precedent proof except that now we consider the sequence $(r_k)_{k\geq 0}$ and we replace at each stage the sub-Gaussian parameter σ_*^2 by $\overline{\sigma}_*^2 = \frac{18\sigma_*^2\Theta}{\ell_k}$ since we are using mini-batches of increasing size ℓ_k (cf. Lemma 2.4.3 presented in the previous chapter). In other words, for $k \in \{1, \dots, \overline{K}_2\}$ the claim (3.3.4) becomes

$$\|\widehat{x}^k_{m_k^{(A)}} - x_*\| \leq \delta \left[{\kappa_k^{(A)}}^{q-1} s^{\frac{q}{2}} \rho + \frac{\nu r_{k-1}^2}{m_k^{(A)} \kappa_k^{(A)}} (4\Theta + 60t) + \frac{r_{k-1}}{m_k^{(A)}} + \frac{18\sigma_*^2\Theta}{\kappa_k^{(A)} \ell_k \nu} \left(\frac{7}{4} + \frac{6t}{m_k^{(A)}} \right) \right],$$

We choose the parameters $\kappa_k^{(A)}$ and $m_k^{(A)}$ using the same arguments as presented in the proof of Lemma 3.2.1, except that now, after having chosen $\kappa_k^{(A)}$ and plugged its value in the previous relationship we obtain

$$\|\widehat{x}^k_{m_k^{(A)}} - x_*\| \delta \left[q r_{k-1}^{\frac{2}{p}} \sqrt{s} \rho^{\frac{1}{q}} \left(\frac{\nu (4\Theta + 60t)}{(q-1) m_k^{(P)}} \right)^{\frac{1}{p}} + \frac{r_{k-1}}{m_k^{(P)}} + \frac{18\sigma_*^2\Theta}{\kappa_k^{(A)} \ell_k \nu} \left(\frac{7}{4} + \frac{6t}{m_k^{(P)}} \right) \right]$$

The parameter $m_k^{(A)}$ is then chosen to bound each of the appearing first two terms of the last inequality by $\frac{r_{k-1}}{8}$. The latter choice of parameter still implies that

$$\frac{r_{k-1}}{m_k^{(A)}} \leq \frac{r_{k-1}}{8}, \text{ and } \frac{6t}{m_k^{(A)}} < \frac{1}{4}$$

since $r_0 \leq R \leq R_{\mathrm{RUC}} \wedge R^{(P)}$, and it results in the following bound

$$\|\widehat{x}^k_{m_k^{(A)}} - x_*\| \leq \frac{r_{k-1}}{4} + \frac{36\delta\sigma_*^2\Theta}{\kappa_k^{(A)} \ell_k \nu}.$$

Then we choose $\ell_k = \lceil \frac{144\delta\sigma_*^2\Theta}{r_{k-1}\kappa_k^{(A)}\nu} \rceil$ in order to bound the last term by $\frac{r_{k-1}}{4}$, this finally results in

$$\forall k \in \{1, \dots, \overline{K}_2\}, \quad \|\widehat{x}^k_{m_k^{(A)}} - x_*\| \leq \frac{r_{k-1}}{4} + \frac{r_{k-1}}{4} = \frac{r_{k-1}}{2} = r_k = 2^{-k} r_0. \tag{3.3.9}$$

This immediately results in (3.2.12). Values of the constants are provided below for the interested reader:

$$C_2(p) = \left(\frac{C_1(p)}{4q} \right)^{p-1}, \quad C_3(p) = 2(p-1)(4q)^p C_1(p)^{2-p}, \quad C_4(p) = 72 \times \frac{(4q)^{p-1}}{C_1(p)^p}.$$

Now let us express the total count of oracle call after k stages of asymptotic stage.

$$\begin{aligned} N_k &= \sum_{i=1}^k m_i^{(A)} \ell_i = C_3(p) C_4(p) \Theta (4\Theta + 60t) s \delta^2 (\rho\nu)^{\frac{2}{q}} \sigma_*^{\frac{2(q-2)}{q}} \sum_{i=1}^k 2^{2i(p-1)} \\ &\leq \frac{C_3(p) C_4(p)}{4^{p-1} - 1} \Theta (4\Theta + 60t) s \delta^2 (\rho\nu)^{\frac{2}{q}} \sigma_*^{\frac{2(q-2)}{q}} 2^{2(p-1)k}. \end{aligned}$$

By inverting the last inequality, we express 2^{-k} as a function of N_k up to a constant depending only on p, and result (3.2.12) follows after plugging this value and the value of r_0 within result (3.3.9). The multiplicative term appearing in (3.3.9) and denoted $C_5(p)$ is as follows :

$$C_5(p) := \frac{C_1(p)^{2(p-1)} C_3(p) C_4(p)}{4^{p-1} - 1}.$$

□

3.3.3 Proof of Theorem 3.2.1

Proof. Let assume that the "total observation budget" N is such that only the preliminary phase of the procedure is implemented. This is the case when either $\sum_{i=1}^{\overline{K}_1} m_i^{(P)} \geq N$, or $\sum_{i=1}^{\overline{K}_1} m_i^{(P)} < N$ and $\sum_{i=1}^{\overline{K}_1} m_i^{(P)} + m_1^{(A)}\ell_1 > N$. The output $\widehat{x}_N$ of the algorithm is then the last update of the Preliminary phase, and by Lemma 3.2.1 it satisfies $\|\widehat{x}_N - x_*\| \leq R_k \leq 2^{-k+1}R$ where k is the count of completed stages. In the first case we have that

$$N \leq \sum_{i=1}^{k} m_i^{(P)} < \tfrac{2}{q-1}\left(4q\sqrt{s}\rho^{\frac{1}{q}}\delta\right)^p \nu\,(4\Theta + 60t)\sum_{i=0}^{k-1} R_i^{2-p}$$

Note that a proof by induction readily demonstrates that $\forall k \in \{1,\ldots,\overline{K}_1\}$, we have $2^{-k}R \leq R_k$. Consequently, it follows that $2^{-k(p-2)}R^{p-2} \leq R_k^{p-2}$, which yields the following result:

$$\begin{aligned} N &< \tfrac{2}{q-1}\left(4q\sqrt{s}\rho^{\frac{1}{q}}\delta\right)^p \nu\,(4\Theta + 60t)\,R^{2-p}\sum_{i=0}^{k-1} 2^{i(p-2)} \\ &= \tfrac{2}{q-1}\left(4q\sqrt{s}\rho^{\frac{1}{q}}\delta\right)^p \nu\,(4\Theta + 60t)\,R^{2-p}\frac{2^{k(p-2)}-1}{2^{p-2}-1} \qquad (3.3.10) \\ &< \tfrac{2}{q-1}\left(4q\sqrt{s}\rho^{\frac{1}{q}}\delta\right)^p \nu\,(4\Theta + 60t)\,R^{2-p}\frac{2^{k(p-2)}}{2^{p-2}-1}. \qquad (3.3.11) \end{aligned}$$

The following bound can be obtained by rearranging the last equation

$$2^{-k} \leq \frac{1}{R}\left(\tfrac{2}{q-1}\left(4q\sqrt{s}\rho^{\frac{1}{q}}\delta\right)^p \frac{\nu\,(4\Theta+60t)}{N(2^{p-2}-1)}\right)^{\frac{1}{p-2}}.$$

Finally, we have shown that with probability at least $1 - 4ke^{-t}$

$$\|\widehat{x}_N - x_*\| \leq \left(\frac{C_6(p)\delta^p s^{\frac{p}{2}}\rho^{p-1}\nu(\Theta+t)}{N}\right)^{\frac{1}{p-2}}, \qquad (3.3.12)$$

where $C_6(p) := 30(p-1)\frac{(8q)^p}{2^{p-2}-1}$.
On the other hand, when $\sum_{i=1}^{\overline{K}_1} m_i^{(P)} < N < \sum_{i=1}^{\overline{K}_1} m_i^{(P)} + m_1^{(A)}\ell_1$, by using the definition of $m_i^{(P)}, \overline{K}_1, m_1^{(A)}$ and ℓ_1, and by giving a similar reasoning as stated above, one has

$$N < 4^{p-1}C_4(p)C_3(p)\delta^2(\nu\rho)^{\frac{2}{q}}s\Theta(4\Theta+60t)\sigma_*^{\frac{2(q-2)}{q}} + \tfrac{2}{q-1}\left(4q\sqrt{s}\rho^{\frac{1}{q}}\delta\right)^p \nu\,(4\Theta+60t)\,R^{2-p}\frac{2^{\overline{K}_1(p-2)}}{2^{p-2}-1}.$$

Recall that we have

$$\overline{K}_1 = \left\lceil \log_2\left(\frac{2R\nu^{\frac{1}{p}}}{C_1(p)\delta\sigma_*^{\frac{2}{p}}\sqrt{s}\rho^{\frac{1}{q}}}\right)\right\rceil \leq \log_2\left(\frac{2R\nu^{\frac{1}{p}}}{C_1(p)\delta\sigma_*^{\frac{2}{p}}\sqrt{s}\rho^{\frac{1}{q}}}\right) + 1.$$

Plugging this into the last equation and using the fact that $2\Theta \geq 1$, we have

$$N < \Big[\underbrace{4^{p-1}C_4(p)C_3(p) + \frac{2^{4p+5}q^p}{(q-1)(2^{p-2}-1)C_1(p)^{p-2}}}_{=:\widetilde{C}_7(p)}\Big]\delta^2(\nu\rho)^{\frac{2}{q}}s\Theta(4\Theta+60t)\sigma_*^{\frac{2(q-2)}{q}}.$$

Or similarly,

$$\nu^{-\frac{1}{p}} < \left(\frac{\widetilde{C}_7(p)\delta^2 \rho^{\frac{2}{q}} s\Theta(4\Theta + 60t)\sigma_*^{\frac{2(q-2)}{q}}}{N} \right)^{\frac{1}{2(p-1)}}.$$

Note that as the Preliminary phase is terminated, bound (3.2.10) is valid with probability greater than $1 - 4\overline{K}_1 e^{-t}$, this together with the previous inequality results in the following bound

$$\|\widehat{x}_N - x_*\| \leq C_1(p)\sigma_*^{\frac{2}{p}} \delta \sqrt{s} \rho^{\frac{1}{q}} \nu^{-\frac{1}{p}} \leq \left(\frac{\overline{C}_7(p)\sigma_*^2 \delta^{2p} s^p \rho^{2(p-1)} \Theta\left(\Theta + t\right)}{N} \right)^{\frac{1}{2(p-1)}}$$

where $\overline{C}_7(p) := 60 C_1(p)^{2(p-1)} \widetilde{C}_7(p)$.

Now, consider the case where at least one asymptotic stage has been completed. When $\sum_{i=1}^{\overline{K}_1} m_i^{(P)} > \frac{N}{2}$ we still have $N \leq 2\sum_{i=1}^{\overline{K}_1} m_i^{(P)}$, so that the bound (3.3.12) holds for the approximate solution $\widehat{x}_N^{(b)}$ at the end of the Asymptotic stage with the same multiplicative constant. Otherwise, the number of oracle calls N_k of asymptotic stages satisfies $N_k \geq N/2$, and by (3.2.12) this implies that with probability $\geq 1 - 4(\overline{K}_1 + \overline{K}_2)e^{-t}$,

$$\|\widehat{x}_N^{(b)} - x_*\| \leq \left(\frac{120 C_5(p)\sigma_*^2 \delta^{2p} s^p \rho^{2(p-1)} \Theta\left(\Theta + t\right)}{N} \right)^{\frac{1}{2(p-1)}}.$$

We then set $C_7(p) := \overline{C}_7(p) \bigvee 120 C_5(p)$ to obtain the value of the last multiplicative term.

To summarize, in both cases, after termination of the algorithm, the bound of Theorem 3.2.1 holds with probability at least $1 - 4(\overline{K}_1 + \overline{K}_2)e^{-t}$. This concludes the proof. □

Part III

Non-Euclidean Accelerated Methods for Sparse Recovery

Chapter 4

Accelerated Stochastic Approximation with State-Dependent Noise

Chapter Abstract

We consider a class of stochastic smooth convex optimization problems under rather general assumptions on the noise in the stochastic gradient observation. As opposed to the classical problem setting in which the variance of noise is assumed to be uniformly bounded, herein we assume that the variance of stochastic gradients is related to the "sub-optimality" of the approximate solutions delivered by the algorithm. Such problems naturally arise in a variety of applications, in particular, in the well-known generalized linear regression problem in statistics. However, to the best of our knowledge, none of the existing stochastic approximation algorithms for solving this class of problems attain optimality in terms of the dependence on accuracy, problem parameters, and mini-batch size.

We discuss two non-Euclidean accelerated stochastic approximation routines—stochastic accelerated gradient descent (SAGD) and stochastic gradient extrapolation (SGE)—which carry a particular duality relationship. We show that both SAGD and SGE, under appropriate conditions, achieve the optimal convergence rate, attaining the optimal iteration and sample complexities simultaneously. However, corresponding assumptions for the SGE algorithm are more general; they allow, for instance, for efficient application of the SGE to statistical estimation problems under heavy tail noises and discontinuous score functions. We also discuss the application of the SGE to problems satisfying quadratic growth conditions, and show how it can be used to recover sparse solutions. Finally, we report on some simulation experiments to illustrate numerical performance of our proposed algorithms in high-dimensional settings.

4.1 Introduction

This paper focuses on the stochastic optimization problem given by

$$f^* := \min_{x \in X} f(x) \tag{4.1.1}$$

where X is a closed convex subset of a Euclidean space E and $f : X \to \mathbf{R}$ is a smooth convex function with Lipschitz continuous gradient, i.e., for some $L \geq 0$,

$$0 \leq f(y) - f(x) - \langle \nabla f(x), x - y \rangle \leq \frac{L}{2}\|y - x\|^2, \quad \forall x, y \in X. \tag{4.1.2}$$

We assume throughout the paper that the set of optimal solutions X^* is nonempty.

We consider the stochastic setting where only stochastic first-order information about f is available for solving problem (4.1.1). Specifically, at the current search point $x_t \in X$, a stochastic oracle (SO) generates the stochastic operator $\mathcal{G}(x_t, \xi_t)$, where $\xi_t \in \Xi$ denotes a random variable, whose probability distribution is supported on a set Ξ. We assume ξ_t is independent of $x_0, ..., x_t$, and $\{\xi_t\}_{t\geq 0}$ are mutually independent. We assume that $\mathcal{G}(x_t, \xi_t)$ is an unbiased estimator of $g(x_t) = \nabla f(x_t)$ satisfying

$$\mathbf{E}_{\xi_t}[\mathcal{G}(x_t, \xi_t)] = g(x_t) \tag{4.1.3}$$

(expectation w.r.t. to the distribution ξ_t).

Stochastic approximation (SA) and stochastic mirror descent (SMD) methods are routinely used to solve stochastic optimization problems; see, e.g., [33, 51, 104, 128]. More specifically, it was shown in [33] that SMD can achieve the optimal sample complexity for general nonsmooth optimization and saddle point problems. For smooth stochastic optimization problems, Lan [24, 129] introduced an accelerated stochastic approximation (AC-SA), also known as stochastic accelerated gradient descent (SAGD), which was obtained by replacing exact gradients with their unbiased estimators in the celebrated accelerated gradient methods [54] (see also [130–133] for early developments). It was shown in [24] that AC-SA achieves the optimal sample complexity for smooth, nonsmooth and stochastic convex optimization (see [25, 26] for generalization to the strongly convex settings). It should be noted that the original analysis of AC-SA in [24] was carried out under *uniformly bounded variance condition*

$$\mathbf{E}_{\xi_t}[\|\mathcal{G}(x_t, \xi_t) - g(x_t)\|_*^2] \leq \sigma^2 \tag{4.1.4}$$

where $\|\cdot\|_*$ is the norm conjugate to $\|\cdot\|$. However, it has been observed recently (see, e.g., [59, 87]) that this uniformly bounded variance condition is not necessarily satisfied in some important applications in which the variance of the stochastic gradient depends on the search point x_t. As a motivation, consider the fundamental to Statistical Learning problem of parameter estimation in the *generalized linear regression* (GLR) model in which one aims to estimate an unknown parameter vector $x^* \in X \subset \mathbf{R}^n$ given observations (ϕ_t, η_t),

$$\eta_t = u(\phi_t^T x^*) + \zeta_t, \quad t = 1, 2, ..., \tag{4.1.5}$$

where, in generalized linear models terminology, $\eta_t \in \mathbf{R}$ are responses, $\phi_t \in \mathbf{R}^n$ are random regressors, $\zeta_t \in \mathbf{R}$ are zero-mean random noises which are assumed to be mutually independent and independent of ϕ_t, and $u : \mathbf{R} \to \mathbf{R}$ is the (generally nonlinear) "activation function". Then it follows directly that

$$\mathbf{E}[\phi_t(u(\phi_t^T x^*) - \eta_t)] = \mathbf{E}[\phi_t\, \zeta_t] = 0. \tag{4.1.6}$$

Thus, the problem of recovery of x^* from observations η_t and ϕ_t may be formulated as a stochastic optimization problem. Specifically, when denoting $v : \mathbf{R} \to \mathbf{R}$ the primitive of u, i.e., $v'(t) = u(t)$ and assuming that $x^* \in \operatorname{int} X$, (4.1.6) may be seen as as the optimality condition for the problem

$$\min_{x\in X} \left\{ f(x) := \mathbf{E}[v(\phi^T x) - \phi^T x \eta] \right\}. \tag{4.1.7}$$

Clearly, the gradient of f is given by $g(x) = \mathbf{E}[\phi(u(\phi^T x) - \eta)]$, and one of its unbiased stochastic estimator is $\mathcal{G}(x, \overbrace{(\phi, \zeta)}^{=:\xi}) = \phi(u(\phi^T x) - \eta)$. Under mild assumptions, one can show (see Section 4.6.1) that the noise

$$\mathcal{G}(x_t, (\phi_t, \zeta_t)) - g(x_t) = \phi_t[u(\phi_t^T x_t) - u(\phi_t^T x^*)] - \phi_t \zeta_t + g(x^*) - g(x_t)$$

(here $g(x^*) = 0$) of the gradient observation $\mathcal{G}(x_t, (\phi_t, \zeta_t))$ at x_t satisfies the condition

$$\mathbf{E}_{\xi_t}\left[\|\mathcal{G}(x_t, \xi_t) - g(x_t)\|_*^2\right] \le \sigma_t^2(x_t) = \mathcal{L}[f(x_t) - f^*] + \sigma_*^2 \tag{SN}$$

for some $\mathcal{L}, \sigma_* > 0$. In what follows, with some terminology abuse, we refer to (SN) and similar conditions as *state-dependent noise* assumptions. More generally, if compared to the "uniform noise" condition (4.1.4), (SN) may be seen as a refined assumption on the structure of the stochastic oracle. Furthermore, one can easily verify that in various settings of the GLR problem, condition (SN) holds, while condition (4.1.4) of uniformly bounded noise variance is violated (e.g., in the case of unbounded X and linear u). Similar conditions have been recently introduced in the context of reduced variance stochastic algorithms for finite sum minimization, see, e.g., [134–137] and references therein. Various stochastic optimization problems in which similar state-dependent noise assumptions apply can also be found in recent literature on machine learning [1, 59, 138] and reinforcement learning, see, e.g., [139–142].

Motivated by these aforementioned statistical applications, Juditsky et al. [59] proposed an SMD algorithm that exploits state-dependent noise assumption to attain optimal convergence rates in the situation of "dominating stochastic error", i.e., when the amplitude of the error of the gradient observation is comparable to the amplitude of the gradient ∇f of the problem objective. In the similar setting, the authors of [138] have recently established sharp lower complexity bounds for stochastic optimization under state-dependent noise in the Euclidean setting for both general and strongly convex situations. However, it is well-known that SMD is suboptimal in the so-call mini-batch setting, where the noise of the gradient estimator is reduced by using a batch of samples. This setting has been widely used for applications of stochastic optimization algorithms especially under a distributed computing environment. To achieve the accelerated convergence, the application of the classical SAGD algorithm of [24] to the state-dependent noise setting was the subject of [138, 143, 144]. The authors of [138] proved (expected) optimal accuracy bound $\mathcal{O}\big(\mathcal{H}/k^2\big)$ after k iterations for the "standard" SAGD under condition of uniform (for all ξ) $\mathcal{H}$-Lipschitz continuity on the stochastic operator $\mathcal{G}(\cdot, \xi)$. However, this assumption impose significant limitations on the form of the stochastic operator in statistical learning applications and is violated in the simple case of unbounded (e.g., Gaussian) regressors ϕ_t, etc. To conclude, in spite of these efforts, to the best of our knowledge, the question of building an optimal stochastic approximation routine of general smooth convex optimization under state-dependent noise assumption (SN) (when only ∇f rather than $\mathcal{G}(\cdot, \xi)$ is Lipschitz continuous) has not received a complete answer.

4.1.1 Contributions and organization

Given the state of affairs, this paper focuses on designing accelerated algorithms and providing sharp analysis for the general stochastic optimization problem (4.1.1) with state-dependent noise. Our contribution is threefold.

1. We analyze the convergence rates of the generic (non-Euclidean) SAGD for solving stochastic optimization with state-dependent noise. We show that under condition (SN), SAGD attains a convergence rate

$$\mathcal{O}\left(\frac{LR^2}{k^2} + \frac{\mathcal{L}R^2}{km} + \frac{\sqrt{L\mathcal{L}}R^2}{k\sqrt{m}} + \sqrt{\frac{\sigma_*^2 R^2}{km}}\right)$$

where k is the number of iterations, R is the initial distance to the optimal solution, and m is the batch size. The terms in the above bound are optimal, except for the third term which

has a sub-optimal dependence on the batch size m. As a consequence, in order to achieve the optimal iteration complexity $\mathcal{O}\big(\sqrt{LR^2/\epsilon}\big)$, SAGD requires a larger batch size, resulting in a sub-optimal sample complexity[1]

$$\mathcal{O}\left(\sqrt{\frac{LR^2}{\epsilon}} + \frac{\sqrt{L}\mathcal{L}R^3}{\epsilon^{3/2}} + \frac{R^2\sigma_*^2}{\epsilon^2}\right).$$

However, imposing the condition of boundness of the second moment of the Lipschitz constant of the gradient observation $\mathcal{G}(\cdot,\xi)$ allows to improve the second term in the above sample complexity bound to $\mathcal{O}(1/\epsilon)$. The corresponding iteration complexity bound $\mathcal{O}\big(\sqrt{LR^2/\epsilon}\big)$ is an improvement w.r.t. the bound $\mathcal{O}\big(\sqrt{\mathcal{H}R^2/\epsilon}\big)$ of [138], and our assumption is more general than the "uniform" Lipschitz continuity assumption in [138]. Moreover, unlike [138], our analysis does not require the feasible region X to be a bounded set.

2. Under state-dependent noise assumption (SN) we analyze an alternative accelerated SA algorithm—stochastic gradient extrapolation method (SGE). The gradient extrapolation method was introduced in [145] by exchanging the primal and the dual variables in a game interpretation of Nesterov's accelerated gradient method (see [145] and Chapter 3 and 4 of [129]). SGE uses the same sequence of points for both gradient estimations and output solutions. This appears to be a significant advantage of the SGE over the SAGD in the present setting and allows for direct "compensation" of the state-dependent noise term by the suboptimality gap of approximate solutions. As a result, SGE achieves the optimal convergence rate after k iterations

$$\mathcal{O}\left(\frac{LR^2}{k^2} + \frac{\mathcal{L}R^2}{km} + \sqrt{\frac{\sigma_*^2R^2}{km}}\right).$$

Consequently, it attains the optimal iteration complexity $\mathcal{O}\big(\sqrt{LR^2/\epsilon}\big)$ along with the optimal sample complexity $\mathcal{O}\big(\sqrt{LR^2/\epsilon} + \mathcal{L}R^2/\epsilon + R^2\sigma_*^2/\epsilon^2\big)$, as supported by lower bounds in [132, 138].

3. We propose a multi-stage algorithm with restarts for solving problems satisfying the quadratic growth condition stating that for some $\mu > 0$ and $x^* \in X$,[2]

$$f(x) - f^* \geq \frac{\mu}{2}\|x - x^*\|^2, \ \forall x \in X. \tag{4.1.9}$$

We show that this algorithm achieves the optimal iteration complexity $\mathcal{O}\big(\sqrt{L/\mu}\ln(1/\epsilon)\big)$ and the optimal sample complexity $\mathcal{O}\big(\sqrt{L/\mu}\ln(1/\epsilon) + \mathcal{L}/\mu\ln(1/\epsilon) + \sigma_*^2/(\mu\epsilon)\big)$ simultaneously.

[1] In what follows we refer to the total number $N = N(\epsilon)$ of calls to the stochastic oracle which are necessary for the approximate solution $\widehat{x}_k$ after $k = k(\epsilon)$ iterations and N oracle calls to attain the expected (in)accuracy ϵ, i.e.,

$$\mathbf{E}[f(\widehat{x}_k)] - f^* \leq \epsilon \tag{4.1.8}$$

as *sample* (or *information*) ϵ-*complexity* of the method. We also call *iteration* ϵ-*complexity* the minimal iteration count k such that (4.1.8) holds.

[2] We suppose for convenience that in this case the optimal solution x^* is unique. Note that (4.1.9) can be seen as a relaxation of the strong convexity assumption, i.e., for any $x, y \in X$,

$$f(y) - f(x) - \langle \nabla f(x), y - x\rangle \geq \mu\|y - x\|^2/2.$$

Furthermore, we specify the multi-stage SGE to solve the sparse recovery problem. This is done by incorporating hard-thresholding of the approximate solution at the end of each algorithm stage to enforce the sparsity. The convergence results match the quadratic growth setting up to a multiplicative factor of the sparsity level s, with only logarithmic dependence on the dimension n. To the best of our knowledge, the corresponding convergence guarantees are new to the optimization literature and lead to extremely efficient algorithms for sparse recovery problems in the distributed setting.

Remaining sections of the paper are organized as follows. In Section 4.2, we formalize the general problem statement and introduce the mini-batch setup. In Section 4.3, we study the SAGD method for solving the general convex problem with state-dependent noise. Section 4.4 introduces SGE and provides convergence guarantees in the general convex problem. We introduce the multi-stage SGE for solving problems satisfying quadratic growth condition in Section 4.5. Section 4.6 further extends the multi-stage SGE to be applied to the sparse recovery problem. Finally, in Section 4.7, we present some results of a preliminary simulation study illustrating the numerical performance of the proposed algorithms in the high-dimensional setting of sparse recovery. Proofs of the statements are postponed until the appendix.

4.1.2 Notation

For any $n \geq 1$, we use $[n]$ to denote the set of integers $\{1, ..., n\}$. For $x \in \mathbf{R}$, we let $(x)_+ = \max(x, 0)$ and $(x)_- = \max(-x, 0)$. In what follows, E is a finite-dimensional real-vector (Euclidean) space. Given a norm $\|\cdot\|$ on E, the associated dual norm $\|\cdot\|_*$ is defined as $\|z\|_* := \sup\{\langle x, z\rangle : \|x\| \leq 1\}$. We define $\omega : E \to \mathbf{R}$, the *distance generating function,* which is a continuously differentiable strongly convex function with modulus 1. Without loss of generality, we assume that $\omega(x) \geq \omega(0) = 0$ and for some $\Omega \geq 1$,

$$\omega(x) \leq \tfrac{\Omega}{2}\|x\|^2, \quad \forall x \in E. \tag{4.1.10}$$

Ideally, we want Ω to be "not too large". Meanwhile, a desired distance generating function should be "prox-friendly", i.e., for any $a \in E$, the minimization problem

$$\min_{x \in X}\{\langle a, x\rangle + \omega(x)\}$$

can be easily solved. Note that when $\|\cdot\|$ is the Euclidean norm, we can set $\omega(x) = \frac{\|x\|_2^2}{2}$, and the corresponding $\Omega = 1$. Another "standard" choice is $\|\cdot\| = \|\cdot\|_1$, $\|\cdot\|_* = \|\cdot\|_\infty$, and one can choose the distance generating function ω (cf. [109])

$$\omega(x) = \tfrac{1}{2}e \ln n \cdot n^{(p-1)(2-p)/p}\|x\|_p^2, \quad p = 1 + \tfrac{1}{\ln n};$$

the corresponding Ω satisfies $\Omega \leq e^2 \ln n$ in this case.

Given an initialization $x_0 \in X$, we define the x_0-associated Bregman's divergence of $x, y \in X$ as

$$V_{x_0}(x, y) = \omega(y - x_0) - \omega(x - x_0) - \langle \nabla\omega(x - x_0), y - x\rangle.$$

Clearly, for any $y, x, x_0 \in X$, we have

$$V_{x_0}(x_0, y) \leq \tfrac{\Omega}{2}\|y - x_0\|^2 \quad \text{and} \quad V_{x_0}(x, y) \geq \tfrac{1}{2}\|x - y\|^2. \tag{4.1.11}$$

We use the shorthand notation $V(x, y)$ for $V_{x_0}(x, y)$ when x_0 is clear in the content.

Unless stated otherwise, all relations between random variables are assumed to hold almost surely.

4.2 Problem statement

We summarize below the setting of the stochastic optimization problem and our assumptions.

4.2.1 Assumptions

The problem under consideration is a stochastic optimization problem (4.1.1) with a convex and smooth objective function as given in (4.1.2).

We assume that stochastic oracle $\mathcal{G}(\cdot,\cdot)$ is unbiased, i.e., satisfies (4.1.3). Furthermore, we consider the following "state-dependent" oracle noise assumption.

- [*State-dependent variance*] We assume that for some $\mathcal{L} < \infty$,

$$\mathbf{E}_{\xi_t}\left[\|\mathcal{G}(x_t,\xi_t) - g(x_t)\|_*^2\right] \le \sigma_t^2 := \mathcal{L}[f(x_t) - f^* - \langle g(x^*), x_t - x^*\rangle] + \sigma_*^2, \quad t \in \mathbb{Z}_+, \tag{SN}$$

  where $x^* \in X^*$.

Assumption (SN) can be further weakened to $\sigma_t^2 := \mathcal{L}[f(x_t) - f^*] + \sigma_*^2$. In the case of unconditional minimizer $x^* \in \operatorname{int} X$, the latter condition is clearly equivalent to (SN). In the case of $g(x^*) \neq 0$, utilizing the relaxed assumption results in convergence guarantees which depend explicitly on $f(x_0) - f^*$. We use (SN) for the sake of convenience, the term $\langle g(x^*), x_t - x^*\rangle$ in the right-hand side when combined with smoothness of f allows us to upper bound the variance of $\mathcal{G}(x_0,\xi_t)$ at x_0 by the term proportional to $\|x_0 - x^*\|^2$; see, e.g., (4.4.8).

When proving the convergence rates for the stochastic variant of Nesterov's accelerated gradient descent method (SAGD; see Section 4.3), we also consider the following assumption:

- [*Lipschitz continuous stochastic gradient*] For each $\xi \in \Xi$, there exists a $\mathcal{K}(\xi) > 0$, such that $\mathbf{E}_\xi[\mathcal{K}(\xi)^2] < \infty$ and

$$\|\mathcal{G}(x,\xi) - \mathcal{G}(y,\xi)\|_* \le \mathcal{K}(\xi)\|x - y\|, \quad \forall x, y \in X. \tag{LP}$$

This assumption relaxes the assumption in [138] that assumes $\mathcal{G}(\cdot,\xi)$ is $\mathcal{H}$-Lipschitz continuous for all $\xi \in \Xi$. For instance, for GLR model with Gaussian/sub-Gaussian regressors ϕ_t, (LP) holds, but the $\mathcal{H}$-uniform Lipschitz continuous condition is violated. Nevertheless, (LP) is not a necessary condition of Assumption (SN) since the latter one may hold in various situations of interest where $\mathcal{G}(\cdot,\xi)$ is not Lipschitz (and even not continuous). In Figure 4.1 we present the plot of the expectation in the left-hand side of (SN) as a function of x for two choices of scalar discontinuous gradient observation $\mathcal{G}_1(x,\phi) = \phi \cdot u(\phi x)$ and $\mathcal{G}_2(x,[\phi,\zeta]) = \phi \cdot u(\phi x + \zeta)$ where ϕ and ζ are independent r.v. with Student t_4 distribution and $u(t) = (\frac{1}{2} + \sqrt{|t|})\operatorname{sign}(t)$.

Given the limitations of Assumption (LP), we will only use it partially in Section 4.3 in order to improve the convergence rates of SAGD. In the following sections, we will propose an alternative accelerated algorithm called SGE that does not rely on Assumption (LP) but attains stronger convergence guarantees; see Section 4.4 for more details.

4.2.2 Mini-batch setup

We consider the mini-batch approach widely used in practice. Specifically, we assume that at each search point u_t, the stochastic oracle is called repeatedly, thus generating m_t i.i.d. samples $\{\xi_{t,i}\}_{i=1}^{m_t}$, m_t being the number of oracle calls. Next, we compute the unbiased estimator $G_t(u_t)$ of $g(u_t)$,

$$G_t(u_t) = \frac{1}{m_t}\sum_{i=1}^{m_t} \mathcal{G}(u_t,\xi_{t,i}).$$

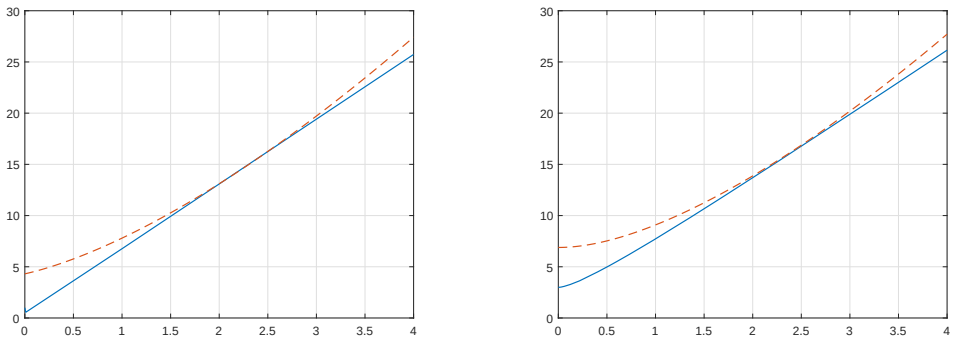


Figure 4.1: Left plot: variance of the stochastic oracle $\mathcal{G}_1(x,\xi)$ as function of x (solid line) and upper bound $4\sigma^2(0)+3[f(x)-f(0)]$ (dashed line); right plot: variance of $\mathcal{G}_2(x,\xi)$ as function of x (solid line) and upper bound $2.3\sigma^2(0)+3[f(x)-f(0)]$ (dashed line).

We define filtration $\mathcal{F}_t = \sigma\big(u_0, \xi_{0,1}, ..., \xi_{0,m_0}, ..., \xi_{t,1}, ..., \xi_{t,m_t}\big)$, so that

- random variables in $\{\xi_{t,i}\}_{i=1}^{m_t}$ are $\mathcal{F}_t$-measurable.
- search points with index t which are deterministic functions of $G_\tau(u_\tau), \tau \le t-1$, are $\mathcal{F}_{t-1}$-measurable.[3]

We use the shorthand notation $\mathbf{E}_{\lceil t \rceil}$ to denote the conditional expectation with respect to the filtration $\mathcal{F}_t$.

Based on the state-dependent noise Assumption (SN), we have the following characterization of the properties of the mini-batch estimator:

Lemma 4.2.1 *The mini-batch estimator G_t satisfies*

$$\mathbf{E}_{\lceil t-1 \rceil}[\|G_t(u_t) - g(u_t)\|_*^2] \le \frac{\overline{\Omega}}{m_t} \cdot \mathbf{E}_{\lceil t-1 \rceil}[\|\mathcal{G}_t(u_t,\xi_{t,1}) - g(u_t)\|_*^2], \tag{4.2.1}$$

where $\overline{\Omega} := \begin{cases} 1, & \text{when } m_t = 1 \\ \Omega, & \text{when } m_t \ge 2, \end{cases}$ *and Ω is defined in (4.1.10). Consequently, under Assumption (SN), we have*

$$\mathbf{E}_{\lceil t-1 \rceil}[\|G_t(u_t) - g(u_t)\|_*^2] \le \frac{\overline{\Omega}}{m_t}\left\{\mathcal{L}[f(u_t) - f^* - \langle g(x^*), u_t - x^*\rangle] + \sigma_*^2\right\}. \tag{4.2.2}$$

In the bound (4.2.2), one has $\Omega = 1$ in the case of Euclidean setup (when $\|\cdot\| = \|\cdot\|_* = \|\cdot\|_2$); in the ℓ_1-setup, one has $\Omega = \mathcal{O}(\ln n)$. One can easily see that, in general, the logarithmic factor is unavoidable in this case. Indeed, when $\mathcal{G}(x,\xi_i) \sim \mathcal{R}(n)$ (are n-dimensional Rademacher vectors), one has $\|\mathcal{G}(x,\xi_i)\|_\infty \le 1$, while

$$\mathbf{E}\left[\left\|\frac{1}{m}\sum_{i=1}^m \mathcal{G}(x,\xi_i)\right\|_\infty^2\right] \asymp \frac{\ln n}{m}.$$

[3]Note that u_t here is a general place holder for a $\mathcal{F}_{t-1}$-measurable search point. With a slight ambiguity of notation, search points z_t and x_t in Algorithms 3 and 4 are $\mathcal{F}_t$-measurable.

On the other hand, for $\mathcal{G}(x,\xi_i) \sim \mathcal{N}(0, I_n)$, we have that $\mathbf{E}[\|\mathcal{G}(x,\xi_i)\|_\infty^2] = 2\ln n$, and $\frac{1}{m}\sum_{i=1}^m \mathcal{G}(x,\xi_i) \sim \mathcal{N}(0, I_n/m)$ with

$$\mathbf{E}\left[\left\|\frac{1}{m}\sum_{i=1}^m \mathcal{G}(x,\xi_i)\right\|_\infty^2\right] = \frac{2\ln n}{m},$$

and there is no extra logarithmic factor in the bound (4.2.2) for the norm of the sum in this situation (in a sense, it is "already paid" in the bound for the expectation of the norm of a summand).

In the following sections, we present accelerated algorithms equipped with mini-batches that achieve the optimal iteration complexity for stochastic optimization with state-dependent noise.

4.3 SAGD for general convex problem with state-dependent noise

Algorithm 1 describes a mini-batch variant of the standard stochastic accelerated gradient descent (SAGD) method. The SAGD method, also called accelerated stochastic approximation (AC-SA), maintains three sequences of points. Specifically, $\{z_t\}$ is the sequence of "prox-centers" for the prox-mapping updates (4.3.1b), and y_t and x_t are weighted averages of the past z_t's; $\{y_t\}$ is the sequence of search points where stochastic gradients are estimated using mini-batches, and points $\{x_t\}$ is the trajectory of approximate solutions (outputs) at each iteration.

Algorithm 3 Stochastic Accelerated Gradient Descent method (SAGD)

Input: initial point $z_0 = x_0$, nonnegative nonrandom parameters $\{\beta_t\}$ and $\{\eta_t\}$, and batch size $\{m_t\}$.

for $t = 1, 2, \ldots,$ **do**

$$y_t = (1-\beta_t)x_{t-1} + \beta_t z_{t-1} \tag{4.3.1a}$$

$$G_t = \tfrac{1}{m_t}\textstyle\sum_{i=1}^{m_t} \mathcal{G}(y_t, \xi_{t,i}).$$

$$z_t = \operatorname*{argmin}_{z\in X}\{\langle G_t, z\rangle + \eta_t V(z_{t-1}, z)\}, \tag{4.3.1b}$$

$$x_t = (1-\beta_t)x_{t-1} + \beta_t z_t, \tag{4.3.1c}$$

end for

We start with the following characterization of the approximate solution x_k of Algorithm 3.

Theorem 4.3.1 *Suppose Assumption (SN) is satisfied. Let the algorithmic parameters β_t and η_t satisfy for some $\theta_t \geq 0$,*

$$\theta_t\beta_t\eta_t \leq \theta_{t-1}\beta_{t-1}\eta_{t-1},\ t = 2, ..., k \tag{4.3.2a}$$

$$\eta_t > L\beta_t,\ t = 1, ..., k. \tag{4.3.2b}$$

Furthermore, suppose that $\beta_1 = 1$ and

$$\theta_t(1-\beta_t)(1 + r_t\mathcal{L}) \leq \theta_{t-1},\ t = 2, ..., k. \tag{4.3.3}$$

Then

$$\theta_k\mathbf{E}[f(x_k) - f^*] + \theta_k\beta_k\eta_k\mathbf{E}[V(z_k, x^*)]$$
$$\leq \theta_1\beta_1\eta_1 V(z_0, x^*) + \sum_{t=1}^k \theta_t r_t\beta_t L\mathcal{L}\mathbf{E}[V(z_{t-1}, x^*)] + \sum_{t=1}^k \theta_t r_t\sigma_*^2 \tag{4.3.4}$$

where $r_t := \frac{\beta_t \overline{\Omega}}{2(\eta_t - L\beta_t)m_t}$ *and* $\overline{\Omega}$ *is defined in Lemma 4.2.1.*

Corollary 4.3.1 *In the premise of Theorem 4.3.1, suppose that* $k \geq 2$, $V(z_0, x^*) \leq D^2$, *and let* $\theta_t = (t+1)(t+2)$, $\beta_t = \frac{3}{t+2}$, $m_t = m$, *and* $\eta_t = \frac{\eta}{t+1}$ *with*

$$\eta = \max\left\{4L, \frac{6\overline{\Omega}(k-1)\mathcal{L}}{m}, \sqrt{\frac{9(k+1)^2\overline{\Omega}L\mathcal{L}}{m}}, \sqrt{\frac{2(k+2)^3\overline{\Omega}\sigma_*^2}{3D^2m}}\right\}.$$

Then $\mathbf{E}[V(z_t, x^*)] \leq 3D^2$ *for all* $1 \leq t \leq k$, *and*

$$\mathbf{E}[f(x_k) - f^*] \leq \frac{12LD^2}{(k+1)(k+2)} + \frac{6\overline{\Omega}\mathcal{L}D^2}{(k+2)m} + \frac{18\sqrt{\overline{\Omega}L\mathcal{L}}D^2}{(k+1)\sqrt{m}} + \frac{4\sqrt{2\overline{\Omega}\sigma_*^2 D^2}}{\sqrt{(k+1)m}}. \tag{4.3.5}$$

Remarks. Bound (4.3.5) of Corollary 4.3.1 allows us to establish the complexity bounds for the SAGD algorithm when solving a general convex problem with state-dependent noise. Let us first consider the case when $m = 1$. In this case, we have $\overline{\Omega} = 1$, thus the total number of iterations/oracle calls used by SAGD to find an ϵ-optimal solution, i.e., $\widehat{x} \in X$ such that $\mathbf{E}[f(\widehat{x}) - f^*] \leq \epsilon$, is bounded by

$$\mathcal{O}\left(\sqrt{\frac{LD^2}{\epsilon}} + \frac{\mathcal{L}D^2}{\epsilon} + \frac{\sqrt{L\mathcal{L}}D^2}{\epsilon} + \frac{D^2\sigma_*^2}{\epsilon^2}\right). \tag{4.3.6}$$

Considering the setting with $L = \mathcal{O}(\mathcal{L})$, as in the context of [138], this upper bound matches the optimal sample complexity under Assumption (SN), supported by the lower bound in Theorem 4 of [138].

In the mini-batch setting (where $\overline{\Omega} = \Omega$), the bound in (4.3.5) means that in order to achieve the optimal iteration complexity of $\mathcal{O}\left(\sqrt{LD^2/\epsilon}\right)$, the batch size $m \geq \max\left\{1, \frac{k\Omega\mathcal{L}}{L}, \frac{k^2\Omega\mathcal{L}}{L}, \frac{k^3\Omega\sigma_*^2}{D^2L^2}\right\} = \max\left\{1, \frac{k^2\Omega\mathcal{L}}{L}, \frac{k^3\Omega\sigma_*^2}{D^2L^2}\right\}$ is needed. Consequently, the total sample complexity is bounded by

$$\mathcal{O}\left(\sqrt{\frac{LD^2}{\epsilon}} + \frac{\sqrt{L}\mathcal{L}\Omega D^3}{\epsilon^{3/2}} + \frac{\Omega D^2\sigma_*^2}{\epsilon^2}\right) \tag{4.3.7}$$

in this situation. When comparing (4.3.7) to (4.3.6), we observe that the second term in (4.3.7) is sub-optimal. This implies that, based on our analysis, SAGD does not simultaneously achieve the optimal iteration complexity and sample complexity under Assumption (SN).

The analysis on the convergence of SAGD in the state-dependent noise setting reveals the "bottleneck": in the recursion (4.3.1), different points y_t and x_t are used for gradient estimations and output solutions. Improving the convergence rates in Corollary 4.3.1 requires better control of the objective value at the points of stochastic gradient estimation. This can be achieved by imposing the Lipschitz regularity assumption in (LP) on stochastic gradients.

Lemma 4.3.1 *Suppose Assumptions (SN) and (LP) hold. Let* $\{x_t\}$, $\{y_t\}$ *and* $\{z_t\}$ *be generated by Algorithm 3. We have*

$$\mathbf{E}_{[t-1]}\left[\|G_t - g(y_t)\|_*^2\right] \leq \frac{\overline{\Omega}}{m_t}\left\{\bar{\mathcal{K}}^2\beta_t^2\|x_{t-1} - z_{t-1}\|^2 + 3\mathcal{L}[f(x_{t-1}) - f^* - \langle g(x^*), x_{t-1} - x^*\rangle] + 3\sigma_*^2\right\},$$

where

$$\bar{\mathcal{K}} := \left(3\mathbf{E}_{\xi_{t,1}}[\mathcal{K}(\xi_{t,1})^2] + 3L^2\right)^{1/2}. \tag{4.3.8}$$

We have the following analog of Theorem 4.3.1 under Assumptions (SN) and (LP).

Theorem 4.3.2 *Let Assumptions (SN) and (LP) hold. Let β_t, η_t satisfy (4.3.2) for some $\theta_t \geq 0$, and let also $\beta_1 = 1$ and*

$$\theta_t(1 - \beta_t + 3r_t\mathcal{L}) \leq \theta_{t-1},\ t = 2, ..., k. \tag{4.3.9}$$

Then

$$\begin{aligned}&\theta_k \mathbf{E}[f(x_k) - f^*] + \theta_k\beta_k\eta_k\mathbf{E}[V(z_k, x^*)]\\ &\quad\leq \theta_1\beta_1\eta_1 V(z_0, x^*) + \tfrac{3\theta_1 r_1 \mathcal{L}L}{2}\|z_0 - x^*\|^2 + \sum_{t=1}^{k}\theta_t r_t\beta_t^2\bar{\mathcal{K}}^2\mathbf{E}\|\bar{z}_{t-1} - z_{t-1}\|^2 + \sum_{t=1}^{k} 3\theta_t r_t\sigma_*^2\end{aligned} \tag{4.3.10}$$

where $\bar{\mathcal{K}}$ is defined in (4.3.8) and $r_t := \frac{\beta_t\overline{\Omega}}{2(\eta_t - L\beta_t)m_t}$.

Corollary 4.3.2 *In the premise of Theorem 4.3.2, suppose that $k \geq 2$, $V(z_0, x^*) \leq D^2$, and let $\beta_t = \frac{3}{t+2}$, $m_t = m$ and $\eta_t = \frac{\eta}{t+1}$ with*

$$\eta = \max\left\{4L, \frac{18\overline{\Omega}(k+1)\mathcal{L}}{m}, 12\sqrt{\frac{k\overline{\Omega}\bar{\mathcal{K}}^2}{m}}, \sqrt{\frac{2(k+2)^3\overline{\Omega}\sigma_*^2}{D^2 m}}\right\}.$$

Then $\mathbf{E}[V(z_t, z_0)] \leq 3D^2$, $t = 1, ..., k$, and

$$\mathbf{E}[f(x_k) - f^*] \leq \frac{13LD^2}{(k+1)(k+2)} + \frac{54\overline{\Omega}\mathcal{L}D^2}{(k+2)m} + \frac{72\bar{\mathcal{K}}D^2}{(k+2)}\cdot\sqrt{\frac{\overline{\Omega}}{m(k+1)}} + 4\sqrt{\frac{6\overline{\Omega}\sigma_*^2 D^2}{(k+1)m}}. \tag{4.3.11}$$

Remarks. Let $m = 1$ (and $\overline{\Omega} = 1$ in Lemmas 4.2.1 and 4.3.1). By (4.3.11), the iteration/sample complexity of the SAGD is bounded by

$$\mathcal{O}\left\{\sqrt{\frac{LD^2}{\epsilon}} + \frac{\mathcal{L}D^2}{\epsilon} + \left(\frac{\bar{\mathcal{K}}D^2}{\epsilon}\right)^{\frac{2}{3}} + \frac{D^2\sigma_*^2}{\epsilon^2}\right\}.$$

When $\epsilon = \mathcal{O}(L^3D^2/\bar{\mathcal{K}}^2)$, this sample complexity is optimal.

In the mini-batch setting (when $\overline{\Omega} = \Omega$ in Lemmas 4.2.1 and 4.3.1), in order to obtain the optimal iteration complexity $\mathcal{O}\left(\sqrt{LD^2/\epsilon}\right)$, SAGD batch size m should be in the order of

$$\max\left\{1, \frac{k\Omega\mathcal{L}}{L}, \frac{k\Omega\bar{\mathcal{K}}^2}{L^2}, \frac{k^3\Omega\sigma_*^2}{L^2D^2}\right\}.$$

As a result, the corresponding sample complexity becomes

$$\mathcal{O}\left\{\sqrt{\frac{LD^2}{\epsilon}} + \frac{\Omega\mathcal{L}D^2}{\epsilon} + \frac{\Omega\bar{\mathcal{K}}^2D^2}{L\epsilon} + \frac{\Omega D^2\sigma_*^2}{\epsilon^2}\right\}. \tag{4.3.12}$$

When $\bar{\mathcal{K}}^2 = \mathcal{O}(L\mathcal{L})$, this complexity bound is optimal (cf. Theorem 4 of [138]). The result of Corollary 4.3.2 refines the corresponding statement of [138] in three aspects. First, the corresponding iteration complexity bound $\mathcal{O}\left(\sqrt{LD^2/\epsilon}\right)$ is stated in terms of the Lipschitz constant of the expected gradient. Second, it relies upon Assumption (LP) which is significantly weaker than the assumption of uniform Lipschitz continuity of the stochastic gradient observation $\mathcal{G}(x, \cdot)$ used in [138]. Third, unlike [138], our analysis does not require the feasible region X to be a bounded set.

4.4 SGE for general convex problem with state-dependent noise

In this section, we discuss an alternative acceleration scheme to solve the general smooth convex problem with state-dependent noise which we refer to as stochastic gradient extrapolation (SGE). SGE (Algorithm 4) is a variant of the gradient extrapolation method proposed in [145]. We consider here the mini-batch version of the routine.

Algorithm 4 Stochastic Gradient Extrapolation method (SGE)

Input: initial point $x_0 = z_0$, nonnegative parameters $\{\alpha_t\}$, $\{\eta_t\}$ and $\{\beta_t\}$, and batch size $\{m_t\}$. $G_0 = G_{-1} = \frac{1}{m_0}\sum_{i=1}^{m_0} \mathcal{G}(x_0, \xi_{0,i})$.

for $t = 1, 2, \ldots,$ **do**

$$\widetilde{G}_t = G_{t-1} + \alpha_t(G_{t-1} - G_{t-2}), \tag{4.4.1a}$$

$$z_t = \operatorname*{argmin}_{x \in X}\{\langle \widetilde{G}_t, x\rangle + \eta_t V_{x_0}(z_{t-1}, x)\}, \tag{4.4.1b}$$

$$x_t = (1-\beta_t)x_{t-1} + \beta_t z_t, \tag{4.4.1c}$$

$$G_t = \tfrac{1}{m_t}\textstyle\sum_{i=1}^{m_t} \mathcal{G}(x_t, \xi_{t,i}).$$

end for

The basic iterative scheme presented in Algorithm 4 is conceptually simple. It involves two sequences of search points $\{z_t\}$ and $\{x_t\}$, the latter being weighted averages of the former. Note that both x_t and z_t are $\mathcal{F}_{t-1}$-measurable. Notably, the stochastic gradients are estimated at the search points $\{x_t\}$, which are also the approximate output solutions generated by the algorithm at each iteration. This property brings benefits for dealing with state-dependent noise of the gradient estimation over the stochastic accelerated gradient descent (SAGD) method, where the output and gradient estimation use different sequences.

The special relationship between SGE and SAGD merits an explanation. It has been discussed in detail in [145, Section 3] and [129, Section 5.2]) in the deterministic setting. For the sake of completeness, we summarize the corresponding argument here.

Let us consider the problem of unconstrained minimization

$$\min_x f(x)$$

where $f : E \to \mathbf{R}$ is strictly convex and continuously differentiable. For $\varsigma \in E$ let us denote

$$\varphi(\varsigma) = \max_x \big\{\langle x, \varsigma\rangle - f(x)\big\},$$

so that $\varphi : E \to \mathbf{R}$ is strictly convex and continuously differentiable on E. Then f has the Fenchel representation

$$f(x) = \max_{\varsigma} \big\{ \underbrace{\langle x, \varsigma\rangle - \varphi(\varsigma)}_{=:F(x,\varsigma)} \big\}, \quad x \in E,$$

and we can reformulate the original minimization problem as a saddle point problem:

$$f^* := \min_x \left\{ \max_{\varsigma} F(x, \varsigma) \right\}.$$

Let us define the Bregman divergence associated with φ according to

$$W_f(\chi, \varsigma) = \varphi(\varsigma) - [\varphi(\chi) + \langle \nabla\varphi(\chi), \varsigma - \chi\rangle], \quad \chi, \varsigma \in E,$$

and the (generalized) prox-mapping

$$\operatorname{argmax}_{\varsigma}\Big\{F(z,\varsigma)-\tau W_f(\chi,\varsigma)\Big\},\quad z,\chi\in E. \tag{4.4.2}$$

As shown in Lemma 1 of [146] (cf. also Lemma 3.6 of [129]), the maximizer of (4.4.2) is the value of ∇f at certain $x\in X$, specifically,

$$\nabla f(x)=\operatorname{argmax}_{\varsigma}\left\{F(z,\varsigma)-\tau W_f(\chi,\varsigma)\right\}$$

where $x=[z+\tau\nabla\varphi(\chi)]/(1+\tau)$.

Using the above observation, we can rewrite the corresponding deterministic version of the SGE recursion in a primal-dual form. It is initialized with $(\varsigma_{-1},\varsigma_0)$ and $x_0=\nabla\varphi(\varsigma_0)$, with the updates (x_t,ς_t) computed according to

$$\tilde{\varsigma}_t=\varsigma_{t-1}+\alpha_t(\varsigma_{t-1}-\varsigma_{t-2}), \tag{4.4.3a}$$
$$z_t=\operatorname*{argmin}_x\left\{\langle\tilde{\varsigma}_t,x\rangle+\eta_t V_{x_0}(z_{t-1},x)\right\}, \tag{4.4.3b}$$
$$\varsigma_t=\operatorname{argmax}_{\varsigma}\left\{F(z_t,\varsigma)-\tau_t W_f(\varsigma_{t-1},\varsigma)\right\}. \tag{4.4.3c}$$

Because $\nabla\varphi(\varsigma_{t-1})=x_{t-1}$, by the above, $\varsigma_t=\nabla f(x_t)$ with $x_t=(z_t+\tau_t x_{t-1})/(1+\tau_t)$ which is the definition of x_t in (4.4.1c) with $\beta_t=1/(1+\tau_t)$. The corresponding stochastic iteration (4.4.1) is obtained from (4.4.3) by replacing ς_t with its estimation G_t—the mean of m_t stochastic gradients $\mathcal{G}(x_t,\xi_{t,i})$. Similarly (see [146, Section 2.2] and [129, Section 3.4]), one can show that SAGD iteration can be viewed as a specific stochastic version of the following (deterministic) primal-dual update:

$$\tilde{z}_t=z_{t-1}+\alpha_t(z_{t-1}-z_{t-2}), \tag{4.4.4a}$$
$$\varsigma_t=\operatorname{argmax}_{\varsigma}\left\{F(\tilde{z}_t,\varsigma)-\tau_t W_f(\varsigma_{t-1},\varsigma)\right\}, \tag{4.4.4b}$$
$$z_t=\operatorname*{argmin}_x\left\{\langle\varsigma_t,x\rangle+\eta_t V_{x_0}(z_{t-1},x)\right\}. \tag{4.4.4c}$$

Recursion (4.4.3) can be seen as a dual version of (4.4.4); the principal difference between the two resides in the extrapolation step which is performed in the dual space in (4.4.3a) and in the primal space in (4.4.4a).

We now establish the convergence guarantees of SGE method in expectation, i.e., $\mathbf{E}[f(x_k)-f^*]$. We should stress here that the convergence analysis of SGE under Assumption (SN) is much more involved than the ones for its basic scheme in [145], and the SAGD method in Section 4.3. Therefore, the details are deferred to the appendix.

Theorem 4.4.1 *Suppose Assumption (SN) holds. Assume that the parameters $\{\alpha_t\}$, $\{\eta_t\}$ and $\{\beta_t\}$ of Algorithm 4 satisfy and a nonnegative sequence $\{\theta_t\}$ satisfy*

$$\theta_{t-1}=\alpha_t\theta_t,\ \eta_t\le\alpha_t\eta_{t-1},\ t=2,\dots,k \tag{4.4.5a}$$
$$\frac{\eta_t(1-\beta_t)}{\alpha_t}\ge 5L\beta_t,\quad t=3,\dots,k \tag{4.4.5b}$$
$$\frac{\eta_1\eta_2}{\alpha_2}\ge 25L^2,\ \eta_k(1-\beta_k)\ge L\beta_k, \tag{4.4.5c}$$

Denote

$$q_t:=\frac{\theta_{t+1}(1+\alpha_{t+1}^2)}{\eta_{t+1}}+\frac{\theta_{t+2}\alpha_{t+2}^2}{\eta_{t+2}}+\frac{\theta_{t+2}^2\alpha_{t+2}^2}{\theta_{t+1}\eta_{t+1}}\quad\text{and}\quad\epsilon_t:=\frac{5q_t\overline{\Omega}}{2m_t}. \tag{4.4.6}$$

If $\beta_1 = 1$ and the parameters satisfy

$$\frac{\theta_t(1-\beta_t)}{\beta_t} + \mathcal{L}\epsilon_{t-1} \leq \frac{\theta_{t-1}}{\beta_{t-1}},\ t \geq 2, \tag{4.4.7}$$

then

$$\frac{\theta_k}{\beta_k}\mathbf{E}[f(x_k) - f^*] \leq \theta_1\eta_1 V(x_0, x^*) + \frac{\mathcal{L}\epsilon_0 L}{2}\|x_0 - x^*\|^2 + \sigma_*^2 \cdot \sum_{t=0}^{k-1}\epsilon_t. \tag{4.4.8}$$

We now specify a particular stepsize policy in order to establish the convergence guarantee of SGE.

Corollary 4.4.1 *Let $\theta_t = t,\ \alpha_t = \frac{t-1}{t},\ m_t = m,\ \beta_t = \frac{3}{t+2}$, and $\eta_t = \frac{\eta}{t}$ for $\eta > 0$. Suppose that $V(x_0, x^*) \leq D^2$ and that*

$$\eta = \max\left\{30L, \frac{30\overline{\Omega}(k+2)\mathcal{L}}{m}, \sqrt{\frac{10\overline{\Omega}(k+1)^3\sigma_*^2}{3mD^2}}\right\}.$$

Then

$$\mathbf{E}[f(x_k) - f^*] \leq \frac{91LD^2}{k(k+2)} + \frac{90\overline{\Omega}\mathcal{L}D^2}{mk} + \sqrt{\frac{120\overline{\Omega}\sigma_*^2 D^2}{mk}}. \tag{4.4.9}$$

Furthermore, when $\|x_0 - x^*\| \leq R$ and

$$\eta = \max\left\{30L, \frac{30\overline{\Omega}(k+2)\mathcal{L}}{m}, \sqrt{\frac{20\overline{\Omega}(k+1)^3\sigma_*^2}{3m\Omega R^2}}\right\}, \tag{4.4.10}$$

we have

$$\mathbf{E}[f(x_k) - f^*] \leq \frac{91L\Omega R^2}{2k(k+2)} + \frac{45\Omega\overline{\Omega}\mathcal{L}R^2}{mk} + \sqrt{\frac{60\Omega\overline{\Omega}\sigma_*^2 R^2}{mk}}. \tag{4.4.11}$$

Remarks. The bounds of Corollary 4.4.1 merit some comments. Observe first that SGE achieves the optimal sample complexity

$$\mathcal{O}\left\{\sqrt{\frac{LD^2}{\epsilon}} + \frac{\mathcal{L}D^2}{\epsilon} + \frac{D^2\sigma_*^2}{\epsilon^2}\right\} \tag{4.4.12}$$

in the case of $m = 1$. In the mini-batch setting (where $\overline{\Omega} = \Omega$), by setting the batch size of $m \geq \max\left\{1, \frac{\Omega\mathcal{L}k}{L}, \frac{\Omega k^3\sigma_*^2}{D^2L^2}\right\}$, the iteration complexity of the algorithm is bounded by $\mathcal{O}\left(\sqrt{LD^2/\epsilon}\right)$ and the overall sample complexity is

$$\mathcal{O}\left\{\sqrt{\frac{LD^2}{\epsilon}} + \frac{\mathcal{L}\Omega D^2}{\epsilon} + \frac{\Omega D^2\sigma_*^2}{\epsilon^2}\right\}. \tag{4.4.13}$$

Notably, the SGE attains the optimal iteration and sample complexity bounds.

Similarly, using the fact that $V(x_0, x^*) \leq \frac{\Omega}{2}\|x_0 - x^*\|^2$ (cf. (4.1.11)), (4.4.11) states the convergence rate under condition $\|x_0 - x^*\| \leq R$. This bound will be further used in the proof of the multi-stage SGE method in the next section.

4.5 SGE for convex problem with quadratic growth condition

In this section, we consider the problem setting in which the smooth objective f in (4.1.1) satisfies the quadratic growth condition (cf. (4.1.9)), i.e., when for some $\mu > 0$ and $x^* \in X$

$$f(x) - f^* \geq \frac{\mu}{2}\|x - x^*\|^2, \ \forall x \in X. \tag{4.5.1}$$

We propose a multi-stage routine with restarts that utilizes Algorithm 4 as a working horse.

Algorithm 5 Multi-stage stochastic gradient extrapolation method

Input: initial point $y^0 \in X$. Let R_0 be a positive real.
for $k = 1, 2, \ldots, K$ **do**

(a) Set $R_k = R_0 2^{-k/2}$, $N = \left\lceil 10\sqrt{\frac{2\Omega L}{\mu}} \right\rceil$. Run N iterations of SGE (Algorithm 4) with $x_0 = z_0 = y^{k-1}$ and

$$\begin{aligned} &\theta_t = t, \ \alpha_t = \tfrac{t-1}{t}, \ \beta_t = \tfrac{3}{t+2}, \ \eta_t = \tfrac{\eta}{t}, \ t = 1, ..., N \\ &\eta = \max\left\{30L, \tfrac{30\Omega(N+2)\mathcal{L}}{m^k}, \sqrt{\tfrac{20(N+1)^3\sigma_*^2}{6m^k R_k^2}}\right\}, \\ &m^k = \max\left\{1, \left\lceil \tfrac{18\Omega\mathcal{L}(N+2)}{L} \right\rceil, \left\lceil \tfrac{15N(N+2)^2\sigma_*^2}{2L^2R_k^2} \right\rceil\right\}, \ t = 0, ..., N. \end{aligned} \tag{4.5.2}$$

(b) Set $y^k = x_N$, where x_N is the SGE solution obtained in Step (a).
end for

Algorithm 5 has a simple structure. Each stage k of the routine, consists of $N_k = N$ iterations of the SGE with initial condition $x_0 = y^{k-1}$ being the approximate solution at the end of the stage $k-1$. Method parameters are selected in such a way that the upper bound R_k^2 for the expected squared distance $\mathbf{E}\|y^k - x^*\|^2$ between the approximate solution y^k at the end of the k-th stage and the optimal solutions reduces by factor 2.

Corollary 4.5.1 *Let $\{y^K\}$ be approximate solution by Algorithm 5 after $K \geq 1$ stages. Assume that $\|y^0 - x^*\|^2 \leq R_0^2$. Then*

$$\mathbf{E}[f(y^K) - f^*] \leq \mu R_0^2 \cdot 2^{-K-1} \quad \textit{and} \quad \mathbf{E}[\|y^K - x^*\|^2] \leq R_0^2 \cdot 2^{-K}.$$

Remarks. By Corollary 4.5.1, the number of stages of Algorithm 5 required to attain the expected inaccuracy ϵ is bounded with $\mathcal{O}\left(\ln(\mu R_0^2/\epsilon)\right)$. When recalling what is the total number N of iterations at each stage, we conclude that the "total" iteration complexity of the method is $\mathcal{O}\left(\sqrt{L\Omega/\mu} \cdot \ln(\mu R_0^2/\epsilon)\right)$. Consequently, the corresponding sample complexity $\sum_{k=1}^K \sum_{t=1}^N m^k$ is order of

$$\mathcal{O}\left\{\sqrt{\frac{L\Omega}{\mu}} \ln\left(\frac{\mu R_0^2}{\epsilon}\right) + \frac{\mathcal{L}\Omega^2}{\mu} \ln\left(\frac{\mu R_0^2}{\epsilon}\right) + \frac{\Omega^2\sigma_*^2}{\mu\epsilon}\right\}.$$

Similarly, the iteration complexity of solution y^K satisfying $\mathbb{E}[\|y^K - x^*\|^2] \leq \epsilon^2$ does not exceed

$$\mathcal{O}\left(\sqrt{\frac{L\Omega}{\mu}} \ln\left(\frac{R_0}{\epsilon}\right)\right);$$

the corresponding sample complexity is then bounded by

$$\mathcal{O}\left\{\sqrt{\frac{L\Omega}{\mu}}\ln\left(\frac{R_0}{\epsilon}\right)+\frac{\mathcal{L}\Omega^2}{\mu}\ln\left(\frac{R_0}{\epsilon}\right)+\frac{\Omega^2\sigma_*^2}{\mu^2\epsilon^2}\right\}.$$

Similar to Corollary 4.4.1, the three terms in the sample complexity bounds represent the deterministic error, the state-dependent stochastic error, and the state-independent stochastic error, respectively. Regardless of the dependence on Ω, the multi-stage SGE achieves the optimal iteration complexity and sample complexity simultaneously, supported by the lower bound in Theorem 5 of [138].

4.6 SGE for sparse recovery

An interesting application of stochastic optimization with state-dependent noise arises in the relation to the problem of sparse recovery when it is assumed that (4.1.1) has a sparse or low-rank solution x^*. This problem is motivated by applications in high-dimensional statistics, where stochastic estimation under sparsity or low-rank constraints has garnered significant attention. To solve this problem, one usually builds a sample average approximation (SAA) of the expected risk and solves the resulting minimization problem. To enhance sparsity, a classical approach consists of incorporating a regularization term, e.g., ℓ_1- or trace-norm penalty (as in Lasso), and minimizing the norm of the solution under constraint (as in Dantzig Selector), see, e.g., [65, 67, 75, 83, 94, 95, 147–150] among many others.

Stochastic approximation also serves as a standard approach to deal with the sparse recovery problem. However, utilizing traditional Euclidean stochastic approximation usually leads to sub-optimal complexity bounds: in this setting, the expected squared ℓ_2-error of the stochastic operator $\mathcal{G}(x,\xi)$ is usually proportional to the problem dimension n. Therefore, non-Euclidean stochastic mirror descent (SMD) methods have been applied to address this issue. In particular, the SMD algorithm in [87, 105] attains the high probability complexity bound $f(\widehat{x})-f^*=\mathcal{O}(\sigma\sqrt{s/N})$ (up to some "logarithmic factors") in recovering an s-sparse signal x^* under the sub-Gaussian noise assumption with subgaussianity parameter σ^2, referred to as "slow rates" for sparse recovery. To improve the convergence rate to $\mathcal{O}(\sigma^2 s/N)$, multi-stage routines exploiting the properties similar to strong/uniform convexity could be used, cf. [25, 37, 48]. In [88, 106], the authors utilize the "restricted" strong convexity condition to establish $\mathcal{O}(\sigma^2 s/N)$ complexity bounds when assuming that $N\gg s^2$. The latter assumption means that the optimal rates are only valid in the range $s\ll\sqrt{N}$ of sparsity parameter. Recently, in [1, 59], new multi-stage stochastic mirror descent algorithms were proposed which rely on the idea of variance reduction. The proposed method improved the required number of iterations in each stage to $\mathcal{O}(s)$, thus attaining the best-known state-dependent stochastic error. However, the iteration complexity of those routines is still sub-optimal in the mini-batch setting, leaving room for further acceleration.

In this section, we suppose that stochastic optimization problem (4.1.1) admits a sparse solution $x^*\in X$. Standard examples of the sparsity assumptions are as follows:

- "Vanilla" sparsity: we assume that an optimal solution $x^*\in X$ has at most $s\ll n$ nonvanishing entries. We put $\|\cdot\|=\|\cdot\|_1$ and $\|\cdot\|_*=\|\cdot\|_\infty$.
- Group sparsity: let us partition the set $[n]$ into b subsets $\{I_1,...,I_B\}$, and let x_b the bth block of x, meaning that $[x_b]_i=0$ for all $i\notin I_b$. We assume that the optimal $x^*\in X$ is a block vector with at most $s\le B$ nonvanishing blocks x_b. We define $\|x\|=\sum_{b=1}^B\|x_b\|_2$ (block ℓ_1/ℓ_2-norm) and $\|x\|_*=\max_{b\le B}\|x_b\|_2$ (block ℓ_∞/ℓ_2-norm).

- Low rank sparsity: consider the matrix space $\mathbf{R}^{p\times q}$ where $p \geq q$ equipped with Frobenius inner product. We assume that the optimal $x^* \in X$ satisfies $\mathrm{rank}(x^*) \leq s$. We consider the nuclear norm $\|x\| = \sum_{i=1}^q \sigma_i(x)$ where $\sigma_i(\cdot)$ are the singular values of x, so that $\|x\|_* = \max_{i\in[q]} \sigma_i(x)$ is the spectral norm.

In what follows we assume that problem (4.1.1) and norms $\|\cdot\|$ and $\|\cdot\|_*$ satisfy conditions (4.1.2), (4.1.3), and state-dependent noise conditions in Assumptions (SN) and (LP) of Section 4.2. Also, instead of μ-quadratic growth (with $\mu > 0$) condition with respect to $\|\cdot\|$ as in the previous section, we suppose f and x^* verify the quadratic growth condition with respect to the Euclidean norm, i.e.,

$$f(x) - f^* \geq \tfrac{1}{2}\underline{\kappa}\|x - x^*\|_2^2, \ \forall x \in X, \tag{4.6.1}$$

for some $\underline{\kappa} > 0$.

4.6.1 Application: sparse generalized linear regression

Let us check that problem assumptions of this section hold in the case of generalized linear regression problem (GLR), as described in the introduction. Let us assume that

- regressors ϕ_t satisfy $\mathbb{E}[\phi_1\phi_1^T] = \Sigma \succeq \kappa I$ with $\kappa > 0$ and $\|\Sigma\|_\infty := \max_{i\in[n],j\in[n]} \Sigma_{i,j} \leq \nu$;
- noises ζ_t are zero mean with bounded variance, i.e., $\mathbb{E}[\zeta_1^2] \leq 1$ without loss of generality;
- activation function u is strongly monotone and Lipschitz continuous, i.e., for some $\bar{r} \geq \underline{r} \geq 0$,

$$\big(u(t) - u(t')\big)\cdot(t - t') \geq \underline{r}(t - t')^2, \quad \text{and} \quad |u(t) - u(t')| \leq \bar{r}|t - t'|, \quad \forall\, t, t' \in \mathbf{R}. \tag{4.6.2}$$

As already explained in the introduction, estimation of $x^* \in \operatorname{int} X$ may be addressed through solving the stochastic optimization problem

$$\min_{x\in X} \big\{ f(x) := \mathbf{E}[v(\phi^T x) - \phi^T x\eta] \big\} \tag{4.6.3}$$

where $v'(t) = u(t)$. The gradient of the problem objective and its stochastic estimate are given by

$$g(x) = \mathbf{E}[\phi\big(u(\phi^T x) - \eta\big)] \quad \text{and} \quad \mathcal{G}(x, (\phi, \zeta)) := \phi\big(u(\phi^T x) - \eta\big) = \phi\big(u(\phi^T x) - u(\phi^T x^*)\big) - \phi\zeta.$$

It is easy to see that condition (4.1.3) is verified in this case, and invoking $\mathbf{E}[\eta] = \mathbf{E}[u(\phi^T x)]$, we conclude that $g(x^*) = 0$. To check the quadratic growth condition (4.6.1), we write

$$\begin{aligned} f(x) - f^* &= \int_0^1 g(x^* + t(x - x^*))^T(x - x^*)dt \\ &= \int_0^1 \mathbf{E}\{\phi[u\big(\phi^T(x^* + t(x - x^*))\big) - u\big(\phi^T x^*\big)]\}^T(x - x^*)dt \\ \text{[by (4.6.2)]} \quad &\geq \int_0^1 \underline{r}\mathbf{E}\{[\phi^T(x - x^*)]^2\}tdt = \tfrac{\underline{r}}{2}\|x - x^*\|_\Sigma^2 \geq \tfrac{\underline{r}\kappa}{2}\|x - x^*\|_2^2. \end{aligned} \tag{4.6.4}$$

Therefore, condition (4.6.1) holds with $\underline{\kappa} = \underline{r}\cdot\kappa$ which is independent of problem dimension n.

Since we are interested in the high-dimensional setting, the desired recovery error should have, at most, logarithmic dependence in the problem dimension n. However, the ℓ_2 variance of the stochastic first-order information $\mathbf{E}\|\mathcal{G} - g\|_2^2$ is proportional to the problem dimension n, making the standard Euclidean SA methods not applicable. To address this issue, we work in the non-Euclidean setting

with $\|\cdot\| = \|\cdot\|_1$ and $\|\cdot\|_* = \|\cdot\|_\infty$. Next, let us examine the smoothness condition (4.1.2) and state-dependent variance condition (SN). Note that for all $x, x' \in \mathbf{R}^n$,

$$\begin{aligned}\|g(x) - g(x')\|_\infty &= \sup_{\|z\|_1\le 1} \langle g(x) - g(x'), z\rangle = \sup_{\|z\|_1\le 1} \mathbf{E}\{\phi^T z[u(\phi^T x) - u(\phi^T x')]\} \\ &\overset{(i)}{\le} \sup_{\|z\|_1\le 1} \bar{r}\mathbf{E}\{|\phi^T z||\phi^T(x - x')|\} \overset{(ii)}{\le} \bar{r} \sup_{\|z\|_1\le 1} \sqrt{\mathbf{E}\{(\phi^T z)^2\}} \cdot \|x - x'\|_\Sigma \\ &\le \bar{r}\sqrt{\nu}\|x - x'\|_\Sigma, \end{aligned} \tag{4.6.5}$$

where (i) is a consequence of (4.6.2) and (ii) follows from the Cauchy inequality. Consequently, we have

$$\begin{aligned}\mathbf{E}\|\mathcal{G}(x,(\phi,\eta)) - g(x)\|_\infty^2 &= \mathbf{E}\|\phi\big(u(\phi^T x) - u(\phi^T x^*)\big) + \phi\zeta + [g(x^*) - g(x)]\|_\infty^2 \\ [\text{by (5.7.45)}] &\le 3\mathbf{E}\|\phi\big(u(\phi^T x) - u(\phi^T x^*)\big)\|_\infty^2 + 3\mathbf{E}\{\|\phi\|_\infty^2\}\sigma^2 + 3\bar{r}^2\nu\|x - x^*\|_\Sigma^2 \\ [\text{by (4.6.2)}] &\le 3\bar{r}^2\mathbf{E}\{\|\phi\|_\infty^2\big(\phi^T x - \phi^T x^*\big)^2\} + 3\mathbf{E}\{\|\phi\|_\infty^2\}\sigma^2 + 3\bar{r}^2\nu\|x - x^*\|_\Sigma^2.\end{aligned}$$

By (4.6.4), we conclude that the condition (SN) holds whenever

$$\mathbf{E}[\|\phi\|_\infty^2\big(\phi^T x - \phi^T x^*\big)^2] \lesssim \mathbf{E}[\big(\phi^T(x - x^*)\big)^2] \lesssim \|x - x^*\|_\Sigma^2;$$

e.g., when the regressor ϕ is bounded or sub-Gaussian. Finally, under similar assumptions on the regressors and sub-Gaussian assumption on the additive noise ζ, condition (LP) naturally follows.

4.6.2 SGE-SR: stochastic gradient extrapolation for sparse recovery

We extend SGE to solve the sparse recovery problem. Similarly to Algorithm 5, sparse recovery routine is organized in stages; each stage represents a run of SGE (Algorithm 4). The principal difference with Algorithm 5, apart from the different choice of algorithm parameters, is the sparsity enforcing step (see, e.g., [84–86]) implemented at the end of each stage.

Observe that for $x \in X$ one can efficiently compute a sparse approximation of x, specifically, $x_s = \mathsf{sparse}(x)$, an optimal solution to

$$\min \|x - z\|_2 \text{ over } s\text{-sparse } z \in X. \tag{4.6.6}$$

For instance, in the "vanilla sparsity" case, when the set X is positive monotone,[4] x_s is obtained by zeroing all but s largest in amplitude entries of x.[5]

[4] For $x \in \mathbf{R}^n$, let $|x|$ denote a vector in $\mathbf{R}^n_+$ whose entries are absolute values of the corresponding entries of x. We say that X is positive monotone if whenever $x \in X$ and $|y| \le |x|$ (the inequality is understood coordinate-wise), one also has $y \in X$. A typical example of a monotone convex set X is a ball of an absolute norm in $\mathbf{R}^n$.

[5] In the block sparsity case, when X is positive block-monotone, the corresponding "sparsification" amounts to zeroing out all but s largest (in ℓ_2-norm) blocks of x; when X is a ball of a Schatten norm in the space of $p \times q$ real matrices, low rank x_s may be obtained from x by trimming the singular values of x.

Algorithm 6 Stochastic Gradient Extrapolation method for Sparse Recovery (SGE-SR)

Input: initial point $\overline{y}^0 \in X$.
for $k = 1, 2, \ldots, K$ **do**
(a) Set $N = 40\sqrt{sL\Omega/\underline{\kappa}}$ and $R_k = 2^{-k/2}R_0$. Run N iterations of SGE (Algorithm 4) with $x_0 = z_0 = \overline{y}^{k-1}$ and

$$\theta_t = t,\ \alpha_t = \tfrac{t-1}{t},\ \beta_t = \tfrac{3}{t+2},\ \eta_t = \tfrac{\eta}{t},\ t = 1, ..., N$$
$$\eta = \max\left\{30L, \tfrac{30\Omega(N+2)\mathcal{L}}{m^k}, \sqrt{\tfrac{20(N+1)^3\sigma_*^2}{6m^k R_k^2}}\right\},$$
$$m^k = \max\left\{1, \left\lceil\tfrac{18\Omega\mathcal{L}(N+2)}{L}\right\rceil, \left\lceil\tfrac{15N(N+2)^2\sigma_*^2}{2L^2R_k^2}\right\rceil\right\},\ t = 0, ..., N$$

(b) Set $y^k = x_N$, where x_N is the solution obtained in Step (a). Calculate

$$\overline{y}^k = \mathsf{sparse}(y^k).$$

end for

The following corollary characterizes the convergence rate of SGE-SR for solving the sparse recovery problem.

Corollary 4.6.1 *Let $\{y^k, \overline{y}^k\}$ be computed by Algorithm 6. Assume $\|\overline{y}^0 - x^*\|^2 \le R_0^2$. Then we have for $k \ge 1$*

$$\mathbf{E}[f(y^k) - f^*] \le \underline{\kappa}s^{-1}R_0^2 \cdot 2^{-k+4} \quad \textit{and} \quad \mathbf{E}[\|\overline{y}^k - x^*\|^2] \le R_0^2 \cdot 2^{-k}.$$

Remarks. From the result of Corollary 4.6.1 we conclude that the SGE-SR algorithm finds an s-sparse $\overline{y}^k \in X$ such that $\mathbf{E}[\|\overline{y}^k - x^*\|^2] \le \epsilon^2$ for any $\epsilon \in (0, R_0)$ in at most $k = \mathcal{O}\left(\ln(R_0/\epsilon)\right)$ stages. The corresponding iteration complexity of the SGE-SR is $\mathcal{O}\left(\sqrt{\frac{sL\Omega}{\underline{\kappa}}}\ln(\frac{R_0}{\epsilon})\right)$, and the overall sample complexity is

$$\mathcal{O}\left\{\sqrt{\frac{sL\Omega}{\underline{\kappa}}}\ln\left(\frac{R_0}{\epsilon}\right) + \frac{s\mathcal{L}\Omega^2}{\underline{\kappa}}\ln\left(\frac{R_0}{\epsilon}\right) + \frac{\Omega^2 s^2\sigma_*^2}{\underline{\kappa}^2\epsilon^2}\right\}.$$

Similarly, the iteration complexity of the solution $y^k \in X$ (which is not s-sparse in general) such that $\mathbf{E}[f(y^k) - f^*] \le \epsilon$ is $\mathcal{O}\left(\sqrt{\frac{sL\Omega}{\underline{\kappa}}}\ln(\frac{\underline{\kappa}R_0^2}{s\epsilon})\right)$, while the total sample complexity is

$$\mathcal{O}\left\{\sqrt{\frac{sL\Omega}{\underline{\kappa}}}\ln\left(\frac{\underline{\kappa}R_0^2}{s\epsilon}\right) + \frac{s\mathcal{L}\Omega^2}{\underline{\kappa}}\ln\left(\frac{\underline{\kappa}R_0^2}{s\epsilon}\right) + \frac{\Omega^2 s\sigma_*^2}{\underline{\kappa}\epsilon}\right\}.$$

The above results may be compared to the convergence guarantees obtained in [59] in the similar setting of the sparse recovery problem. The iteration complexity of the SGE-SR algorithm attains the optimal dependence on the problem's condition number, $\sqrt{sL/\underline{\kappa}}$, which improves over the corresponding result in [59] by a factor of $\mathcal{O}(\sqrt{sL/\underline{\kappa}})$. Moreover, except for the dependence on Ω, the proposed solution matches the best-known sample complexity bounds for the stochastic error; the extra factor Ω being due to the mini-batch use (cf. Lemma 4.2.1 and subsequent remark) and could theoretically hinder the method precision. Note that in the problems of interest, Ω is logarithmic in the problem dimension. It should be mentioned that in our numerical experiments we did not observe any accuracy degradation when using the mini-batch algorithm.

4.7 Numerical experiments

In this section we present a simulation study illustrating numerical performance of the proposed routines. We consider the sparse recovery problem in generalized linear regression (GLR) model with random design as discussed in the previous section. Recall that we are looking to recover the s-sparse vector $x^* \in \mathbf{R}^n$ from i.i.d observations

$$\eta_i = u(\phi_i^T x^*) + \sigma\zeta_i, \ \ i = 1, 2, \ldots, N.$$

In experiments we report on below, the activation function $u(\cdot)$ is of the form

$$u_\alpha(x) = x\mathbf{1}\{|x| \le \alpha\} + \mathrm{sign}(x)[\alpha^{-1}(|x|^\alpha - 1) + 1]\mathbf{1}\{|x| > \alpha\}, \quad \alpha > 0, \ x \in \mathbf{R}.$$

We consider three different activations, namely, the linear link function $u_1(\cdot)$, $u_{1/2}(\cdot)$, and $u_{1/10}(\cdot)$ (cf. Figure 4.2). In our simulations, s nonvanishing components of the signal x^* are sampled

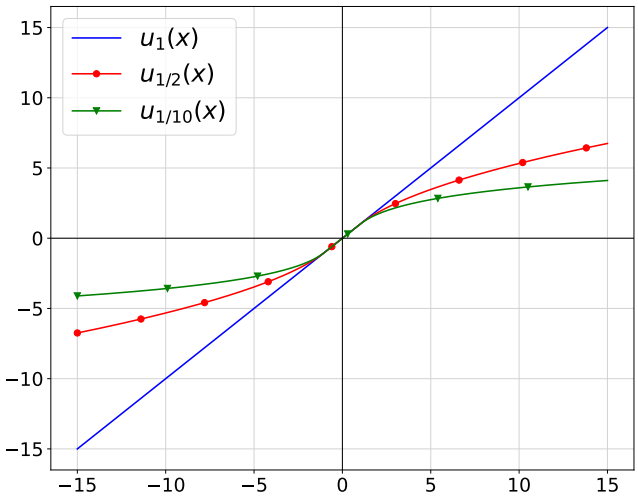


Figure 4.2: Activation functions

from the s-dimensional standard Gaussian distribution. We explore two setups—light tailed and heavy tailed—for generating regressors and additive noises. In the light-tail setup, regressors ϕ_i are independently drawn from a multivariate Gaussian distribution $\phi_i \sim \mathcal{N}(0, \Sigma)$, where Σ is a diagonal covariance matrix with diagonal entries $0 < \Sigma_{1,1} \le \cdots \le \Sigma_{n,n}$. In the heavy-tail setup, regressors are independently drawn from a multivariate Student distribution $\phi_i \sim t_n(\nu, 0, \Sigma)$, ν being the corresponding degree of freedom [151]. The condition number κ of the problem is defined as the ratio of the largest and the smallest eigenvalues of Σ. The additive noise of the model in the light-tail setup is the zero-mean Gaussian noise with variance σ^2; in the heavy-tail setup, the additive noises have (scaled) univariate Student distribution $\eta_i \sim \lambda t(\nu)$, $\nu \ge 3$, with scale parameter $\lambda = \sqrt{(\nu-2)/\nu}$ with unit variance. Because of the memory limitations, observations (η_i, ϕ_i) are generated on the fly at each oracle call.

In all our experiments, we run 50 simulation trials (with randomly generated regressors and noises); then we trace in the plots the median and the first and the last deciles of the error $\|x_t - x^*\|_2$.

The aim of the first series of experiments is to compare the procedure described in Section 4.6 to the SMD-SR algorithm of [59];[6] in the light-tail noise setting, the corresponding results are presented in Figure 4.3. In Figure 4.4, we present results of simulations of the accelerated algorithm in the light-tail and heavy-tail setup. We used the same algorithmic parameters in both simulation setups. In the above experiments, $n = 500\,000$, the maximal number of calls to the stochastic oracle (estimation sample size) $N = 250\,000$, and sparsity level $s = 250$; unless stated otherwise, the problem condition number is set to $\kappa = 1$.

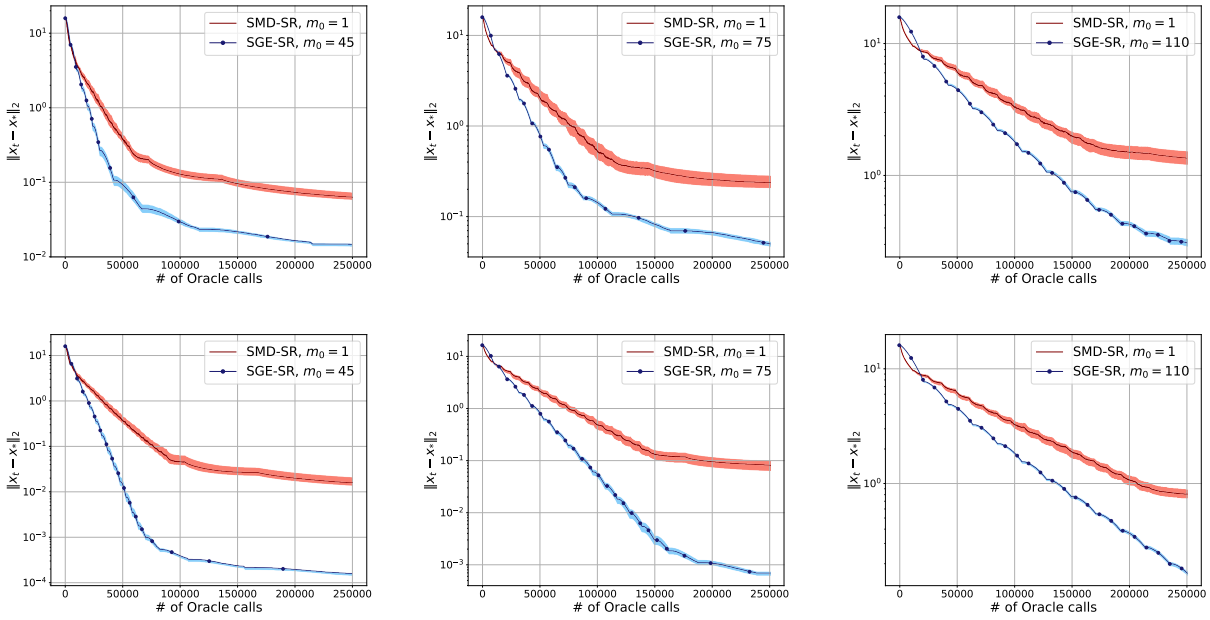


Figure 4.3: Estimation error $\|x_t - x^*\|_2$ against the number of stochastic oracle calls for SGE-SR and SMD-SR algorithms. In the left, middle, and right columns of the plot we show results for the linear activation u_1, and nonlinear $u_{1/2}$ and $u_{1/10}$, respectively. Two figure rows correspond to two different noise levels, $\sigma = 0.1$ (the upper row) and $\sigma = 0.001$ (the bottom row). The legend specifies the value m_0 of the batch size of the preliminary phase of the algorithm for both routines.

In the second series of experiments we put $n = 100\,000$, $N = 200\,000$, and $s = 50$. Experiments reported in Figure 4.5 illustrate the impact of the condition number on the convergence of the SGE-SR algorithm.

Finally, we illustrate the performance of SGE-SR and SMD-SR algorithms which share the same size of mini-batch. Recall that both algorithms if "normally set"—SGE-SR with mini-batch of optimal size and SMD-SR with "trivial" mini-batch (of size $m_0 = 1$)—converge linearly during the preliminary phase. In Figure 4.6 we report on the simulation of algorithms with the same (optimal for the accelerated algorithm) size of the mini-batch [59].

The series of experiments conducted indicates that the SGE-SR algorithm outperforms its non-accelerated counterpart. Despite both algorithms exhibiting linear rates of convergence, SGE-SR reaches a better precision for a fixed number of samples in every setting. This advantage is also observed when the two algorithms are compared in terms of the number of iterations, where the accelerated algorithm clearly outperforms its non-accelerated counterpart by a significant margin.

[6] SMD-SR is a stochastic approximation algorithm for sparse recovery utilizing hard thresholding which relies upon "vanilla" non-Euclidean mirror descent; both algorithms use the same distance generating function $\omega(x) = c(n)\|x\|_p^2$.

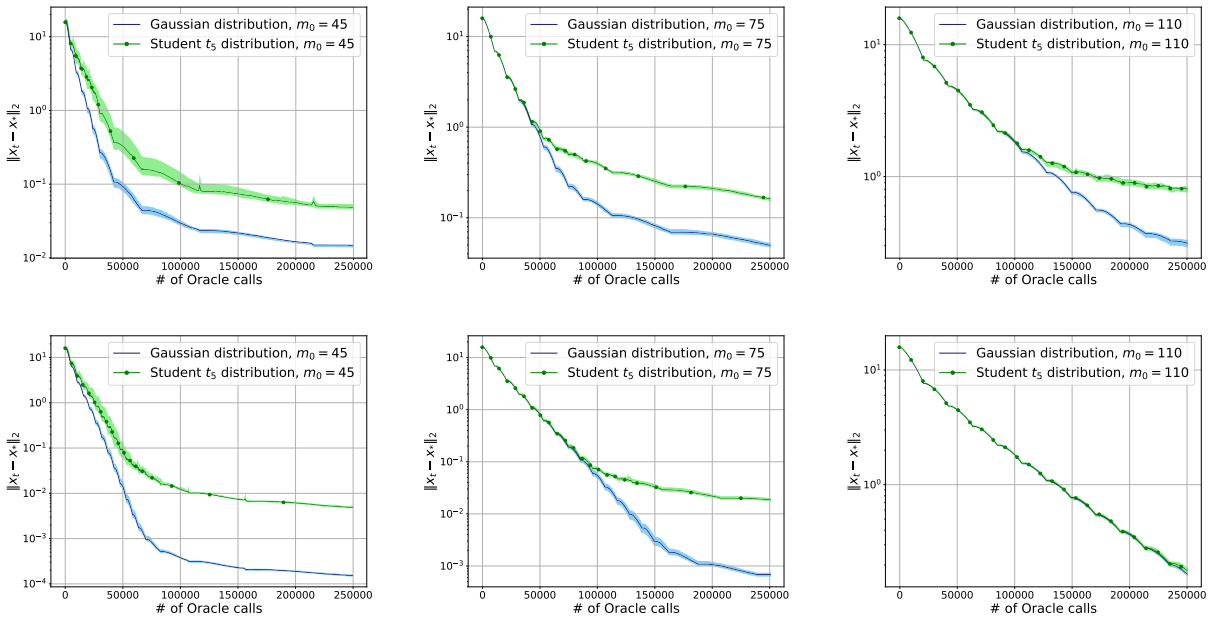


Figure 4.4: Estimation error $\|x_t - x^*\|_2$ against the number of stochastic oracle calls for SGE-SR in Gaussian (light-tail) and Student t_5 (heavy-tail) regressor and noise generation setups. In the left, middle, and right columns of the plot we show results for the linear activation u_1, and nonlinear $u_{1/2}$ and $u_{1/10}$, respectively. Two figure rows correspond to two different noise levels, $\sigma = 0.1$ (the upper row) and $\sigma = 0.001$ (the bottom row). The legend specifies the value m_0 of the batch size of the preliminary phase of the algorithm for both routines.

The improved iteration complexities make SGE-SR a viable solution for distributed sparse recovery problems, where it is crucial to reduce the number of communication rounds between the servers and the clients while keeping high precision.

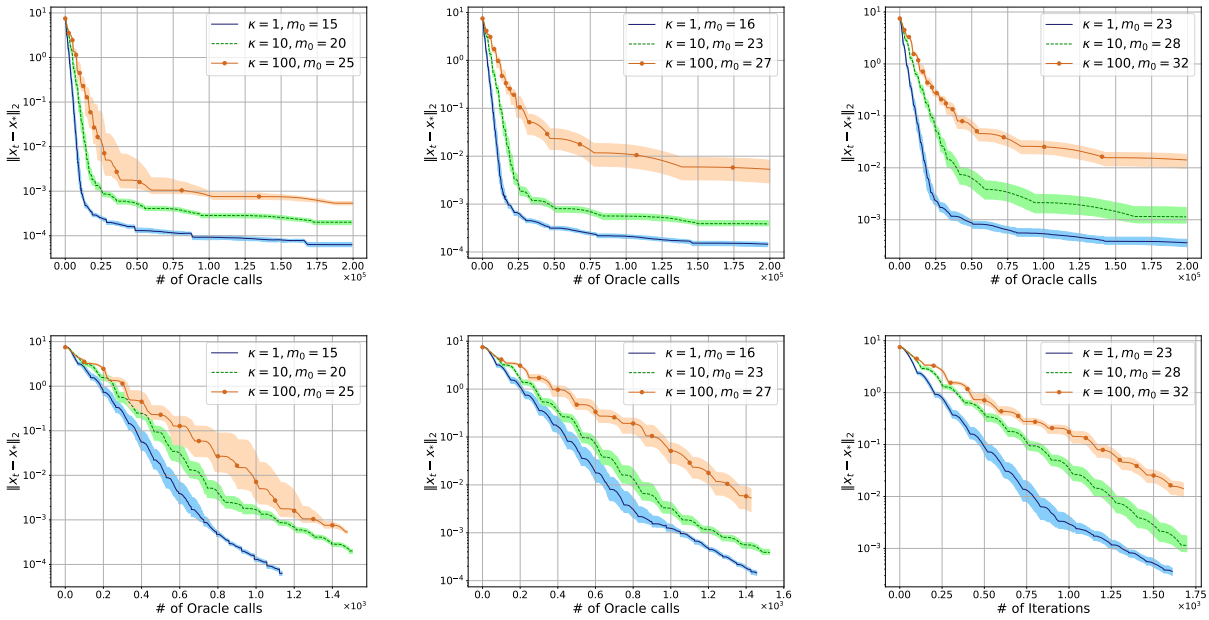


Figure 4.5: Error $\|x_t - x^*\|_2$ of the SGE-SR algorithm for three values of the problem condition number. First row: algorithm error against the number of oracle calls; second row: error against the number of algorithm iterations. Figure columns correspond to the results for u_1, $u_{1/2}$, and $u_{1/10}$ activation functions and $\sigma = 0.001$.

4.8 Concluding remarks

In this paper, we investigate the problem of stochastic smooth convex optimization with "state-dependent" variance of stochastic gradients. We study two non-Euclidean accelerated stochastic approximation algorithms, stochastic accelerated gradient descent (SAGD) and stochastic gradient extrapolation (SGE), and provide optimal iteration and sample complexities for both algorithms under appropriate conditions. However, the optimal convergence guarantees for SGE require less restrictive assumptions, thus leading to wider applications such as statistical estimation problems with heavy tail noises. In addition, we propose a multistage routine of SGE to solve problems that satisfy the quadratic growth condition and further extend it to the sparse recovery problem. Our theoretical guarantees are corroborated by numerical experiments in high-dimensional settings. Further research will be directed to proving large deviation bounds to ensure the reliability and robustness of the solutions.

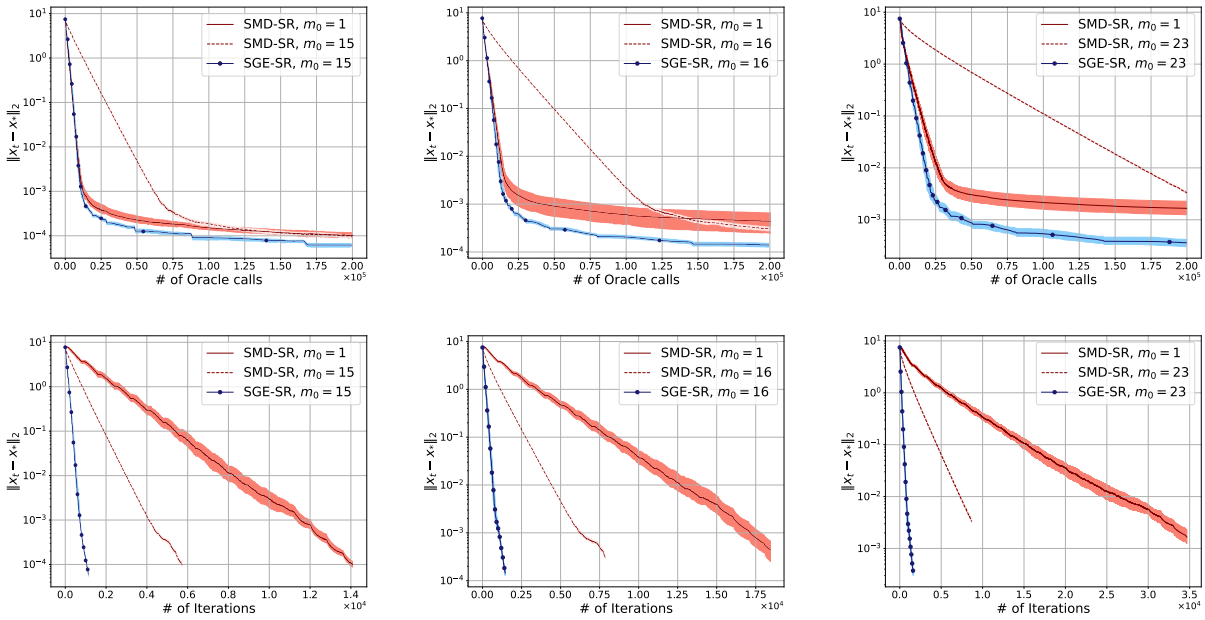


Figure 4.6: SGE-SR compared to "vanilla" SMD-SR and its mini-batch variant. First row: error $\|x_t - x^*\|_2$ against the number of oracle calls; second row: the error against the number of algorithm iterations. Figure columns correspond to the results for u_1, $u_{1/2}$, and $u_{1/10}$ activation functions and $\sigma = 0.001$.

4.9 Appendix: proofs

4.9.1 Proof of Lemmas 4.2.1 and 4.3.1

Proof of Lemma 4.2.1. It is easy to see that (4.2.1) and (4.2.2) hold when $m_t = 1$, so we consider the case when $m_t \geq 2$. For notational simplicity, put

$$\delta_t := G_t(u_t) - g(u_t), \quad \delta_{t,i} := \mathcal{G}_t(u_t, \xi_{t,i}) - g(u_t).$$

Let ω^* denote the convex conjugate of the distance generating function ω. We write

$$\|\delta_t\|_*^2 = \Omega^2 \|\tfrac{\delta_t}{\Omega}\|_*^2 = 2\Omega^2 \max_z [\langle \tfrac{\delta_t}{\Omega}, z\rangle - \tfrac{1}{2}\|z\|^2] \underset{\text{by (4.1.10)}}{\leq} 2\Omega \max_z [\langle \delta_t, z\rangle - \omega(z)] = 2\Omega\, \omega^*\big(\delta_t\big). \quad (4.9.1)$$

By the strong convexity of ω, ω^* is smooth with 1-Lipschitz continuous gradient. Thus

$$\omega^*\big(\delta_t\big) \leq \omega^*\left(\frac{1}{m_t}\sum_{i=1}^{m_t-1}\delta_{t,i}\right) + \left\langle \nabla\omega^*\left(\frac{1}{m_t}\sum_{i=1}^{m_t-1}\delta_{t,i}\right), \frac{\delta_{t,m_t}}{m_t}\right\rangle + \frac{1}{2}\left\|\frac{\delta_{t,m_t}}{m_t}\right\|_*^2.$$

Now, recursively using the above relationship and the independence of $\delta_{t,i},\ i \in [m_t]$, we get

$$\mathbf{E}_{\lceil t-1\rceil}[\omega^*\big(\delta_t\big)] \leq \frac{1}{2m_t} \cdot \mathbf{E}\|\delta_{t,1}\|_*^2, \quad (4.9.2)$$

and we obtain the bound in (4.2.2) by combing the above inequality with (4.9.1). □

Proof of Lemma 4.3.1. Note that x_{t-1} and y_t are $\mathcal{F}_{t-1}$-measurable. By Lemma 4.2.1,

$$\begin{aligned}
&\mathbf{E}_{\lceil t-1\rceil}\left[\|G_t - g(y_t)\|_*^2\right] \le \frac{\overline{\Omega}}{m_t}\mathbf{E}_{\lceil t-1\rceil}\left[\|\mathcal{G}(y_t,\xi_{t,1}) - g(y_t)\|_*^2\right] \\
&\le \frac{3\overline{\Omega}}{m_t}\Big\{\mathbf{E}_{\lceil t-1\rceil}\left[\|\mathcal{G}(y_t,\xi_{t,1}) - \mathcal{G}(x_{t-1},\xi_{t,1})\|_*^2\right] + \mathbf{E}_{\lceil t-1\rceil}\left[\|\mathcal{G}(x_{t-1},\xi_{t,1}) - g(x_{t-1})\|_*^2\right] \qquad (4.9.3)\\
&\qquad + \mathbf{E}_{\lceil t-1\rceil}\left[\|g(x_{t-1}) - g(y_t)\|_*^2\right]\Big\} \\
&\le \frac{3\overline{\Omega}}{m_t}\left\{\left(\mathbf{E}_{\xi_{t,1}}[\mathcal{K}(\xi_{t,1})^2] + L^2\right)\cdot\|y_t - x_{t-1}\|^2 + \left[\mathcal{L}[f(x_{t-1}) - f^* - \langle g(x^*), x_{t-1} - x^*\rangle] + \sigma_*^2\right]\right\}, \qquad (4.9.4)
\end{aligned}$$

the last inequality being the consequence of Assumptions (LP) and (SN). Now, utilizing $x_{t-1} - y_t = x_{t-1} - (1-\beta_t)x_{t-1} - \beta_t z_{t-1} = \beta_t(x_{t-1} - z_{t-1})$, we obtain the desired result. □

4.9.2 Proof of Theorem 4.3.1

For sake of completeness, we start with the following statement similar to the previous results for SAGD (e.g., Proposition 3.1 in [129]).

Proposition 4.9.1 *Let $\{z_t\}, \{y_t\}$ and $\{x_t\}$ be generated by Algorithm 3. Suppose that $\{\beta_t\}$ and $\{\eta_t\}$ satisfy (4.3.2) for some $\theta_t \ge 0$. Then for any $x \in X$ one has*

$$\begin{aligned}
&\sum_{t=1}^k \theta_t[f(x_t) - f(x)] + \theta_k\beta_k\eta_k V(z_k, x) \\
&\le \sum_{t=1}^k \theta_t(1-\beta_t)[f(x_{t-1}) - f(x)] + \theta_1\beta_1\eta_1 V(z_0, x) + \sum_{t=1}^k \theta_t\beta_t\langle\delta_t, x - z_{t-1}\rangle + \sum_{t=1}^k \frac{\theta_t\beta_t\|\delta_t\|_*^2}{2(\eta_t - L\beta_t)}
\end{aligned} \qquad (4.9.5)$$

where $\delta_t := G_t - g(y_t)$.

Proof of the proposition. First, by convexity of f and due to the definition of x_t, we have for all $z \in X$,

$$\begin{aligned}
f(z) + \langle g(z), x_t - z\rangle &= (1-\beta_t)[f(z) + \langle g(z), x_{t-1} - z\rangle] + \beta_t[f(z) + \langle g(z), z_t - z\rangle] \\
&\le (1-\beta_t)f(x_{t-1}) + \beta_t[f(z) + \langle g(z), z_t - z\rangle].
\end{aligned}$$

Now, by the smoothness of f,

$$\begin{aligned}
f(x_t) &\le f(z) + \langle g(z), x_t - z\rangle + \tfrac{L}{2}\|x_t - z\|^2 \\
&\le (1-\beta_t)f(x_{t-1}) + \beta_t[f(z) + \langle g(z), z_t - z\rangle] + \tfrac{L}{2}\|x_t - z\|^2,
\end{aligned}$$

so that for $z = y_t$ we have

$$f(x_t) \le (1-\beta_t)f(x_{t-1}) + \beta_t[f(y_t) + \langle g(y_t), z_t - y_t\rangle] + \tfrac{L}{2}\|x_t - y_t\|^2. \qquad (4.9.6)$$

On the other hand, the optimality condition for (4.3.1b) in Algorithm 3 implies the following relationship (see, e.g., Lemma 3.5 of [129]):

$$\langle G_t, z_t - y_t\rangle + \eta_t V(z_{t-1}, z_t) + \eta_t V(z_t, x) \le \langle G_t, x - y_t\rangle + \eta_t V(z_{t-1}, x), \quad \forall x \in X. \qquad (4.9.7)$$

By combining (4.9.6) and (4.9.7) we obtain for all $x \in X$,

$$\begin{aligned}
&f(x_t) + \beta_t\eta_t V(z_{t-1}, z_t) + \beta_t\eta_t V(z_t, x) \\
&\quad \le (1-\beta_t)f(x_{t-1}) + \beta_t[f(y_t) + \langle g(y_t), x - y_t\rangle] \\
&\qquad + \beta_t\langle \delta_t, x - z_t\rangle + \beta_t\eta_t V(z_{t-1}, x) + \tfrac{L}{2}\|x_t - y_t\|^2 \\
&\quad \le (1-\beta_t)f(x_{t-1}) + \beta_t[f(y_t) + \langle g(y_t), x - y_t\rangle] \\
&\qquad + \beta_t\langle \delta_t, z_{t-1} - z_t\rangle + \beta_t\langle \delta_t, x - z_{t-1}\rangle + \beta_t\eta_t V(z_{t-1}, x) + \tfrac{L}{2}\|x_t - y_t\|^2. \qquad (4.9.8)
\end{aligned}$$

From (4.3.1a) and (4.3.1c) we have $x_t - y_t = \beta_t(z_t - z_{t-1})$; after substituting into (4.9.8) and taking into account (4.1.11), we obtain

$$\begin{aligned}
&f(x_t) + \beta_t\eta_t V(z_t, x) \\
&\quad \le (1-\beta_t)f(x_{t-1}) + \beta_t[f(y_t) + \langle g(y_t), x - y_t\rangle] \\
&\qquad + \beta_t\langle \delta_t, x - z_{t-1}\rangle + \beta_t\eta_t V(z_{t-1}, x) + \beta_t\|\delta_t\|_*\|z_{t-1} - z_t\| + (\tfrac{L\beta_t^2}{2} - \tfrac{\eta_t\beta_t}{2})\|z_t - z_{t-1}\|^2 \\
&\quad \le (1-\beta_t)f(x_{t-1}) + \beta_t[f(y_t) + \langle g(y_t), x - y_t\rangle] + \beta_t\langle \delta_t, x - z_{t-1}\rangle + \beta_t\eta_t V(z_{t-1}, x) + \frac{\beta_t\|\delta_t\|_*^2}{2(\eta_t - L\beta_t)},
\end{aligned}$$

the last inequality being a consequence of (4.3.1c) and Young's inequality. When subtracting $f(x)$ on both sides we get

$$\begin{aligned}
&[f(x_t) - f(x)] + \beta_t\eta_t V(z_t, x) \\
&\le (1-\beta_t)[f(x_{t-1}) - f(x)] + \beta_t\underbrace{[f(y_t) + \langle g(y_t), x - y_t\rangle - f(x)]}_{\le 0} + \beta_t\eta_t V(z_{t-1}, x) + \beta_t\langle \delta_t, x - z_{t-1}\rangle + \tfrac{\beta_t\|\delta_t\|_*^2}{2(\eta_t - L\beta_t)} \\
&\le (1-\beta_t)[f(x_{t-1}) - f(x)] + \beta_t\eta_t V(z_{t-1}, x) + \beta_t\langle \delta_t, x - z_{t-1}\rangle + \frac{\beta_t\|\delta_t\|_*^2}{2(\eta_t - L\beta_t)},
\end{aligned}$$

and after multiplying by θ_t and summing up from $t = 1$ to k we arrive at

$$\begin{aligned}
&\sum_{t=1}^k \theta_t[f(x_t) - f(x)] + \theta_k\beta_k\eta_k V(z_k, x) \\
&\qquad \le \sum_{t=1}^k \theta_t(1-\beta_t)[f(x_{t-1}) - f(x)] + \theta_1\beta_1\eta_1 V(z_0, x) + \sum_{t=1}^k \theta_t\beta_t\langle \delta_t, x - z_{t-1}\rangle + \sum_{t=1}^k \frac{\theta_t\beta_t\|\delta_t\|_*^2}{2(\eta_t - L\beta_t)},
\end{aligned}$$

which is (4.9.5). □

Proof of the theorem. When setting $x = x^*$, using (4.2.2) in Lemma 4.2.1, and taking expectation on both sides of (4.9.5), we get

$$\begin{aligned}
&\sum_{t=1}^k \theta_t\mathbf{E}[f(x_t) - f^*] + \theta_k\beta_k\eta_k\mathbf{E}[V(z_k, x^*)] \\
&\quad \le \sum_{t=1}^k \theta_t(1-\beta_t)\mathbf{E}[f(x_{t-1}) - f^*] + \theta_1\beta_1\eta_1 V(z_0, x^*) + \sum_{t=1}^k \theta_t r_t\mathbf{E}\big[\mathcal{L}\big(f(y_t) - f^* - \langle g(x^*), y_t - x^*\rangle\big) + \sigma_*^2\big].
\end{aligned}$$

Recall that $y_t = (1-\beta_t)x_{t-1} + \beta_t z_{t-1}$. On the other hand, by smoothness of f, when recalling that $\langle g(x^*), x_{t-1} - x^*\rangle \geq 0$,

$$\begin{aligned}
&f(y_t) - f^* - \langle g(x^*), y_t - x^*\rangle \\
&\quad \leq (1-\beta_t)\big[f(x_{t-1}) - f^* - \langle g(x^*), x_{t-1} - x^*\rangle\big] + \beta_t\big[f(z_{t-1}) - f^* - \langle g(x^*), z_{t-1} - x^*\rangle\big] \\
&\quad \leq (1-\beta_t)\big[f(x_{t-1}) - f^* - \langle g(x^*), x_{t-1} - x^*\rangle\big] + \tfrac{L\beta_t}{2}\|z_{t-1} - x^*\|^2 \\
&\quad \leq (1-\beta_t)\big[f(x_{t-1}) - f^*\big] + L\beta_t V(z_{t-1}, x^*).
\end{aligned}$$

When combining the above inequalities we obtain

$$\begin{aligned}
&\sum_{t=1}^{k} \theta_t \mathbf{E}[f(x_t) - f^*] + \theta_k\beta_k\eta_k\mathbf{E}[V(z_k, x^*)] \\
&\leq \sum_{t=1}^{k} \theta_t(1-\beta_t)(1+r_t\mathcal{L})\mathbf{E}[f(x_{t-1}) - f^*] + \theta_1\beta_1\eta_1 V(z_0, x^*) + \sum_{t=1}^{k}\left\{\theta_t r_t\beta_t L\mathcal{L}\mathbf{E}[V(z_{t-1}, x^*)] + \theta_t r_t\sigma_*^2\right\}.
\end{aligned}$$

Due to (4.3.3), and taking into account that $\beta_1 = 1$, we conclude that

$$\theta_k\mathbf{E}[f(x_k) - f^*] + \theta_k\beta_k\eta_k\mathbf{E}[V(z_k, x^*)] \leq \theta_1\beta_1\eta_1 V(z_0, x^*) + \sum_{t=1}^{k}\theta_t r_t\beta_t L\mathcal{L}\mathbf{E}[V(z_{t-1}, x^*)] + \sum_{t=1}^{k}\theta_t r_t\sigma_*^2,$$

what is (4.3.4). □

4.9.3 Proof of Corollary 4.3.1

1^o. Observe first that for $\eta \geq 4L$ and $m_t = m$ one has

$$r_t = \frac{\beta_t\overline{\Omega}}{2[\eta/(t+1) - 3L/(t+2)]m} \leq \frac{2\beta_t\overline{\Omega}(t+1)}{\eta m} = \frac{2\beta_t\overline{\Omega}}{\eta_t m}. \tag{4.9.9}$$

As a result, when $\eta \geq \frac{6\overline{\Omega}(k-1)\mathcal{L}}{m}$ and given the definition of β_t and θ_t, we have

$$\begin{aligned}
\theta_t(1-\beta_t)(1+r_t\mathcal{L}) &\leq (t+1)(t-1)\left(1 + \frac{6\overline{\Omega}(t+1)\mathcal{L}}{(t+2)\eta m}\right) \\
&\leq (t+1)(t-1)\left(1 + \frac{(t+1)}{(t+2)(k-1)}\right) \\
&\leq t(t+1) = \theta_{t-1},
\end{aligned}$$

so that (4.3.3) holds. On the other hand,

$$\theta_t\beta_t\eta_t = (t+1)(t+2)\cdot\frac{3\eta}{(t+2)(t+1)} = 3\eta = \theta_{t-1}\beta_{t-1}\eta_{t-1}, \tag{4.9.10}$$

so condition (4.3.2a) is satisfied, so Theorem 4.3.1 applies. By (4.9.9), we also have

$$\theta_t r_t\beta_t L\mathcal{L} \leq \frac{2\theta_t\beta_t^2\overline{\Omega}L\mathcal{L}}{\eta_t m} \leq \frac{18\overline{\Omega}L\mathcal{L}}{\eta_t m}. \tag{4.9.11}$$

2^o. Next, let us check that $\mathbf{E}[V(z_t, x^*)] \le 3D^2$ for all $t \ge 1$. Indeed, for $t = 1$ we have by (4.3.4), the definition of $\theta_t, \beta_t, \eta_t$ and due to (4.9.9)–(4.9.11)

$$\begin{aligned}\mathbf{E}[V(z_1, x^*)] &\le V(z_0, x^*) + \frac{\theta_1\beta_1 r_1 L\mathcal{L}}{\theta_1\beta_1\eta_1}V(z_0, x^*) + \frac{\theta_1 r_1}{\theta_1\beta_1\eta_1}\sigma_*^2 \\ &\le V(z_0, x^*) + \tfrac{8\overline{\Omega}L\mathcal{L}}{m\eta^2}V(z_0, x^*) + \frac{8\overline{\Omega}}{\eta^2 m}\sigma_*^2 \\ &\overset{(ii)}{\le} V(z_0, x^*) + V(z_0, x^*) + D^2 \le 3D^2.\end{aligned}$$

where the last inequality follows from $\eta^2 \ge \max\left\{\frac{9(k+1)^2\overline{\Omega}L\mathcal{L}}{m}, \frac{2(k+2)^3\overline{\Omega}\sigma_*^2}{3D^2 m}\right\}$.

Now we assume that for $s = 1, ..., t-1$, $\mathbf{E}[V(z_s, x^*)] \le 3D^2$. Then

$$\begin{aligned}\mathbf{E}[V(z_t, x^*)] &\le \frac{\theta_1\beta_1\eta_1}{\theta_t\beta_t\eta_t}V(z_0, x^*) + \sum_{s=1}^{t}\frac{\theta_s r_s\beta_s L\mathcal{L}}{\theta_t\beta_t\eta_t}\mathbf{E}[V(z_{s-1}, x^*)] + \sum_{s=1}^{t}\frac{\theta_s r_s}{\theta_t\beta_t\eta_t}\sigma_*^2 \\ &\le V(z_0, x^*) + \sum_{s=1}^{t}\frac{r_s}{\eta_s}L\mathcal{L}\mathbf{E}[V(z_{s-1}, x^*)] + \sum_{s=1}^{t}\frac{r_s}{\beta_t\beta_s\eta_s}\sigma_*^2 \\ &\le V(z_0, x^*) + \frac{3(t+1)^2\overline{\Omega}L\mathcal{L}}{\eta^2 m}\cdot 3D^2 + \frac{2(t+2)^3\overline{\Omega}}{3\eta^2 m}\cdot\sigma_*^2 \\ &\le V(z_0, x^*) + D^2 + D^2 \le 3D^2\end{aligned}$$

where the last inequality is, again, a consequence of $\eta^2 \ge \max\left\{\frac{9(k+1)^2\overline{\Omega}L\mathcal{L}}{m}, \frac{2(k+2)^3\overline{\Omega}\sigma_*^2}{3D^2 m}\right\}$.

3^o. Finally, when substituting the above estimates into (4.3.4), we obtain

$$\begin{aligned}\mathbf{E}[f(x_k) - f^*] &\le \frac{1}{\theta_k}\cdot\left(\theta_1\beta_1\eta_1 V(z_0, x^*) + \sum_{t=1}^{k}3\theta_t r_t\beta_t L\mathcal{L}D^2 + \sum_{t=1}^{k}\theta_t r_t\sigma_*^2\right) \\ &\overset{(i)}{\le} \frac{1}{(k+1)(k+2)}\cdot\left(3\eta D^2 + \frac{27(k+1)^2\overline{\Omega}L\mathcal{L}D^2}{\eta m} + \frac{2(k+2)^3\overline{\Omega}\sigma_*^2}{\eta m}\right) \\ &\overset{(ii)}{\le} \frac{12LD^2}{(k+1)(k+2)} + \frac{18\overline{\Omega}\mathcal{L}D^2}{(k+2)m} + \frac{18\sqrt{\overline{\Omega}L\mathcal{L}}D^2}{(k+1)\sqrt{m}} + \frac{2\sqrt{6(k+2)\overline{\Omega}\sigma_*^2 D^2}}{(k+1)\sqrt{m}}\end{aligned}$$

where (i) follows from (4.9.9)–(4.9.11), and (ii) is a consequence of the definition of η. This implies (4.3.5) due to $\frac{k+2}{k+1} \le \frac{4}{3}$ for $k \ge 2$. □

4.9.4 Proof of Theorem 4.3.2 and Corollary 4.3.2

The subsequent proofs follow those of Theorem 4.3.1 and Corollary 4.3.1. We present them here for reader's convenience.

Proof of Theorem 4.3.2. Note that assumptions of Proposition 4.9.1 hold. When taking expectation on both sides of (4.9.5) and using the bound Lemma 4.3.1, we obtain for $x = x_*$,

$$\begin{aligned}
&\sum_{t=1}^{k} \theta_t \mathbf{E}[f(x_t) - f^*] + \theta_k \beta_k \eta_k \mathbf{E}[V(z_k, x^*)] \\
&\quad \le \sum_{t=1}^{k} \theta_t (1-\beta_t) \mathbf{E}[f(x_{t-1}) - f^*] + \theta_1 \beta_1 \eta_1 V(z_0, x^*) \\
&\quad + \sum_{t=1}^{k} \theta_t r_t \mathbf{E}\big[\bar{\mathcal{K}}^2 \beta_t^2 \|x_{t-1} - z_{t-1}\|^2 + 3\mathcal{L}\big(f(x_{t-1}) - f^* - \langle g(x^*), x_{t-1} - x^* \rangle\big) + 3\sigma_*^2\big].
\end{aligned}$$

By rearranging the terms and utilizing (4.3.9) and $\beta_1 = 1$, we get

$$\begin{aligned}
&\theta_k \mathbf{E}[f(x_k) - f^*] + \theta_k \beta_k \eta_k \mathbf{E}[V(z_k, x^*)] \\
&\le \theta_1 \beta_1 \eta_1 V(z_0, x^*) + 3\theta_1 r_1 \mathcal{L}\big(f(z_0) - f^* - \langle g(x^*), z_0 - x^* \rangle\big) + \sum_{t=1}^{k} \theta_t r_t \mathbf{E}\big[\bar{\mathcal{K}}^2 \beta_t^2 \mathbf{E}\|x_{t-1} - z_{t-1}\|^2 + 3\sigma_*^2\big] \\
&\le \theta_1 \beta_1 \eta_1 V(z_0, x^*) + \frac{3\theta_1 r_1 \mathcal{L} L}{2} \|z_0 - x^*\|^2 + \sum_{t=1}^{k} \theta_t r_t \mathbf{E}\big[\bar{\mathcal{K}}^2 \beta_t^2 \mathbf{E}\|x_{t-1} - z_{t-1}\|^2 + 3\sigma_*^2\big],
\end{aligned}$$

what is (4.3.10). □

Proof of Corollary 4.3.2. Observe that for $\eta \ge 4L$ and $m_t = m$ (cf. (4.9.9))

$$r_t \le \frac{2\beta_t \overline{\Omega}(t+1)}{\eta m}. \tag{4.9.12}$$

Then, due to $\eta \ge \frac{18\overline{\Omega}(k+1)\mathcal{L}}{m}$,

$$\begin{aligned}
\theta_t(1 - \beta_t + 3r_t\mathcal{L}) &\le (t+1)(t+2)\left(\frac{t-1}{t+2} + \frac{18\overline{\Omega}(t+1)\mathcal{L}}{(t+2)\eta m}\right) \\
&\le (t+1)\left(t - 1 + \frac{t+1}{k+1}\right) \\
&\le t(t+1) = \theta_{t-1},
\end{aligned}$$

thus (4.3.9) holds. We have (cf. (4.9.10))

$$\theta_t \beta_t \eta_t = (t+1)(t+2) \cdot \frac{3\eta}{(t+2)(t+1)} = 3\eta = \theta_{t-1}\beta_{t-1}\eta_{t-1}, \tag{4.9.13}$$

so (4.3.2a) is verified and the conclusion of Theorem 4.3.2 applies. We also have

$$\theta_t r_t \beta_t^2 \le \frac{2\theta_t \beta_t^3 \overline{\Omega}(t+1)}{\eta m} \le \frac{54\overline{\Omega}}{\eta m}. \tag{4.9.14}$$

Let us now check that $\mathbf{E}[V(z_t, x^*)] \leq 3D^2$ for all $t \geq 1$. First, for $t = 1$, applying (4.3.10) and taking into account (4.9.12)–(4.9.14), we get

$$\begin{aligned}\mathbf{E}[V(z_1, x^*)] &\leq V(z_0, x^*) + \frac{3\theta_1 r_1 \mathcal{L} L}{2\theta_1\beta_1\eta_1}\|z_0 - x^*\|^2 + \frac{3\theta_1 r_1}{\theta_1\beta_1\eta_1}\sigma_*^2 \\ &\leq V(z_0, x^*) + \frac{12\overline{\Omega}\mathcal{L}L}{m\eta^2}\|z_0 - x^*\|^2 + \frac{24\overline{\Omega}}{\eta^2 m}\sigma_*^2 \\ &\leq V(z_0, x^*) + \frac{1}{24}V(z_0, x^*) + D^2 \leq 3D^2,\end{aligned}$$

the second inequality being due to (4.1.11) and $\eta \geq \max\left\{4L, \frac{18\overline{\Omega}(k+1)\mathcal{L}}{m}, \sqrt{\frac{2(k+2)^3\overline{\Omega}\sigma_*^2}{D^2 m}}\right\}$.

Now, assume that for $s = 1, ..., t-1$, $\mathbf{E}[V(z_{s-1}, x^*)] \leq 3D^2$. Because x_s is a weighted average of the previous iterates, by convexity of $\|\cdot\|^2$, we have for $s < t$,

$$\begin{aligned}\mathbf{E}\|x_s - z_s\|^2 &\leq 2\mathbf{E}\|x_s - x^*\|^2 + 2\mathbf{E}\|z_s - x^*\|^2 \\ &\leq 2\max_{0\leq i\leq s}\mathbf{E}\|z_i - x^*\|^2 + 2\mathbf{E}\|z_s - x^*\|^2 \\ &\leq 4\max_{0\leq i\leq s}\mathbf{E}[V(z_i, x^*)] + 4\mathbf{E}[V(z_s, x^*)] \leq 24D^2. \qquad (4.9.15)\end{aligned}$$

As a consequence, when substituting into (4.3.10) the bounds of (4.9.12)–(4.9.14)

$$\begin{aligned}\mathbf{E}[V(z_k, x^*)] &\leq V(z_0, x^*) + \frac{3\theta_1 r_1 \mathcal{L} L}{2\theta_k\beta_k\eta_k}\|z_0 - x^*\|^2 + \sum_{t=2}^{k}\frac{24\theta_t r_t \beta_t^2 \mathcal{K}^2 D^2}{\theta_k\beta_k\eta_k} + \sum_{t=1}^{k}\frac{3\theta_t r_t \sigma_*^2}{\theta_k\beta_k\eta_k} \\ &\overset{(ii)}{\leq} V(z_0, x^*) + \frac{12\overline{\Omega}\mathcal{L}L}{m\eta^2}\|z_0 - x^*\|^2 + \frac{144\overline{\Omega}(k-1)\mathcal{K}^2 D^2}{\eta^2 m} + \frac{2(k+2)^3\overline{\Omega}\sigma_*^2}{\eta^2 m} \overset{(i)}{\leq} 3D^2,\end{aligned}$$

due to

$$\eta \geq \max\left\{4L, \frac{18\overline{\Omega}(k+1)\mathcal{L}}{m}, 12\sqrt{\frac{k\overline{\Omega}\mathcal{K}^2}{m}}, \sqrt{\frac{2(k+2)^3\overline{\Omega}\sigma_*^2}{D^2 m}}\right\}.$$

Finally, from (4.3.10) and the definition of η we conclude that

$$\begin{aligned}\mathbf{E}[f(x_k) - f^*] &\leq \frac{1}{\theta_k}\left(\theta_1\beta_1\eta_1 V(z_0, x^*) + \frac{3\theta_1 r_1 \mathcal{L} L}{2}\|z_0 - x^*\|^2 + \sum_{t=2}^{k}\theta_t r_t \beta_t^2 \mathcal{K}^2 \mathbf{E}\|\bar{z}_{t-1} - z_{t-1}\|^2 + \sum_{t=1}^{k} 3\theta_t r_t \sigma_*^2\right) \\ &\leq \frac{1}{(k+1)(k+2)}\left(3\eta D^2 + \frac{36\overline{\Omega}\mathcal{L}L}{m\eta} + \frac{432(k-1)\overline{\Omega}\mathcal{K}^2 D^2}{\eta m} + \frac{6(k+2)^3\overline{\Omega}\sigma_*^2}{\eta m}\right) \\ &\leq \frac{13LD^2}{(k+1)(k+2)} + \frac{54\overline{\Omega}\mathcal{L}D^2}{(k+2)m} + \frac{72\mathcal{K}D^2}{(k+2)}\sqrt{\frac{\overline{\Omega}}{m(k+1)}} + \frac{6}{k+1}\sqrt{\frac{2(k+2)\overline{\Omega}\sigma_*^2 D^2}{m}}.\end{aligned}$$

This completes the proof due to $\frac{k+2}{k+1} \leq \frac{4}{3}$ for $k \geq 2$. □

4.9.5 Proof of Theorem 4.4.1

For $0 \leq \beta_t \leq 1$, denote $\tau_t = \frac{1-\beta_t}{\beta_t}$. Relationships (4.4.5) in variables $\theta_t, \alpha_t, \eta_t$, and τ_t become

$$\theta_{t-1} = \alpha_t\theta_t, \ \eta_t \leq \alpha_t\eta_{t-1}, \ t = 2, \ldots, k \qquad (4.9.16a)$$

$$\frac{\eta_t\tau_{t-1}}{\alpha_t} \geq 5L, \quad t = 3, \ldots, k \qquad (4.9.16b)$$

$$\frac{\eta_1\eta_2}{\alpha_2} \geq 25L^2, \ \eta_k\tau_k \geq L, \qquad (4.9.16c)$$

with (4.4.1c) of Algorithm 4 replaced with

$$x_t = (z_t + \tau_t x_{t-1})/(1+\tau_t). \tag{4.9.17}$$

We need the following technical statement.

Proposition 4.9.2 *If the algorithm parameters $\{\alpha_t\}$, $\{\eta_t\}$ and $\{\tau_t\}$ satisfy $\tau_1 \geq 0$, $\tau_t > 0$ for all $t \geq 2$, along with relations (4.9.16a)–(4.9.16c) for some $\theta_t \geq 0$. Then for all $x \in X$*

$$\sum_{t=1}^{k} \theta_t \left\{\tau_t[f(x_t) - f(x_{t-1})] - \langle g(x_t), x - x_t\rangle\right\} \leq \theta_1\eta_1 V(x_0, x) + \sum_{t=0}^{k-1}\left[p_t\langle\delta_t, x - z_t\rangle + \frac{5q_t}{2}\|\delta_t\|_*^2\right] \tag{4.9.18}$$

where $p_t := \begin{cases} \theta_t, & t \leq k-2, \\ \theta_{t+1} + \theta_t, & t = k-1 \end{cases}$ *and* $\delta_t := G_t - g(x_t)$.

Proof of the proposition.

1^o. Denote $g_t := g(x_t)$. By the smoothness of f,

$$\frac{1}{2L}\|g_t - g_{t-1}\|_*^2 \leq f(x_{t-1}) - [f(x_t) + \langle g_t, x_{t-1} - x_t\rangle].$$

Then for any $x \in X$, using the above inequality and the fact that $x - x_t = x - z_t - \tau_t(x_{t-1} - x_t)$ due to (4.9.17), we get

$$\begin{aligned}\tau_t f(x_t) - \langle g_t, x - x_t\rangle &= \tau_t[f(x_t) + \langle g_t, x_{t-1} - x_t\rangle] + \langle g_t, z_t - x\rangle \\ &\leq \tau_t\left[f(x_{t-1}) - \frac{1}{2L}\|g_t - g_{t-1}\|_*^2\right] + \langle g_t, z_t - x\rangle.\end{aligned}$$

By the optimality condition of (4.4.1b), we have

$$\langle \widetilde{G}_t, z_t - x\rangle \leq \eta_t V(z_{t-1}, x) - \eta_t V(z_t, x) - \eta_t V(z_{t-1}, z_t).$$

Combining two previous inequalities, we obtain

$$\begin{aligned}&\tau_t f(x_t) - \langle g_t, x - x_t\rangle - \tau_t f(x_{t-1}) \\ &\leq \eta_t V(z_{t-1}, x) - \eta_t V(z_t, x) + \langle g_t - \widetilde{G}_t, z_t - x\rangle - \frac{\tau_t}{2L}\|g_t - g_{t-1}\|_*^2 - \eta_t V(z_{t-1}, z_t).\end{aligned} \tag{4.9.19}$$

Note that

$$\begin{aligned}\langle g_t - \widetilde{G}_t, z_t - x\rangle &= \langle g_t - g_{t-1} - \alpha_t(g_{t-1} - g_{t-2}), z_t - x\rangle - \langle \delta_{t-1} + \alpha_t(\delta_{t-1} - \delta_{t-2}), z_t - x\rangle \\ &= \langle g_t - g_{t-1}, z_t - x\rangle - \alpha_t\langle g_{t-1} - g_{t-2}, z_{t-1} - x\rangle \\ &\quad + \alpha_t\langle g_{t-1} - g_{t-2}, z_t - z_{t-1}\rangle - \langle \delta_{t-1} + \alpha_t(\delta_{t-1} - \delta_{t-2}), z_t - x\rangle.\end{aligned} \tag{4.9.20}$$

When taking the θ_t-weighted sum of the above inequalities for $t = 1, \ldots, k$, noting that $x_0 = z_0$, and using (4.9.16a), we obtain

$$\begin{aligned}&\sum_{t=1}^{k} \theta_t \left\{\tau_t[f(x_t) - f(x_{t-1})] - \langle g_t, x - x_t\rangle\right\} \\ &\leq \theta_1\eta_1 V(x_0, x) - \theta_k\eta_k V(z_k, x) + \theta_k\langle g_k - g_{k-1}, z_k - x\rangle + \Delta_k,\end{aligned} \tag{4.9.21}$$

where

$$\Delta_k := \sum_{t=1}^{k} \theta_t \big[\alpha_t \langle g_{t-1} - g_{t-2}, z_t - z_{t-1}\rangle - \frac{\tau_t}{2L}\|g_t - g_{t-1}\|_*^2 \\ - \langle \delta_{t-1} + \alpha_t(\delta_{t-1} - \delta_{t-2}), z_t - x\rangle - \eta_t V(z_{t-1}, z_t)\big].$$

Observe that

$$\theta_t\alpha_t\langle g_{t-1} - g_{t-2}, z_t - z_{t-1}\rangle \le \frac{\theta_t^2\alpha_t^2 L}{2\theta_{t-1}\tau_{t-1}}\|z_t - z_{t-1}\|^2 + \frac{\theta_{t-1}\tau_{t-1}}{2L}\|g_{t-1} - g_{t-2}\|_*^2$$

so that, after rearranging the terms,

$$\Delta_k \le \underbrace{\theta_2\alpha_2\langle g_1 - g_0, z_2 - z_1\rangle + \sum_{t=3}^{k}\frac{L\theta_t^2\alpha_t^2}{2\theta_{t-1}\tau_{t-1}}\|z_t - z_{t-1}\|^2 - \frac{\theta_k\tau_k}{2L}\|g_k - g_{k-1}\|_*^2 - \tfrac{1}{5}\sum_{t=1}^{k}\theta_t\eta_t V(z_{t-1}, z_t)}_{=:\Delta_{k,1}}$$
$$\underbrace{-\sum_{t=1}^{k}\theta_t\langle \delta_{t-1} + \alpha_t(\delta_{t-1} - \delta_{t-2}), z_t - x\rangle - \frac{4}{5}\sum_{t=1}^{k}\theta_t\eta_t V(z_{t-1}, z_t)}_{=:\Delta_{k,2}}.$$

2^o. Let us bound $\Delta_{k,1}$. By the Young's inequality,

$$\alpha_2\langle g_1 - g_0, z_2 - z_1\rangle \le \frac{5\alpha_2^2}{2\eta_2}\|g_1 - g_0\|_*^2 + \frac{\eta_2}{10}\|z_2 - z_1\|^2 \le \frac{5\alpha_2^2}{2\eta_2}\|g_1 - g_0\|_*^2 + \frac{\eta_2}{5}V(z_1, z_2),$$

thus

$$\Delta_{k,1} \le \frac{5\theta_2\alpha_2^2\|g_1 - g_0\|_*^2}{2\eta_2} - \frac{1}{5}\theta_1\eta_1 V(z_1, z_0) + \sum_{t=3}^{k}\left(\frac{L\theta_t^2\alpha_t^2}{2\theta_{t-1}\tau_{t-1}} - \frac{\theta_t\eta_t}{10}\right)\|z_t - z_{t-1}\|^2 - \frac{\theta_k\tau_k}{2L}\|g_k - g_{k-1}\|_*^2$$
$$\le -\frac{\theta_k\tau_k}{2L}\|g_k - g_{k-1}\|_*^2 \tag{4.9.22}$$

where the last inequality follows from (5.3.2) and the bound

$$\frac{5\theta_2\alpha_2^2\|g_1 - g_0\|_*^2}{2\eta_2} \le \frac{5\theta_2\alpha_2^2 L^2\|x_1 - x_0\|^2}{2\eta_2} = \frac{5\theta_2\alpha_2^2 L^2\|z_1 - z_0\|^2}{2\eta_2} \le \frac{5\theta_2\alpha_2^2 L^2 V(z_1, z_0)}{\eta_2} \le \frac{\theta_1\eta_1 V(z_1, z_0)}{5}.$$

Note that

$$-\eta_k V(z_k, x) + \langle g_k - g_{k-1}, z_k - x\rangle - \frac{\tau_k}{2L}\|g_k - g_{k-1}\|_*^2$$
$$\le -\frac{\eta_k}{2}\|z_k - x\|^2 + \langle g_k - g_{k-1}, z_k - x\rangle - \frac{\tau_k}{2L}\|g_k - g_{k-1}\|_*^2$$
$$\le -\left(\frac{\eta_k}{2} - \frac{L}{2\tau_k}\right)\|z_k - x\|^2 \le 0$$

due to the second relationship in (4.9.16c). Now, substituting the bound (4.9.22) into (4.9.21) results in

$$\sum_{t=1}^{k}\theta_t\left\{\tau_t[f(x_t) - f(x_{t-1})] - \langle g_t, x - x_t\rangle\right\}$$
$$\le \theta_1\eta_1 V(x_0, x) - \theta_k\eta_k V(z_k, x) + \theta_k\langle g_k - g_{k-1}, z_k - x\rangle - \frac{\theta_k\tau_k}{2L}\|g_k - g_{k-1}\|_*^2 + \Delta_{k,2}$$
$$\le \theta_1\eta_1 V(x_0, x) + \Delta_{k,2}. \tag{4.9.23}$$

3^o. To bound $\Delta_{k,2}$ we act as follows. Observe that

$$\begin{aligned}\Delta_{k,2} \le &-\sum_{t=1}^{k}\theta_t[\langle\delta_{t-1}, z_t - z_{t-1}\rangle + \langle\delta_{t-1}, z_{t-1} - x\rangle]\\ &-\sum_{t=1}^{k}\theta_t\alpha_t[\langle\delta_{t-1}, z_t - z_{t-1}\rangle + \langle\delta_{t-1}, z_{t-1} - x\rangle]\\ &+\sum_{t=2}^{k}\theta_t\alpha_t[\langle\delta_{t-2}, z_t - z_{t-1}\rangle + \langle\delta_{t-2}, z_{t-1} - z_{t-2}\rangle + \langle\delta_{t-2}, z_{t-2} - x\rangle]\\ &-\frac{2}{5}\sum_{t=1}^{k}\theta_t\eta_t\|z_{t-1} - z_t\|^2.\end{aligned}$$

Recall that z_t is $\mathcal{F}_{t-1}$-measurable. When using the bound

$$\langle\delta_s, z_r - z_{r-1}\rangle - a\|z_{r-1} - z_r\|^2 \le \|\delta_s\|_*^2/(4a)$$

for $s < r$ and $a > 0$, we get

$$\begin{aligned}\Delta_{k,2} &\le \sum_{t=1}^{k}\left[\frac{5\theta_t}{2\eta_t}\|\delta_{t-1}\|_*^2 - \theta_t\langle\delta_{t-1}, z_{t-1} - x\rangle\right]\\ &\quad+\sum_{t=1}^{k}\left[\frac{5\theta_t\alpha_t^2}{2\eta_t}\|\delta_{t-1}\|_*^2 - \theta_t\alpha_t\langle\delta_{t-1}, z_{t-1} - x\rangle\right]\\ &\quad+\sum_{t=2}^{k}\left[\frac{5\theta_t\alpha_t^2}{2\eta_t}\|\delta_{t-2}\|_*^2 + \frac{5\theta_t^2\alpha_t^2}{2\theta_{t-1}\eta_{t-1}}\|\delta_{t-2}\|_*^2 + \theta_t\alpha_t\langle\delta_{t-2}, z_{t-2} - x\rangle\right]\\ &\le \sum_{t=1}^{k}\frac{5}{2}\left[\frac{\theta_t(1+\alpha_t^2)}{\eta_t} + \frac{\theta_{t+1}\alpha_{t+1}^2}{\eta_{t+1}} + \frac{\theta_{t+1}^2\alpha_{t+1}^2}{\theta_t\eta_t}\right]\|\delta_{t-1}\|_*^2\\ &\quad-\sum_{t=1}^{k}\theta_t(1+\alpha_t)\langle\delta_{t-1}, z_{t-1} - x\rangle + \sum_{t=1}^{k-1}\theta_t\langle\delta_{t-1}, z_{t-1} - x\rangle\\ &\le \sum_{t=1}^{k}\frac{5}{2}\left[\frac{\theta_t(1+\alpha_t^2)}{\eta_t} + \frac{\theta_{t+1}\alpha_{t+1}^2}{\eta_{t+1}} + \frac{\theta_{t+1}^2\alpha_{t+1}^2}{\theta_t\eta_t}\right]\|\delta_{t-1}\|_*^2\\ &\quad-\sum_{t=0}^{k-1}\theta_t\langle\delta_t, z_t - x\rangle - \theta_k\langle\delta_{k-1}, z_{k-1} - x\rangle. \qquad (4.9.24)\end{aligned}$$

When substituting the bound (4.9.24) for $\Delta_{k,2}$ into (4.9.23) we obtain (4.9.18). □

Proof of the theorem. By taking expectation on both sides of (4.9.18), and using (SN), (4.2.2), and the fact that

$$\langle g(x_t), x - x_t\rangle \le f(x) - f(x_t)$$

we have for all $x \in X$,

$$\sum_{t=1}^{k} \theta_t \mathbf{E}\left\{\tau_t[f(x_t) - f(x_{t-1})] - [f(x) - f(x_t)]\right\}$$

$$\leq \theta_1\eta_1 V(x_0, x) + \sum_{t=1}^{k} \epsilon_{t-1}\left\{\mathcal{L}\mathbf{E}[f(x_{t-1}) - f^* - \langle g(x^*), x_{t-1} - x^*\rangle] + \sigma_*^2\right\}.$$

Note that (4.4.7) implies that

$$\theta_t\tau_t + \mathcal{L}\epsilon_{t-1} \leq \theta_{t-1}(1 + \tau_{t-1}),\ t \geq 2. \tag{4.9.25}$$

Thus, by setting $x = x^*$ and using $\langle g(x^*), x_t - x^*\rangle \geq 0$, $f(x_0) - f^* - \langle g(x^*), x_0 - x^*\rangle \leq \frac{L}{2}\|x_0 - x^*\|^2$, and $\tau_1 = 0$, we obtain

$$\begin{aligned}
\theta_k(1 + \tau_k)\mathbf{E}[f(x_k) - f^*] &\leq \sum_{t=1}^{k} \theta_t(1 + \tau_t)\mathbf{E}[f(x_t) - f^*] - \sum_{t=1}^{k-1} \theta_t(1 + \tau_t)\mathbf{E}[f(x_t) - f^*] \\
&\leq \sum_{t=1}^{k} \theta_t(1 + \tau_t)\mathbf{E}[f(x_t) - f^*] - \sum_{t=2}^{k} (\theta_t\tau_t + \epsilon_{t-1}\mathcal{L})\mathbf{E}[f(x_{t-1}) - f^*] \\
&\leq \theta_1\eta_1 V(x_0, x^*) + \frac{\epsilon_0\mathcal{L}L}{2}\|x_0 - x^*\|^2 + \sum_{t=1}^{k} \epsilon_{t-1}\sigma_*^2
\end{aligned}$$

which is (4.4.8). □

4.9.6 Proof of Corollary 4.4.1

1^o. Note that in the premise of the corollary one has $\tau_t = \frac{t-1}{3}$. Let us check that with the present choice of stepsize parameters conditions (4.9.16a)–(4.9.16c) and (5.3.7) (which are equivalents (4.4.5) and (4.4.7)) are satisfied. It is easy to see that (4.9.16a) holds. Because $\eta \geq 30L$, we have

$$\begin{aligned}
\eta_1\eta_2 &= \frac{1}{2} \cdot (30L)^2 \geq 25\alpha_2 L^2, \\
\eta_t\tau_{t-1} &= \frac{\eta}{t} \cdot \frac{t-2}{3} \geq \frac{10L(t-2)}{t} \geq 5L\alpha_t,\ t = 3, ..., k \\
\eta_k\tau_k &\geq \frac{\eta}{k} \cdot \frac{k-1}{3} \geq \frac{10L(k-1)}{k} \geq L.
\end{aligned}$$

To check (5.3.7), notice that for $t \geq 1$,

$$q_t = \frac{\theta_{t+1}(1 + \alpha_{t+1}^2)}{\eta_{t+1}} + \frac{\theta_{t+2}\alpha_{t+2}^2}{\eta_{t+2}} + \frac{\theta_{t+2}^2\alpha_{t+2}^2}{\theta_{t+1}\eta_{t+1}} = \frac{3(t+1)^2 + t^2}{\eta} \leq \frac{4(t+1)^2}{\eta},$$

thus

$$\epsilon_t = \frac{5\overline{\Omega}q_t}{2m_t} \leq \frac{10\overline{\Omega}(t+1)^2}{\eta m}.$$

Given the above inequality, for $t \geq 3$ we have

$$\theta_t \tau_t + \mathcal{L}\epsilon_{t-1} = t \cdot \frac{t-1}{3} + \mathcal{L}\epsilon_{t-1} \leq \frac{t(t-1)}{3} + \frac{10\overline{\Omega}t^2\mathcal{L}}{\eta m} \leq \frac{t(t-1)}{3} + \frac{t-1}{3} = \frac{(t+1)(t-1)}{3},$$

where the second inequality follows from $\eta \geq \frac{30\overline{\Omega}(k+2)\mathcal{L}}{m}$. Combining the above bound with $\theta_{t-1}(1+\tau_{t-1}) = \frac{(t-1)(t+1)}{3}$, we arrive at

$$\theta_{t-1}(1+\tau_{t-1}) - (\theta_t\tau_t + \mathcal{L}\epsilon_{t-1}) \geq 0, \qquad t \geq 2,$$

which is (5.3.7). We conclude that the bound (4.4.8) of Theorem (4.4.1) holds.

On the other hand, when $\eta \geq \sqrt{\frac{10\overline{\Omega}(k+1)^3\sigma_*^2}{3mD^2}}$ we have

$$\sum_{t=0}^{k-1} \epsilon_t\sigma_*^2 \leq \sum_{t=0}^{k-1} \frac{10\overline{\Omega}(t+1)^2\sigma_*^2}{\eta m} \leq \frac{10\overline{\Omega}(k+1)^3\sigma_*^2}{3\eta m} \leq \sqrt{\frac{10\overline{\Omega}(k+1)^3\sigma_*^2 D^2}{3m}}.$$

When substituting the above bound into (4.4.8) and noticing that $\theta_k(1+\tau_k) = \frac{k(k+2)}{3}$ and $\mathcal{L}\epsilon_0 \leq \frac{1}{3}$, we conclude that

$$\mathbf{E}[f(x_k) - f^*] \leq \frac{91LD^2}{k(k+2)} + \frac{90\overline{\Omega}\mathcal{L}D^2}{mk} + \sqrt{\frac{120\overline{\Omega}\sigma_*^2 D^2}{mk}},$$

which completes the proof of (4.4.9).

2^o. The "Furthermore" part of the statement immediately follows from (4.1.11) and the fact that when $\eta \geq \sqrt{\frac{20(k+1)^3\overline{\Omega}\sigma_*^2}{3m\Omega R^2}}$ one has

$$\sum_{t=0}^{k-1} \epsilon_t\sigma_*^2 \leq \frac{10\overline{\Omega}(k+1)^3\sigma_*^2}{3\eta m} \leq \sqrt{\frac{5\Omega\overline{\Omega}(k+1)^3\sigma_*^2 R^2}{3m}}.$$ □

4.9.7 Proofs of Corollaries 4.5.1 and 4.6.1

Proof of Corollary 4.5.1. Note that bound (4.4.11) of Corollary 4.4.1 implies that whenever size m of the mini-batch satisfies

$$m \geq \max\left\{1, \frac{18(k+2)\Omega\mathcal{L}}{L}, \frac{15N(N+2)^2\sigma_*^2}{L^2R^2}\right\},$$

we have for the approximate solution x_N by SGE after N iterations,

$$\begin{aligned}\mathbf{E}[f(x_N) - f^*] &\leq \frac{91L\Omega R^2}{2N(N+2)} + \frac{45\Omega^2\mathcal{L}R^2}{mk} + \sqrt{\frac{60\Omega^2\sigma_*^2R^2}{mk}} \\ &\leq \frac{91L\Omega R^2}{2N(N+2)} + \frac{5L\Omega R^2}{2N(N+2)} + \frac{2L\Omega R^2}{N(N+2)} \\ &= \frac{50L\Omega R^2}{N(N+2)}\end{aligned} \tag{4.9.26}$$

where R is an upper bound for the "initial distance to x^*".

Note that $\|y^0 - x^*\| \le R_0$. Let us now assume that for $1 \le k \le K$, $\|y^{k-1} - x^*\| \le R_k$, so that at the beginning of the kth stage $\|x_0 - x^*\| \le R_k$. Based on the above inequality, by the choice of N,

$$\mathbf{E}[f(x_N) - f^*] \le \frac{50L\Omega R_{k-1}^2}{N(N+2)} \le \frac{\mu R_{k-1}^2}{4} = 2^{-k-1}\mu R_0^2 = \frac{1}{2}\mu R_k^2.$$

Due to (4.5.1) this also means that

$$\mathbf{E}[\|x_N - x^*\|^2] \le R_k^2 = 2^{-k}R_0^2. \qquad \square$$

Proof of Corollary 4.6.1. We have $\|\overline{y}^0 - x^*\| \le R_0$. Let us assume that $\|\overline{y}^{k-1} - x^*\| \le R_{k-1}$ for some $1 \le k \le K$. From (4.9.26) we conclude that

$$\mathbf{E}[f(y^k) - f^*] \le \frac{50L\Omega R_{k-1}^2}{N(N+2)} \le \frac{\underline{\kappa} R_k^2}{16s} = 2^{-k-4}s^{-1}\underline{\kappa}R_0^2$$

by the choice of N. Furthermore, recalling that $\overline{y}^k = \mathsf{sparse}(y^k)$, we have

$$\|\overline{y}^k - x^*\| \le \sqrt{2s}\|\overline{y}^k - x^*\|_2 \le 2\sqrt{2s}\|y^k - x^*\|_2. \tag{4.9.27}$$

Indeed, given that both $\overline{y}^k$ and x^* are s-sparse, we conclude that $\overline{y}^k - x^*$ is $2s$-sparse, thus $\|\overline{y}^k - x^*\| \le \sqrt{2s}\|\overline{y}^k - x^*\|_2$. On the other hand, by the optimality of $\overline{y}^k$ for (4.6.6),

$$\|\overline{y}^k - x^*\|_2 \le \|\overline{y}^k - y^k\|_2 + \|y^k - x^*\|_2 \le 2\|y^k - x^*\|_2.$$

We conclude that

$$\mathbf{E}[\|\overline{y}^k - x^*\|^2] \le 8s\|y^k - x^*\|_2^2 \le 16s\underline{\kappa}^{-1}\mathbf{E}[f(y^k) - f^*] \le R_k^2 = 2^{-k}R_0^2$$

which completes the proof. $\square$

Chapter 5

Large Deviation Bounds, Accuracy Certificates and Composite Algorithm

Chapter Abstract

In this final chapter, we build upon the accelerated algorithms introduced and discussed in the previous chapter. We continue to explore stochastic optimization algorithms over a class of smooth convex objectives. To ensure reliability of the solutions computed by the SGE method and its multistage conterparts, we provide high-probability guarantees under the assumption of sub-exponential state-dependent gradient noise. These bounds match the optimal in-expectation bounds up to logarithmic factors of the confidence level. Following this, we derive accuracy certificates for the SGE algorithm. This provides an interesting procedure to compute on-the-fly stochastic upper bounds for the unknown suboptimality gap, ultimately leading to a practical stopping criterion for the SGE algorithm.

Additionally, we propose an adaptive method to address the statistical problem of sparse parameters estimation. We draw inspiration from the framework introduced in Chapter 2 of the manuscript to offer an analysis of a multistage algorithm for minimizing over the set of sparse minimizers. The proposed method involves solving a sequence of norm-penalized stochastic composite optimization problems at each stage of the multistage routine. Each stage is solved up to a prescribed accuracy by the non-Euclidean *Composite Stochastic Gradient Extrapolation* (CSGE) algorithm. Finally, we apply the multistage algorithm to solve the sparse GLR problem under the "vanilla" sparsity structure and provide numerical validation of the approach.

5.1 Introduction

This chapter studies stochastic optimization problems of the form

$$f^* := \min_{x \in X} f(x) \tag{5.1.1}$$

where X is assumed to be a closed convex subset of a Euclidean space E and $f : X \to \mathbf{R}$ is a smooth convex function with Lipschitz continuous gradient, i.e., for some $L \geq 0$,

$$0 \leq f(y) - f(x) - \langle \nabla f(x), x - y \rangle \leq \frac{L}{2}\|x - y\|^2 \tag{5.1.2}$$

We assume that the set of minimizers X^* is nonempty.

We consider the stochastic setting where only stochastic first-order information about f is available for solving problem (5.1.1). Specifically, at the current search point $x_t \in X$, a stochastic

oracle (SO) generates the stochastic operator $\mathcal{G}(x_t, \xi_t)$, where $\xi_t \in \Xi$ denotes a random variable whose probability distribution is supported on a set Ξ. We assume ξ_t is independent of $x_0, ..., x_t$, and $(\xi_t)_{t\geq 0}$ are mutually independent.

We assume that $\mathcal{G}(x_t, \xi_t)$ is an unbiased estimator of $g(x_t) = \nabla f(x_t)$ satisfying:

$$\mathbb{E}_{\xi_t}[\mathcal{G}(x_t, \xi_t)] = g(x_t) \tag{5.1.3}$$

In this chapter we are, again, motivated by the problem of parameter estimation in the *generalized linear regression* (GLR) model. We want to estimate the unknown parameter vector $x^* \in X \subset \mathbf{R}^n$ given observations (ϕ_t, η_t),

$$\eta_t = u(\phi_t^T x^*) + \zeta_t, \quad t = 1, 2, ..., \tag{5.1.4}$$

where, $\eta_t \in \mathbf{R}$ are the responses, $\phi_t \in \mathbf{R}^n$ are random regressors, $\zeta_t \in \mathbf{R}$ are zero-mean random noises which are assumed to be mutually independent and independent of ϕ_t, and $u : \mathbf{R} \to \mathbf{R}$ is the "activation function". Given the settings, one clearly has

$$\mathbf{E}[\phi_t(u(\phi_t^T x^*) - \eta_t)] = \mathbf{E}[\phi_t\, \zeta_t] = 0. \tag{5.1.5}$$

Thus, the problem of recovery of x^* from observations η_t and ϕ_t may be formulated as a stochastic optimization problem. Specifically, when denoting $v : \mathbf{R} \to \mathbf{R}$ the primitive of u, i.e., $v'(t) = u(t)$ and assuming that $x^* \in \operatorname{int} X$, (5.1.5) may be seen as as the optimality condition for the problem

$$\min_{x\in X} \left\{ f(x) := \mathbf{E}[v(\phi^T x) - \phi^T x \eta] \right\}. \tag{5.1.6}$$

Expanding on the work laid out in the previous chapter, our current motivation is to provide high-probability guarantees for the SGE method and its multistage variant. Large deviation bounds are desirable to ensure the reliability and robustness of the solutions. A sub-Gaussian tail condition, e.g.,

$$\mathbf{E}_{\xi_t}\left[\exp\{\|\mathcal{G}(x_t, \xi_t) - g(x_t)\|_*^2 / \sigma_t^2(x_t)\} \middle| x_t \right] \leq \exp(1), \quad t \in \mathbf{Z}_+,$$

is often assumed when proving such bounds (cf., e.g., [87, 88, 105], among others).[1] However, this assumption is violated in many well-known applications. For instance, in the GLR problem, it implies the boundedness of the regressors; one can easily check that if ϕ_t are Gaussian (see Section 5.5.4), the stochastic oracle noise becomes sub-exponential:

$$\mathbf{E}_{\xi_t}\left[\exp\{\|\mathcal{G}(x_t, \xi_t) - g(x_t)\|_* / \sigma_t(x_t)\} \middle| x_t \right] \leq \exp(1), \quad t \in \mathbf{Z}_+.$$

Besides the sub-exponential structure, another challenging problem when proving large deviation bounds under state-dependent noise assumption stems from the fact that iterates x_t of the stochastic algorithm are random variables, thus the normalizing factor $\sigma_t(x_t)$ is a random variable itself. As a consequence, classical large deviation inequalities for martingales (cf., e.g., [113, 118, 153, 154]) are not well suited to obtain precise concentration bounds for the trajectories of stochastic approximation.

A second aspect explored in this chapter is the derivation of accuracy certificates for the SGE algorithm. Accuracy certificates usually provide upper and lower bounds on the unknown optimal objective value in optimization problems. In particular, we focus on stochastic certificates, which are

[1] It is worth noting that authors of [59] proposed a Median-of-Means type technique [152] to obtain high probability guarantees without additional assumptions on the tail distribution. However, this technique requires data splitting and post-manipulation after running stochastic approximation algorithms, introducing additional computational efforts and implementation issues.

generated iteratively as the optimization algorithm progresses and are used to evaluate the quality of the current solution. This approach, combined with a specified inaccuracy level ϵ, allows the computation of upper and lower estimates for the unknown suboptimality gap. These estimates can ultimately be used to formulate a stopping criterion for the optimization procedure. Online accuracy certificates have previously been derived for the Mirror Descent SA algorithm in [155]. In their work, the authors provide in-expectation and high-probability guarantees for these certificates by, notably exploiting the assumption of sub-Gaussian stochastic gradients. In our study, we extend similar ideas to the SGE algorithm. We offer a solution to compute an accuracy certificate for the objective's suboptimality. This upper bound possesses a comparable optimal convergence rate of $\mathcal{O}(1/k^2)$, characteristic of smooth convex programming, and offers the distinct advantage of being computable "on-the-fly" during the execution of the SGE algorithm, providing a real-time measure for controlling the objective's suboptimality gap and thus providing a termination criterion for the algorithm. The theoretical guarantees for this approach are established within the context of sub-exponential stochastic gradient noise

Finally, we focus on the development of a novel algorithm for the problem of sparse recovery. Inspired by the work developed in Chapter 2, we propose in Section 5.5.3, a multistage procedure based on the *Composite Stochastic Gradient Extrapolation* (CSGE) algorithm. The CSGE algorithm is a variant of the SGE algorithm in which the usual proximal operator is replaced by the same composite proximal operator used within the CSMD algorithm. Each stage of the multistage routine is a specific run of the CSGE algorithm solving a proxy sub-composite stochastic problem. By considering a sparsity-inducing penalization function, the algorithm produces estimates with a sparse structure. We derive an analysis for the sparse recovery algorithm, the CSGE-SR algorithm, by building on the *Reduced Strong Convexity* assumption introduced in Chapter 2. The latter assumption offers a general framework to provide an analysis of the multistage procedure adaptable to different types of sparsity structures. By presenting an application in the context of sparse GLR with minima x^* possessing a "vanilla" sparsity structure, we prove that our accelerated multistage algorithm achieves the optimal iteration complexity and the optimal sample complexity up to some logarithmic terms. Similar to the analysis in the previous chapter under noise (SN) which revealed a sample complexity with three terms: the deterministic error, the state-dependent stochastic error, and the state-independent stochastic error, the analysis of our CSGE-SR algorithm also reveals these same three terms in the sample complexity. We show that the CSGE-SR algorithm improves on the CSMD-SR algorithm, where the algorithm has the optimal dependence on the condition number of the problem, while achieving a similar state-independent stochastic error. The multistage method also has the same advantages as CSMD-SR over the SGE-SR algorithm presented in Chapter 4, since the CSGE-SR algorithm can be made adaptive to two unknown parameters at the same time using Lepski's adaptation procedure (see Chapter 3).

The organization of the following sections of this chapter is as follows. In Section 5.2 we present the state-dependent sub-exponential stochastic noise assumption and we introduce the mini-batch setup. Additionally, we discuss an important technical result on large deviations for handling sequences of sub-exponential random variables. This result serves as the basis for providing the high-probability guarantees in our setup. In Section 5.3, we provide the high-probability guarantees for the SGE method. In addition, by exploiting the quadratic growth condition on the objective function, we propose a variant of Algorithm 5 with a shrinking domain, for which we also derive high-probability bounds. Section 5.4 focuses on the derivation of accuracy certificates for the SGE algorithm. In Section 5.5 we present a setup, inspired by the results developed in Chapter 2 of the manuscript, to design an accelerated multistage composite algorithm for sparse recovery based on the *Composite Stochastic Gradient Extrapolation* (CSGE) method. Finally, in Section 5.6, we present a simulation study to illustrate the performance of our multistage accelerated algorithm in a

high-dimensional sparse recovery problem. Until Section 5.5, we will use the same notation as in the previous chapter.

5.2 Problem statement

We provide below a brief summary of the setting under which we derive the analysis of the accelerated methods in order to provide bounds that are valid in high-probability.

5.2.1 Assumptions

We consider the problem of stochastic optimization with a smooth and convex objective.

To derive the high-probability guarantees presented in Sections 5.3, 5.4 and 5.5, Assumption (SN) introduced in the previous analysis is replaced with the following one:

- [*Sub-exponential tails*] One has

$$\mathbf{E}_{\xi_t}\left[\exp\left\{\frac{\|\mathcal{G}(x_t,\xi_t)-g(x_t)\|_*}{\sigma_t}\right\}\right]\leq \exp(1), \quad t\in\mathbf{Z}_+, \tag{SEN}$$

  where, similarly to (SN), for some $\mathcal{L}<\infty$ and $x^*\in X^*$ we have,

$$\sigma_t^2 := \mathcal{L}[f(x_t)-f(x^*)-\langle g(x^*),x_t-x^*\rangle]+\sigma_*^2.$$

5.2.2 Mini-batch setup

Here again, we consider the mini-batch approach widely used in practice. Specifically, we assume that at each search point x_t, the stochastic oracle is called repeatedly, thus generating m_t i.i.d. samples $\{\xi_{t,i}\}_{i=1}^{m_t}$, m_t being the number of oracle calls. Next, we compute the unbiased estimate $G_t(x_t)$ of $g(x_t)$,

$$G_t(x_t)=\frac{1}{m_t}\sum_{i=1}^{m_t}\mathcal{G}(x_t,\xi_{t,i}).$$

We define filtration $\mathcal{F}_t=\sigma\big(x_0,\xi_{0,1},...\xi_{0,m_0},\xi_{1,1},...\xi_{1,m_1},...,\xi_{t,1},...\xi_{t,m_t}\big)$, so that

- random variables in $\{\xi_{t,i}\}_{i=1}^{m_t}$ are $\mathcal{F}_t$-measurable.
- search points with index t which are deterministic functions of $G_\tau(x_\tau),\tau\leq t-1$, are $\mathcal{F}_{t-1}$-measurable.

We use the shorthand notation $\mathbf{E}_{\lceil t\rceil}$ to denote the conditional expectation with respect to the filtration $\mathcal{F}_t$.

Under Assumption (SEN), we can establish sub-exponential bounds for the mini-batch operator $G_t(x_t)$.

Lemma 5.2.1 *Let Assumption (SEN) hold, then for* $\lambda\in\left[0,\frac{m_t}{2\sigma_t}\right]$,

$$\mathbf{E}_{\lceil t-1\rceil}\left[\exp(\lambda\|G_t(x_t)-g(x_t)\|_*)\right]\leq\exp\left(1.06\lambda\bar{\sigma}_t+\frac{3\lambda^2\overline{\sigma}_t^2}{m_t}\right) \tag{5.2.1}$$

almost surely and where

$$\bar{\sigma}_t^2 := \frac{\Omega}{m_t}\left\{\mathcal{L}[f(x_t) - f(x^*) - \langle g(x^*), x_t - x^*\rangle] + \sigma_*^2\right\}.$$

Consequently, for all $m_t \geq 1$,

$$\mathbf{E}_{\lceil t-1\rceil}\left\{\exp\left(\frac{\|G_t(x_t) - g(x_t)\|_*}{3\bar{\sigma}_t}\right)\right\} \leq \exp(1), \tag{5.2.2}$$

almost surely.

5.2.3 A preliminary large deviation bound

Notice that we have proved in Lemma 5.2.1 that $\delta_t := G_t - g(x_t)$ follows a sub-exponential distribution with a variance bounded by $\mathcal{O}(\bar{\sigma}_t^2)$. However, the challenge arises from the fact that the variance is a $\mathcal{F}_{t-1}$-measurable random variable, which makes it difficult to establish sharp large deviation bounds. A straightforward approach to handle the random variance is to apply a fixed upper bound, such as $\bar{\sigma}_t \leq \widehat{\sigma}$, and then utilize Bernstein's inequality. However, this approach fails to capture the reduction in variance as x_t converges to x^*. Therefore, it is necessary to have a "data-driven" bound that explicitly depends on the random variance of the noise. Towards this end, we first prove a preliminary result for handling a sequence of sub-exponential random variables with $\mathcal{F}_{t-1}$-measurable variance. We believe that this result is of independent interest to the statistics and optimization community, so we present it in the following lemma using general notation.

Lemma 5.2.2 *Let* $\mathfrak{y}_t, t = 0, 1, 2, ...$ *be a sequence of* $\mathcal{F}_t$*-measurable random variables such that for some* $v > 0$ *and* $\lambda \in [0, (v\mathfrak{s}_t)^{-1}]$,

$$\mathbf{E}_{\lceil t-1\rceil}\{e^{\lambda\mathfrak{y}_t}\} \leq \exp\{\lambda\mathfrak{r}_t + \tfrac{1}{2}\lambda^2\mathfrak{s}_t^2\}, \tag{5.2.3}$$

where $\mathfrak{s}_t$ *and* $\mathfrak{r}_t$ *are* $\mathcal{F}_{t-1}$*-measurable. Let us set* $\mathfrak{Y}_j := \sum_{t=0}^{j}(\mathfrak{y}_t - \mathfrak{r}_t)$, $\mathfrak{u}_j^2 := \sum_{t=0}^{j}\mathfrak{s}_t^2$, *and* $\widehat{\mathfrak{s}}_j := \max_{0\leq i\leq j}\mathfrak{s}_i$. *Then we have for* $y, \underline{s}, \bar{s}, \bar{u} > 0$,

$$\begin{aligned}&\mathrm{Prob}\left\{\exists\ t \leq k,\ s.t.\ \mathfrak{Y}_t \geq 2\mathfrak{u}_t\sqrt{2y} + 4v(\widehat{\mathfrak{s}}_t + \underline{s})y,\ \widehat{\mathfrak{s}}_t \leq \bar{s},\ \mathfrak{u}_t \leq \bar{u}\right\}\\ &\leq \left[\left(\log_2\left(\tfrac{\bar{u}}{\sqrt{2y}v\underline{s}}\right) + 1\right)\left(\log_2\left(\tfrac{\bar{s}}{\underline{s}}\right) + 1\right) + 1\right]e^{-y}.\end{aligned}$$

Consequently, if $\widehat{\mathfrak{s}}_k \leq \bar{s}$ *and* $\mathfrak{u}_k \leq \bar{u}$ *almost surely, we have*

$$\mathrm{Prob}\left\{\mathfrak{Y}_k \geq 2\mathfrak{u}_k\sqrt{2y} + 4v(\widehat{\mathfrak{s}}_k + \underline{s})y\right\} \leq \left[\left(\log_2\left(\tfrac{\bar{u}}{\sqrt{2y}v\underline{s}}\right) + 1\right)\left(\log_2\left(\tfrac{\bar{s}}{\underline{s}}\right) + 1\right) + 1\right]e^{-y}. \tag{5.2.4}$$

5.3 High-probability bounds for SGE

In this section we provide high-probability guarantees for the Stochastic Gradient Extrapolation algorithm introduced in the previous chapter. We derive high-probability bounds under the sub-exponential tail Assumption (SEN) by using the result presented in Lemma 5.2.2.

5.3.1 General smooth convex problem

Before diving into the development of the high-probability convergence guarantees, we first present an important result verified by the SGE method that will be used in the subsequent analysis. Recall that in the preceding chapter, we provide in Theorem 4.4.1 a result on the convergence of the SGE algorithm when the feasible region X is unbounded. Next proposition adapts this result to the setting when X is bounded.

Proposition 5.3.1 *If the algorithm parameters $\{\alpha_t\}$, $\{\eta_t\}$ and $\{\tau_t\}$ in Algorithm 4 satisfy $\tau_1 \geq 0$, $\tau_t > 0$ for all $t \geq 2$, and for some $\theta_t \geq 0$,*

$$\theta_{t-1} = \alpha_t \theta_t, \ \eta_t \leq \alpha_t \eta_{t-1}, \ t = 2, \ldots, k \tag{5.3.1}$$

$$\frac{\eta_t \tau_{t-1}}{\alpha_t} \geq 5L, \quad t = 3, \ldots, k \tag{5.3.2}$$

$$\frac{\eta_1 \eta_2}{\alpha_2} \geq 25L^2, \ \eta_k \tau_k \geq L, \tag{5.3.3}$$

then for a bounded feasible region X, i.e, $\max_{x,x' \in X} \|x - x'\| \leq R_X$, we have for any $x \in X$,

$$\sum_{t=1}^{k} \theta_t \left\{ \tau_t f(x_t) - \langle g(x_t), x - x_t \rangle - \tau_t f(x_{t-1}) \right\} \leq \theta_1 \eta_1 V(x_0, x) + \sum_{t=0}^{k-1} \chi_t(x), \tag{5.3.4}$$

where $\delta_t := G_t - g(x_t)$, $q_t := \frac{\theta_{t+1}(1+\alpha_{t+1}^2)}{\eta_{t+1}} + \frac{\theta_{t+2}\alpha_{t+2}^2}{\eta_{t+2}} + \frac{\theta_{t+2}^2 \alpha_{t+2}^2}{\theta_{t+1}\eta_{t+1}}$, $p_t := \begin{cases} \theta_t, & t \leq k-2 \\ \theta_{t+1} + \theta_t, & t = k-1 \end{cases}$, *and*

$$\chi_t(x) := \min \left\{ p_t \langle \delta_t, x - z_t \rangle + \frac{5q_t}{2} \|\delta_t\|_*^2, \ (2\theta_{t+1} + \theta_t) \|\delta_t\|_* R_X \right\}. \tag{5.3.5}$$

Proof of Proposition 5.3.1 The proof of the proposition follows exactly the same steps as in the proof of Theorem 4.4.1, except that when X is bounded, we can upper bound $\Delta_{k,2}$ appearing in Ineq. directly by using Cauchy-Schwarz inequality and $\delta_0 = \delta_{-1}$, we obtain

$$\Delta_{k,2} \leq \sum_{t=1}^{k} \theta_t [(1 + \alpha_t) \|\delta_{t-1}\|_* + \alpha_t \|\delta_{t-2}\|_*] R_X.$$

□

Given Lemma 5.2.2, let us now focus on the high-probability convergence results for the SGE method. Recalling Ineq (5.3.4) in Lemma 4.9.2, we need to properly bound the term $\sum_{t=0}^{k-1} \chi_t(x^*)$, where

$$\chi_t(x^*) := \min \left\{ p_t \langle \delta_t, x^* - z_t \rangle + \frac{5q_t}{2} \|\delta_t\|_*^2, \ (2\theta_{t+1} + \theta_t) \|\delta_t\|_* R_X \right\}, \text{ and } p_t := \begin{cases} \theta_t, & t \leq k-2 \\ \theta_{t+1} + \theta_t, & t = k-1 \end{cases}.$$

We have the following lemma which characterizes the sub-exponential distribution of $\chi_t(x^*)$.

Lemma 5.3.1 *Under condition (SEN), we have for $\lambda \in [0, \frac{1}{6\bar{\sigma}_t(2\theta_{t+1}+\theta_t)R_X}]$,*

$$\mathbf{E}_{\lceil t-1 \rceil} [\exp(\lambda \cdot \chi_t(x^*))] \leq \exp \left\{ 2.8 \ \lambda q_t \bar{\sigma}_t^2 + 27\lambda^2 (2\theta_{t+1} + \theta_t)^2 R_X^2 \bar{\sigma}_t^2 \right\}. \tag{5.3.6}$$

By applying Lemma 5.2.2 to upper bound the term $\sum_{t=0}^{k-1}\chi_t(x^*)$ in Proposition 5.3.1, we can derive the following theorem that characterizes the convergence of the SGE method with high-probability.

Theorem 5.3.1 *Suppose that the algorithm parameters $\{\theta_t\}$, $\{\alpha_t\}$, $\{\eta_t\}$ and $\{\tau_t\}$ in Algorithm 4 satisfy (4.9.16a)-(4.9.16c), $\tau_1 = 0$ and*

$$\theta_t\tau_t + \mathcal{L}\epsilon_{t-1} \le \theta_{t-1}(1+\tau_{t-1}),\ t \ge 2, \tag{5.3.7}$$

where

$$\epsilon_t := \frac{2.8\ q_t\Omega}{m_t} \tag{5.3.8}$$

Also assume that Assumption (SEN) holds and that the feasible region X is bounded, i.e., $\max_{x,x'\in X}\|x-x'\|^2 \le R_X^2$. For $\delta\in(0,1)$ and $k\in\mathbf{Z}_+$, we denote

$$\varsigma_t := \frac{(2\theta_{t+1}+\theta_t)^2R_X^2\Omega}{m_t},\quad \text{and}\quad \widehat{\delta} := \ln\left\{\frac{\left(\frac{1}{2}\log_2(k\cdot\frac{\mathcal{L}LR_X^2/2+\sigma_*^2}{\sigma_*^2})+1\right)^2+1}{\delta}\right\}. \tag{5.3.9}$$

Then, with probability at least $1-\delta$,

$$\begin{aligned}
&\theta_k(1+\tau_k)[f(x_k)-f(x^*)]\\
&\le \theta_1\eta_1V(x_0,x^*) + \frac{\mathcal{L}\epsilon_0L}{2}\|x_0-x^*\|^2 + 35\widehat{\delta}\sqrt{\frac{\mathcal{L}\varsigma_0L\|x_0-x^*\|^2}{2}}\\
&\quad + \sum_{t=0}^{k-1}\epsilon_t\sigma_*^2 + 6\sqrt{\sum_{t=0}^{k-1}6\varsigma_t\sigma_*^2\widehat{\delta}} + 48\widehat{\delta}\max_{0\le t\le k-1}\sqrt{\varsigma_t\sigma_*^2} + \frac{307\widehat{\delta}^2\sum_{t=1}^{k-1}\varsigma_t\mathcal{L}[f(x_t)-f(x^*)]}{\sum_{t=1}^{k-1}[\theta_t(1+\tau_t)-\theta_{t+1}\tau_{t+1}-\epsilon_t\mathcal{L}][f(x_t)-f(x^*)]}.
\end{aligned}$$

We now specify a particular stepsize policy with the convergence guarantee in high-probability. In fact, this stepsize policy is nearly the same as the stepsize policy introduced in the expectation bound, i.e., Corollary 4.4.1. The only difference is that the batch size m_t in the following corollary depends explicitly on the variable $\widehat{\delta}\sim\mathcal{O}(\ln(1/\delta))$.

Corollary 5.3.1 *Set*

$$\theta_t = t,\ \alpha_t = \frac{t-1}{t},\ \text{and}\quad \eta_t = \frac{24L}{t}.$$

Also let us set $\tau_1 = 0$, $\tau_t = \frac{t-1}{2} - \frac{t}{24}$, $t\ge 2$, and

$$m_t = \max\left\{1, \left\lceil\frac{216\mathcal{L}(t+2)(\widehat{\delta}^2+\Omega)}{L}\right\rceil, \left\lceil\frac{5(k+1)^3(\widehat{\delta}+\Omega)\sigma_*^2}{L^2\Omega R_X^2}\right\rceil\right\},\quad t\ge 0.$$

For $\delta\in(0,1)$, we define $\widehat{\delta} := \ln\left\{\frac{\left(\frac{1}{2}\log_2(k\cdot\frac{\mathcal{L}LR_X^2/2+\sigma_*^2}{\sigma_*^2})+1\right)^2+1}{\delta}\right\}$. Then for any $k\ge\widehat{\delta}$, with probability at least $1-\delta$,

$$f(x_k)-f(x^*) \le \frac{797\Omega LR_X^2}{k(k+1)}.$$

Notice that $\widehat{\delta} = \mathcal{O}\{\ln(1/\delta)\}$ if we neglect terms with log(log) dependence on the problem parameters. Given Corollary 5.3.1, and assuming $k \geq \widehat{\delta}$, the total number of iterations performed by the SGE method to find an (ϵ, δ)-optimal solution, i.e., $\bar{x} \in X$ such that $f(\bar{x}) - f(x^*) \leq \epsilon$ with probability at least $1 - \delta$, is bounded by $\mathcal{O}\left(\frac{\sqrt{L\Omega R_X^2}}{\sqrt{\epsilon}}\right)$. The dependence on Ω arises from applying the inequality $V(x, x') \leq \frac{\Omega}{2}\|x - x'\|^2 \leq \frac{\Omega R_X^2}{2}$. Consequently, the overall sample complexity of the SGE method to achieve an (ϵ, δ)-optimal solution is bounded by

$$\mathcal{O}\left\{\sqrt{\frac{L\Omega R_X^2}{\epsilon}} + \frac{\mathcal{L}\Omega[\Omega + \ln^2(1/\delta)]R_X^2}{\epsilon} + \frac{\Omega[\Omega + \ln(1/\delta)]R_X^2\sigma_*^2}{\epsilon^2}\right\}. \tag{5.3.10}$$

It is important to note that the extra dependence on Ω in the second and third terms comes from the application the mini-batch operator, as stated in Lemma 4.2.2.

Comparing with the sample complexity under expectation in Eq. (4.4.13), Eq. (5.3.10) has additional logarithmic dependencies on the probability of confidence δ. However, the dependence on $\ln(1/\delta)$ appears to be a combination of multiplication and summation, rather than simple multiplicative factors. Specifically, the first term (deterministic error) in (5.3.10) is the same as its analog in (4.4.13). The second and third terms in (5.3.10) reduce to their counterparts in (4.4.13) when $\ln(1/\delta) \leq \mathcal{O}(\sqrt{\Omega})$ and $\ln(1/\delta) \leq \mathcal{O}(\Omega)$, respectively.

5.3.2 Smooth convex problem with quadratic growth condition

In this section, we consider the case when the smooth convex objective function f also satifies the μ-quadratic growth condition, i.e., when for some $\mu > 0$ and $x^* \in X$

$$f(x) - f(x^*) \geq \frac{\mu}{2}\|x - x^*\|, \quad \forall x \in X. \tag{5.3.11}$$

In order to generate the high probability convergence guarantee, we introduce a shrinking multi-stage SGE method, which modifies Algorithm 5 by shrinking the feasible region in each stage.

Algorithm 7 Shrinking multi-stage Stochastic Gradient Extrapolation method

Input: initial point $\widetilde{x}_0 \in X$, accuracy parameter $\delta \in (0, 1)$. Assume $\|\widetilde{x}_0 - x^*\| \leq R_0^2$.
for $k = 1, 2, \ldots, K$ **do**
(a) Run $\overline{N}$ iterations of the SGE method (Algorithm 4) with feasible region

$$X_k := \left\{x \in X : \|\widetilde{x}_{k-1} - x\|^2 \leq R_{k-1}^2 := R_0^2 \cdot 2^{-k+1}\right\}$$

The algorithm parameters are set as: $x_0 = \bar{x}_0 = \widetilde{x}_{k-1}$ and

$$\begin{aligned}
&\theta_t = t,\ \alpha_t = \tfrac{t-1}{t},\ \eta_t = \tfrac{24L}{t},\ t = 1, ..., \overline{N}\\
&m_t^{(k)} = \max\{1, \lceil\tfrac{216\mathcal{L}(t+2)(\widehat{\delta}_K^2+\Omega)}{L}\rceil, \lceil\tfrac{5(\overline{N}+1)^3(\widehat{\delta}_K^2+\Omega)\sigma_*^2}{L^2\Omega D_0^2\cdot 2^{-k+1}}\rceil\},\ t = 0, ..., \overline{N}\\
&\tau_1 = 0,\ \tau_t = \tfrac{t-1}{2} - \tfrac{t}{24},\ t = 2, ..., \overline{N}
\end{aligned}$$

and

$$\overline{N} = \tfrac{57\sqrt{\Omega L}}{\sqrt{\mu}}. \tag{5.3.12}$$

(b) Set $\widetilde{x}_k = x_{\overline{N}}$, where $x_{\overline{N}}$ is the solution obtained in Step (a).
end for

Notice that the confidence level parameter $\widehat{\delta}_K$ is defined as

$$\widehat{\delta}_K := \ln\left\{\frac{K\left(\frac{1}{2}\log_2(\bar{N}\frac{\mathcal{L}LR_X^2/2+\sigma_*^2}{\sigma_*^2})+1\right)^2+K}{\delta}\right\},$$

which is different from the definition in Corollary 5.3.1 because of the uniform probability taken over the stages. The following corollary characterizes the convergence property of the shrinking multi-stage SGE method in high probability.

Corollary 5.3.2 *Let $\{\widetilde{x}_K\}$ be computed by Algorithm 7. Assume $\|\widetilde{x}_0 - x^*\|^2 \le R_0^2$. Then we have with probability greater than $1-\delta$,*

$$f(\widetilde{x}_K) - f(x^*) \le \mu R_0^2 \cdot 2^{-K-1} \quad \text{and} \quad \|\widetilde{x}_K - x^*\|^2 \le R_0^2 \cdot 2^{-K}.$$

In view of Corollary 5.3.2, the number of stages required by the multi-stage SGE method to find a solution $\bar{x} \in X$ such that $f(\bar{x}) - f(x^*) \le \epsilon$ with probability greater than $1-\delta$ is bounded by $\mathcal{O}\left(\ln(\frac{\mu R_0^2}{\epsilon})\right)$. Considering the iteration number in each stage, the algorithm achieves an iteration complexity of $\mathcal{O}\left(\sqrt{\frac{L\Omega}{\mu}}\ln(\frac{\mu R_0^2}{\epsilon})\right)$. By further considering the batch size, the total number of samples $\sum_{k=1}^{K}\sum_{t=1}^{\overline{N}} m_t^{(k)}$ is bounded by

$$\mathcal{O}\left\{\sqrt{\frac{L\Omega}{\mu}}\ln\left(\frac{\mu R_0^2}{\epsilon}\right) + \frac{\mathcal{L}\Omega(\Omega+\widehat{\delta}_K^2)}{\mu}\ln\left(\frac{\mu R_0^2}{\epsilon}\right) + \frac{\Omega(\Omega+\widehat{\delta}_K^2)\sigma_*^2}{\mu\epsilon}\right\}.$$

Similarly, we can obtain the iteration complexity for finding a solution $\bar{x} \in X$ such that $\|\bar{x} - x^*\|^2 \le \epsilon$ with high-probability $1-\delta$ by $\mathcal{O}\left(\sqrt{\frac{L\Omega}{\mu}}\ln(\frac{R_0^2}{\epsilon})\right)$. Consequently, the sample complexity is bounded by

$$\mathcal{O}\left\{\sqrt{\frac{L\Omega}{\mu}}\ln\left(\frac{R_0^2}{\epsilon}\right) + \frac{\mathcal{L}\Omega(\Omega+\widehat{\delta}_K^2)}{\mu}\ln\left(\frac{R_0^2}{\epsilon}\right) + \frac{\Omega(\Omega+\widehat{\delta}_K^2)\sigma_*^2}{\mu^2\epsilon}\right\}.$$

Clearly, the above iteration complexities are optimal. To the best of our knowledge, this is the first high probability bound for strongly convex or quadratic growth stochastic optimization problem with sub-exponential state-dependent noise.

5.4 Accuracy certificate

In this section we derive a procedure to estimate in an online fashion, i.e., while running the SGE algorithm, a computationable accuracy certificate for the suboptimality gap. We will see in the development of this section that this procedure provides a practical termination criterion for the latter stochastic method.

As a first step, we provide the following proposition, which is similar in the idea to Proposition 4.9.2 and Proposition 5.3.1 but has a telescoping sum from t_0 to k, where $t_0 \in \mathbf{Z}_+$.

Proposition 5.4.1 *Suppose that the algorithm parameters $\{\theta_t\}$, $\{\alpha_t\}$, $\{\eta_t\}$ and $\{\tau_t\}$ in Algorithm 4 satisfy (4.9.16a)-(4.9.16c). Also assume that the feasible region X is bounded, i.e.,* $\max_{x,x'\in X}\|x -$

$x'\| \leq R_X$. *Then for any* $t_0 \geq 2$ *and* $k > t_0$, *we have*

$$\begin{aligned}
&\sum_{t=t_0}^{k} \theta_t \langle G_t, x_t - x\rangle + \theta_k \tau_k [f(x_k) - f(x^*)] \\
&\leq \theta_{t_0}\eta_{t_0} V(z_{t_0-1}, x) + \theta_{t_0}\eta_{t_0} R_X^2 + \frac{\theta_k \eta_k}{5} R_X^2 + \theta_{t_0}\tau_{t_0}[f(x_{t_0-1}) - f(x^*)] \\
&\quad + \sum_{t=t_0}^{k-1} (\theta_{t+1}\tau_{t+1} - \theta_t \tau_t)[f(x_t) - f(x^*)] + \sum_{t=t_0-2}^{k} \widetilde{\chi}_t, \qquad (5.4.1)
\end{aligned}$$

where

$$\widetilde{\chi}_t := \min\left\{ \widetilde{p}_t \langle \delta_t, x_t - z_t\rangle + \frac{5q_t}{2}\|\delta_t\|_*^2,\ (2\theta_t + 2\theta_{t+1})\|\delta_t\|_* R_X \right\}, \qquad (5.4.2)$$

and $\widetilde{p}_t := \begin{cases} \theta_t, & t_0 \leq t \leq k-1 \\ 0, & \text{otherwise} \end{cases}$, *and* $q_t := \frac{\theta_{t+1}(2+\alpha_{t+1}^2)}{\eta_{t+1}} + \frac{\theta_{t+2}\alpha_{t+2}^2}{\eta_{t+2}}$.

Next, by choosing the parameters of the SGE algorithm as stated in Corollary 5.3.1 we obtain the following result.

Corollary 5.4.1 *Set the parameters as in Corollary 5.3.1 with*

$$\widetilde{\delta} := \ln\left\{ k \frac{\left(\frac{1}{2}\log_2(k \cdot \frac{\mathcal{L}LR_X^2/2 + \sigma_*^2}{\sigma_*^2}) + 1\right)^2 + 1}{\delta} \right\}.$$

Then for any $k \geq \widetilde{\delta}$, *with probability at least* $1 - \delta$, *we have*

$$\max_{x \in X}\left\{ \sum_{t=\lceil k/2\rceil}^{k} t\langle G_t, x_t - x\rangle \right\} \leq 1207 L\Omega R_X^2.$$

An immediate consequence of the latter corollary is that while the SGE algorithm is running, it is possible to compute certificates that provide upper bounds on the suboptimality gap while having the optimal convergence rate. To that matter, we use a particular averaging of the iterates output by the SGE algorithm. Indeed, consider $k \in \mathbf{Z}_+$ such that $k \geq \widetilde{\delta}$ and let us denote by

$$\overline{x}_k := \frac{\sum_{t=\lceil k/2\rceil}^{k} t\ x_t}{\sum_{t=\lceil k/2\rceil}^{k} t} = \frac{2}{\left(\lceil \frac{k}{2}\rceil + k\right)\left(k - \lceil \frac{k}{2}\rceil + 1\right)} \sum_{t=\lceil k/2\rceil}^{k} t\ x_t.$$

Now observe that since the objective function f is convex and that

$$\frac{2}{\left(\lceil \frac{k}{2}\rceil + k\right)\left(k - \lceil \frac{k}{2}\rceil + 1\right)} \sum_{t=\lceil k/2\rceil}^{k} t = 1,$$

we have

$$\begin{aligned} f(\overline{x}_k) - f(x^*) &\leq \frac{2}{\left(\lceil \frac{k}{2} \rceil + k\right)\left(k - \lceil \frac{k}{2} \rceil + 1\right)} \sum_{t=\lceil k/2 \rceil}^{k} t\left(f(x_t) - f(x^*)\right) \\ &\leq \frac{2}{\left(\lceil \frac{k}{2} \rceil + k\right)\left(k - \lceil \frac{k}{2} \rceil + 1\right)} \left[\sum_{t=\lceil k/2 \rceil}^{k} t\langle g_t, x_t - x^* \rangle\right] \\ &= \frac{2}{\left(\lceil \frac{k}{2} \rceil + k\right)\left(k - \lceil \frac{k}{2} \rceil + 1\right)} \left[\sum_{t=\lceil k/2 \rceil}^{k} t\langle G_t + \delta_t, x_t - x^* \rangle\right] \\ &\lesssim \frac{1}{k^2}\left(\max_{x \in X}\left\{\sum_{t=\lceil k/2 \rceil}^{k} t\langle G_t, x_t - x\rangle\right\} + \sum_{t=\lceil k/2 \rceil}^{k} t\langle \delta_t, x_t - x^* \rangle\right) \end{aligned} \tag{5.4.3}$$

Observe that the first term of Ineq. (5.4.3) can be computed while the SGE algorithm is running. Indeed, computing the first term boils down to solving a simple linear programming problem. Additionally, the latter term can be computed recursively by keeping track of the two quantities,

$$\sum_{t=\lceil k/2 \rceil}^{k} \langle G_t, x_t \rangle, \text{ and } \sum_{t=\lceil k/2 \rceil}^{k} t\ G_t.$$

The accuracy certificate is motivated both by the fact that the first term in Ineq. (5.4.3) can be computed "on-the-fly" but also by the fact that it is possible to provide a better bound for the second martingale term than the bound provided in Corollary 5.3.1, as shown in the next Corollary.

Corollary 5.4.2 *Let the stepsize parameters as in Corollary 5.3.1 with a batch size*

$$m_t = \max\left\{1, \left\lceil \frac{216\mathcal{L}(t+2)(\widehat{\delta}^2 + \Omega)}{L} \right\rceil, \left\lceil \frac{5(k+1)^3 \widehat{\delta}\sigma_*^2}{L^2 R_X^2} \right\rceil\right\}, \quad t \geq 0. \tag{5.4.4}$$

Then for any $k \geq \widehat{\delta}$, with probability at least $1 - \delta$, we have

$$f(\overline{x}_k) - f(x^*) \leq \frac{16}{3k^2} \cdot \max_{x \in X}\left\{\sum_{t=\lceil k/2 \rceil}^{k} t\langle G_t, x_t - x\rangle\right\} + \frac{203 L R_X^2}{k^2}. \tag{5.4.5}$$

where

$$\overline{x}_k = \frac{\sum_{t=\lceil k/2 \rceil}^{k} t\ x_t}{\sum_{t=\lceil k/2 \rceil}^{k} t}.$$

To obtain the termination criterion, we fix a desired inacuracy level $\epsilon > 0$ and then we use the upper bound for the second martingale term in (5.4.5) to determine the number of iterations needed to make this term smaller than $\epsilon/2$. Specifically, we choose $K_1 \in \mathbf{Z}_+$ such that

$$K_1 = \left\lceil \sqrt{\frac{406 L R_X^2}{\epsilon}} \right\rceil.$$

This being decided, we can now run the SGE algorithm and keep recursively track of the first quantity appearing in (5.4.5). We stop the algorithm at time $K = \max\{K_1, K_2\}$, where $K_2 \in \mathbf{Z}_+$ is such that

$$\frac{16}{3K_2^2} \cdot \max_{x \in X} \left\{ \sum_{t=\lceil K_2/2 \rceil}^{K_2} t\langle G_t, x_t - x\rangle \right\} \leq \frac{\epsilon}{2}.$$

After K iterations we finally have

$$f(\overline{x}_K) - f(x^*) \leq \epsilon.$$

5.5 A Multistage Accelerated Algorithm for Sparse Recovery

In this section we aim at solving the following sparse estimation problem

$$x^* \in \operatorname*{argmin}_{x \in X} \; f(x), \tag{5.5.1}$$

where x^* is assumed to be s-sparse, i.e., with at most s-nonvanishing components.

In the optimization literature, addressing sparsity is often approached using "composite" techniques. Accordingly, we will reformulate Problem (5.5.1) into a composite optimization problem. Consider the following auxiliary composite problem:

$$\min_{x \in X} \left\{ \Psi(x) := \tfrac{1}{2} f(x) + h(x) \right\}, \tag{5.5.2}$$

where h is a positive convex function and f is the initial smooth convex objective function. In the subsequent development we introduce a non-Euclidean composite version of the SGE algorithm. The latter algorithm is used to control the gap $\Psi(\widehat{x}) - \Psi(x^*)$ and it does not minimize the composite function Ψ. We rather use the composite function and more specifically the penalty function h to infer some sparse structure on the optimum of the main objective function x^*.

5.5.1 Local Proximal Setup.

Here we introduce the local proximal setup which will be the foundation to build the main algorithm of this work. Consider $B := \{x \in E : \|x\| \leq 1\}$, the unit ball associated to the general norm $\|\cdot\|$, and $\omega : B \to \mathbf{R}$ a *distance-generating function* (d.-g.f.) on B, i.e., a continuously differentiable convex function that is also strongly convex w.r.t. the norm $\|\cdot\|$,

$$\langle \nabla\omega(x) - \nabla\omega(y), x - y\rangle \geq \|x - y\|^2, \quad \forall x, y \in X.$$

We assume w.l.o.g. that $\omega(x) \geq \omega(0) = 0$ and Ω is still such that $\Omega = \max_{\|z\| \leq 1} \omega(z)$.

Instead of using function ω, we will consider its local renormalized version defined on $X_R(x_0) := X \cap B(x_0, R) = \{x \in X : \|x - x_0\| \leq R\}$, for some $x_0 \in X$ and $R > 0$, by

$$\widetilde{\omega}_{x_0}^R(z) := R^2 \omega\left(\frac{z - x_0}{R}\right).$$

Observe that $\widetilde{\omega}_{x_0}^R(\cdot)$ is 1-strongly convex on $X_R(x_0)$ and we have $\widetilde{\omega}_{x_0}^R(x_0) = 0$, and $\widetilde{\omega}_{x_0}^R(z) \leq \Omega R^2$.

For a given initialization point $x_0 \in X$ and $R > 0$, we can also define the local Bregman Divergence associated to the d.-g.f. $\widetilde{\omega}_{x_0}^R(\cdot)$ such as

$$\forall x, y \in X_R(x_0), \quad V_{x_0}^R(x, y) := \widetilde{\omega}_{x_0}^R(y) - \widetilde{\omega}_{x_0}^R(x) - \langle \nabla\widetilde{\omega}_{x_0}^R(x), y - x\rangle.$$

For any $y, x, x_0 \in X$, we have

$$V_{x_0}^R(x_0, y) \leq \frac{\Omega}{2}\|y - x_0\|^2 \quad \text{and} \quad V_{x_0}^R(x, y) \geq \frac{1}{2}\|x - y\|^2. \tag{5.5.3}$$

5.5.2 Composite Stochastic Gradient Extrapolation (CSGE).

Below we present the CSGE algorithm, it is the work horse for our main multistage routine.

Algorithm 8 Composite Stochastic Gradient Extrapolation method (CSGE)

Input: initial point $x_0 = z_0$, nonnegative parameters $\{\alpha_t\}$, $\{\eta_t\}$ and $\{\beta_t\}$, batch size $\{m_t\}$ and finite horizon k.
$G_0 = G_{-1} = \frac{1}{m_0}\sum_{i=1}^{m_0} \mathcal{G}(x_0, \xi_{0,i})$.
for $t = 1, 2, \ldots, k$ **do**

$$\widetilde{G}_t = G_{t-1} + \alpha_t(G_{t-1} - G_{t-2}), \tag{5.5.4a}$$
$$z_t = \operatorname*{argmin}_{x\in X}\{\langle \widetilde{G}_t, x\rangle + h(x) + \eta_t V_{x_0}^R(z_{t-1}, x)\}, \tag{5.5.4b}$$
$$x_t = (1-\beta_t)x_{t-1} + \beta_t z_t, \tag{5.5.4c}$$
$$G_t = \tfrac{1}{m_t}\textstyle\sum_{i=1}^{m_t} \mathcal{G}(x_t, \xi_{t,i}).$$

end for
Output: $\widehat{x}_k = \frac{\theta_k(1+\tau_k)x_k + \sum_{t=1}^{k-1}[\theta_t(1+\tau_t) - \theta_{t+1}\tau_{t+1}]x_t}{\theta_k(1+\tau_k) + \sum_{t=1}^{k-1}\theta_t(1+\tau_t) - \theta_{t+1}\tau_{t+1}}$

In the next proposition, we present an important result verified by the estimates provided by Algorithm 8. For the sake of compactness of the following proposition we introduce the function $\widetilde{\Psi} := f + h$.

Proposition 5.5.1 *For $0 \le \beta_t \le 1$, denote $\tau_t = \frac{1-\beta_t}{\beta_t}$. We consider the following relations*

$$\theta_{t-1} = \alpha_t\theta_t,\ \eta_t \le \alpha_t\eta_{t-1},\ t = 2, \ldots, k \tag{5.5.5a}$$
$$\frac{\eta_t\tau_{t-1}}{\alpha_t} \ge 5L, \quad t = 3, \ldots, k \tag{5.5.5b}$$
$$\frac{\eta_1\eta_2}{\alpha_2} \ge 25L^2,\ \eta_k\tau_k \ge L. \tag{5.5.5c}$$

If the algorithm parameters $\{\alpha_t\}$, $\{\eta_t\}$ and $\{\tau_t\}$ satisfy $\tau_1 \ge 0$, $\tau_t > 0$ for all $t \ge 2$, along with relations (5.5.5a) - (5.5.5c) for some $\theta_t \ge 0$. Then for all $x \in X$

$$\sum_{t=1}^{k}\theta_t\left\{(1+\tau_t)\widetilde{\Psi}(x_t) - \widetilde{\Psi}(x) - \tau_t\widetilde{\Psi}(x_{t-1})\right\} \le \theta_1\eta_1 V(x_0, x) + \sum_{t=0}^{k-1}\chi_t(x) \tag{5.5.6}$$

where

$$\chi_t(x) := \min\left\{p_t\langle\delta_t, x - z_t\rangle + \tfrac{5q_t}{2}\|\delta_t\|_*^2,\ (2\theta_{t+1} + \theta_t)\|\delta_t\|_* R_X\right\}. \tag{5.5.7}$$

with

$$p_t := \theta_t\mathbf{1}\{t \le k-2\} + (\theta_t + \theta_{t+1})\mathbf{1}\{t = k-1\} \quad \textit{and} \quad \delta_t := G_t - g(x_t).$$

By applying Lemma 5.2.2 to upper bound the term $\sum_{t=0}^{k-1}\chi_t(x^*)$ in Proposition 5.5.1, we can derive the following theorem that characterizes the convergence of the CSGE method with high-probability.

Now we can state the main result of this section

Theorem 5.5.1 *Suppose that the algorithm parameters $\{\theta_t\}$, $\{\alpha_t\}$, $\{\eta_t\}$ and $\{\tau_t\}$ in Algorithm 8 satisfy (5.5.5a)-(5.5.5c),*

$$\tau_1 = 0 \text{ and } \forall t \geq 1, \quad \tfrac{1}{2}(\theta_{t+1}\tau_{t+1} - \theta_t(1+\tau_t)) + \epsilon_t \mathcal{L} < 0. \tag{5.5.8}$$

Also assume that Assumption (SEN) holds and that the feasible region X is bounded, i.e., $\max_{x,x' \in X} \|x - x'\|^2 \leq R_X^2$. For $\delta \in (0,1)$ and $k \in \mathbf{Z}_+^*$, we denote

$$\epsilon_t := \frac{2.8\ q_t \Omega}{m_t}, \quad \varsigma_t := \frac{(2\theta_{t+1} + \theta_t)^2 R_X^2 \Omega}{m_t}, \quad \text{and} \quad \widehat{\delta} := \ln\Big\{ \frac{\left(\frac{1}{2}\log_2(k \cdot \frac{\mathcal{L}LR_X^2/2 + \sigma_*^2}{\sigma_*^2}) + 1\right)^2 + 1}{\delta} \Big\}. \tag{5.5.9}$$

Then with probability at least $1 - \delta$ we have

$$\begin{aligned} \Psi(\widehat{x}_k) - \Psi^* \leq & \Gamma_k^{-1} \Bigg[\frac{\theta_1 \eta_1 \Omega + \epsilon_0 \mathcal{L} L}{2} R_X^2 + 35\widehat{\delta} \sqrt{\frac{\varsigma_0 \mathcal{L} L R_X^2}{2}} + 6\sigma_* \sqrt{6\widehat{\delta}} \sqrt{\sum_{t=0}^{k-1} \varsigma_t} + 48\widehat{\delta} \max_{0 \leq t \leq k-1} \sqrt{\sigma_*^2 \varsigma_t} \\ & + \sigma_*^2 \sum_{t=0}^{k-1} \epsilon_t + \frac{(35\widehat{\delta})^2 \sum_{t=1}^{k-1} \varsigma_t \mathcal{L}[f(x_t) - f(x^*)]}{2\sum_{t=1}^{k-1} \big(\theta_t(1+\tau_t) - \theta_{t+1}\tau_{t+1} - \epsilon_t \mathcal{L}\big)[f(x_t) - f(x^*)]} \Bigg], \end{aligned} \tag{5.5.10}$$

where $\Gamma_k := \theta_k(1+\tau_k) + \sum_{t=1}^{k-1} \theta_t(1+\tau_t) - \theta_{t+1}\tau_{t+1}$.

Corollary 5.5.1 *Set*

$$\theta_t = t, \ \alpha_t = \frac{t-1}{t}, \ \text{and} \quad \eta_t = \frac{24L}{t}.$$

Also let us set $\tau_1 = 0$, $\tau_t = \frac{t-1}{2} - \frac{t}{24}$, $t \geq 2$, and

$$m_t = \max\left\{1, \left\lceil \frac{216\mathcal{L}(t+2)(\widehat{\delta}^2 + \Omega)}{L} \right\rceil, \left\lceil \frac{5(k+1)^3(\widehat{\delta} + \Omega)\sigma_*^2}{L^2 \Omega R_X^2} \right\rceil \right\}, \quad t \geq 0.$$

Consider $\widehat{x}_k$, the output of Algorithm 8, and for $\delta \in (0,1)$, we consider

$$\widehat{\delta} = \ln\left\{ \frac{\left(\frac{1}{2}\log_2(k \cdot \frac{\mathcal{L}LR_X^2/2 + \sigma_*^2}{\sigma_*^2}) + 1\right)^2 + 1}{\delta} \right\}.$$

Then for any $k \geq \widehat{\delta}$, with probability at least $1 - \delta$,

$$\Psi(\widehat{x}_k) - \Psi(x^*) \leq \frac{1811 \Omega L R_X^2}{k(k+1)}. \tag{5.5.11}$$

5.5.3 Composite Stochastic Gradient Extrapolation for Sparse Recovery

In this section we will present our main method for solving the original sparse stochastic optimization problem (4.1.1). We consider the penalization function of the form $h(x) = \kappa\|x\|$ with $\kappa \geq 0$. Composite problem (5.5.2) becomes

$$\min_{x \in X} \left\{ \Psi_\kappa(x) := \tfrac{1}{2} f(x) + \kappa\|x\| \right\}, \tag{5.5.12}$$

Our proposed algorithm is based on a multistage procedure where at the k-th stage we form an auxiliary composite problem of the form (5.5.12) defined by a triplet of parameters (x_0, R_k, κ_k).

Subsequently, we run Algorithm 8 for a fixed number of iterations T to form an output $\widehat{x}_T$ which will be set as the initialization point x_0 of the next stage. It is crucial to emphasize the fact that x^* is a sparse minimizer of the function f. This implies that our interest does not lie in approximating $\Psi_\kappa(x^*)$ through the approximate solution generated by the multistage procedure. Instead, we utilize the composite problem as an auxiliary tool to infer some structure to the approximate solution. To assess the quality of the solution approximation, which is provided by the main multistage procedure, we will use the *Reduced Strong Convexity* (RSC) assumption introduced in [1]. This assumption provides a quite simple framework for analyzing sparse recovery problems across various general sparsity structures. In the next section we discuss and study an example where the RSC assumption is verified.

Assumption 4 (Reduced Strong Convexity) *There exist some positive constants $\Upsilon, \rho < \infty$ such that for any feasible solution $\widehat{x} \in X$ to the composite problem (5.5.12) satisfying, with probability at least $1 - \varepsilon$,*

$$\Psi_\kappa(\widehat{x}) - \Psi_\kappa(x^*) \leq \Lambda,$$

it holds, with probability at least $1 - \varepsilon$, that

$$\|\widehat{x} - x_*\| \leq \Upsilon^{-1} \left[\rho s \kappa + \Lambda \kappa^{-1}\right]. \tag{5.5.13}$$

Given the number of stages K and the parameters of the problem along with the confidence parameter $\widehat{\delta}_K$, Algorithm 9 works as follows, at stage $k \in [K]$ we instantiate the CSGE method with a point x_0 and run the algorithm for a fixed number of iterations $\overline{N}$ to solve the composite problem defined as

$$\min_{x \in X_{R_{k-1}}(x_0)} \{\Psi_{\kappa_k}(x) := \tfrac{1}{2} f(x) + \kappa_k \|x\|\}.$$

This run of the CSGE-SR algorithm provides an estimate $x_{\overline{N}}$ which verifies the bound presented in (5.5.13). Finally, the parameters of the multistage method are appropriately chosen to halve the bound on $\|\widehat{x}_{\overline{N}} - x^*\|^2$ at the end of the stage and we set the initialization point of the following stage to the value of $\widehat{x}_{\overline{N}}$. We provide below the multistage method along with its parameter values.

Algorithm 9 Composite Stochastic Gradient Extrapolation method for Sparse Recovery (CSGE-SR)

Input: initial point $\widetilde{x}_0 \in X$, accuracy parameter $\delta \in (0,1)$. Assume $\|\widetilde{x}_0 - x^*\| \le R_0$.
for $k = 1, 2, \ldots, K$ **do**
(a) Run $\overline{N}$ iterations of the CSGE method (Algorithm 8) with feasible region

$$X_k := X_{R_{k-1}}(\widetilde{x}_{k-1}) = \left\{x \in X : \|\widetilde{x}_{k-1} - x\|^2 \le R_{k-1}^2 := R_0^2 \cdot 2^{-k+1}\right\}$$

The algorithm parameters are set as: $x_0 = \bar{x}_0 = \widetilde{x}_{k-1}$ and

$$\begin{aligned}
&\theta_t = t,\ \alpha_t = \tfrac{t-1}{t},\ \eta_t = \tfrac{24L}{t},\ t = 1, ..., \overline{N} \\
&m_t^{(k)} = \max\{1, \lceil \tfrac{216\mathcal{L}(t+2)(\widehat{\delta}_K^2 + \Omega)}{L} \rceil, \lceil \tfrac{5(\overline{N}+1)^3(\widehat{\delta}_K + \Omega)\sigma_*^2}{L^2 \Omega R_0^2 \cdot 2^{-k+1}} \rceil\},\ t = 0, ..., \overline{N} \\
&\tau_1 = 0,\ \tau_t = \tfrac{t-1}{2} - \tfrac{t}{24},\ t = 2, ..., \overline{N}
\end{aligned}$$

and

$$\overline{N} = \left\lceil \tfrac{121}{\Upsilon}\sqrt{\rho s \Omega L} \right\rceil,\ \kappa_k = R_{k-1}\sqrt{\tfrac{1811\Omega L}{\rho s \overline{N}(\overline{N}+1)}}. \tag{5.5.14}$$

(b) Set $\widetilde{x}_k = \widehat{x}_{\overline{N}}$, where $\widehat{x}_{\overline{N}}$ is the solution obtained in Step (a).
end for

The following theorem characterizes the convergence rate of the CSGE-SR method applied for solving the main sparse stochastic optimization problem.

Theorem 5.5.2 *Let $\{\widetilde{x}_K\}$ be computed by Algorithm 9 and assume that $\|\widetilde{x}_0 - x^*\|^2 \le R_0^2$. Then we have with probability greater than $1 - \delta$,*

$$f(\widetilde{x}_K) - f(x_*) \le \frac{\Upsilon(\Upsilon+1)}{\rho s} R_0^2 \cdot 2^{-K-1} \quad \text{and} \quad \|\widetilde{x}_K - x^*\|^2 \le R_0^2 \cdot 2^{-K}.$$

Remark. The result presented in the last theorem states that the CSGE-SR algorithm approach the s-sparse optimum x^* by an estimate $\widetilde{x}_K$ with probability at least $1 - \delta$, such that $\|\widehat{x} - x^*\| \le \epsilon$ for any $\epsilon \in (0, R_0)$ in a number of stages bounded by $\mathcal{O}(\ln(R_0/\epsilon))$. The corresponding iteration complexity in each stage is of order $\mathcal{O}(\sqrt{s\rho\Omega L}\ln(R_0/\epsilon))$. In terms of the batch size, the total number of samples used by the CSGE-SR method is bounded by

$$\mathcal{O}\left(\sqrt{s\rho\Omega L}\ln\left(\frac{R_0}{\epsilon}\right) + s\rho\mathcal{L}\Omega\left(\ln^2\left(\delta^{-1}\right) + \Omega\right)\ln\left(\frac{R_0}{\epsilon}\right) + \frac{(s\rho\sigma_*)^2\Omega\left(\widehat{\delta}_K + \Omega\right)}{\epsilon^2}\right).$$

5.5.4 Application : Sparse Generalized Linear Regression

Problem setup

Recall that the original problem of sparse generalized linear regression problem we want to solve is as follows. We want to recover the s-sparse signal $x^* \in \operatorname{int} X \subset \mathbf{R}^n$ from observations

$$\eta_t = u(\phi_t^T x^*) + \sigma\zeta_t, \quad t = 1, 2, \ldots, N, \tag{5.5.15}$$

where $u : \mathbf{R} \to \mathbf{R}$ is some non-decreasing and continuous "activation function", $\phi_t \in \mathbf{R}^n$ and $\zeta_t \in \mathbf{R}$ are mutually independent. On top of this setting we also consider the following assumptions

- regressors ϕ_t are independent sub-Gaussian r.v., $\phi_t \sim \mathcal{SG}(0, S)$ i.e., for some $S \in \mathbb{S}^n$ and all $z \in \mathbf{R}^n$,
$$\mathbf{E}\left[e^{z^T\phi_t}\right] \leq e^{\frac{1}{2}z^TSz}.$$
Furthermore, we suppose that $\|\Sigma\|_\infty =: \max_i \Sigma_{ii} \leq \nu$, and that for some $\kappa_\Sigma, \varkappa > 0$, $S \preceq \varkappa\Sigma$ where
$$\Sigma = \mathbf{E}[\phi_1\phi_1^T] \succeq \kappa_\Sigma I;$$
- noises ζ_t are mutually independent and independent of ϕ_t zero mean sub-Gaussian, $\zeta_t \sim \mathcal{SG}(0, 1)$: $\mathbf{E}[e^{s\zeta_t}] \leq e^{s^2/2}$;
- activation function u is strongly monotone and Lipschitz continuous, i.e., for all $t \geq t'$
$$\underline{r}(t - t') \leq u(t) - u(t') \leq \bar{r}(t - t'). \tag{5.5.16}$$

As already explained in the introduction, estimation of $x^* \in \operatorname{int} X$ may be addressed through solving the stochastic optimization problem
$$\min_{x\in X}\left\{f(x) := \mathbf{E}[v(\phi^Tx) - \phi^Tx\eta]\right\} \tag{5.5.17}$$
where $v'(t) = u(t)$. The gradient of the problem objective and its stochastic estimate are given by
$$g(x) = \mathbf{E}[\phi\big(u(\phi^Tx) - \eta\big)] \quad \text{and} \quad \mathcal{G}(x, \underbrace{(\phi, \zeta)}_{=:\xi}) := \phi\big(u(\phi^Tx) - \eta\big) = \phi\big(u(\phi^Tx) - u(\phi^Tx^*)\big) - \phi\zeta.$$

This being introduced, we now place ourselves in the "vanilla" sparsity setting. In this setting we have $\|\cdot\| = \|\cdot\|_1$ and consequently $\|\cdot\|_* = \|\cdot\|_\infty$. We also consider to have at our disposal an initialization point $x_0 \in X$ such that $\|x_0 - x^*\|_1 \leq R_0$ for some $R_0 \geq 0$. In the next proposition we provide the verification of several assumptions made during this chapter.

Proposition 5.5.2 *Assume that the conditions of the setting we have described are valid then*

1. *[Smoothness] Objective function f is L-smooth with $L = \bar{r}\nu$.*
2. *[Quadratic minoration] f satisfies*
$$f(x) - f(x_*) \geq \tfrac{1}{2}\underline{r}\|x - x_*\|_\Sigma^2. \tag{5.5.18}$$
3. *[Sub-exponential noise] The dual norm of the stochastic noise $\|\mathcal{G}(x, \xi) - g(x)\|_*$ is $\sigma(x)$-sub exponential with parameter verifying*
$$\begin{aligned}\sigma^2(x) &\leq \bar{r}^2\nu\left(2.32\ \sqrt{2\varkappa(1 + \ln(n))} + \sqrt{2}\right)^2\|x - x^*\|_\Sigma^2 + 10.77\ \sigma^2\nu(1 + \ln(n))\\ &\leq \frac{\bar{r}^2\nu}{\underline{r}}\left(4.64\ \sqrt{\varkappa(1 + \ln(n))} + 2\right)^2(f(x) - f(x^*)) + 10.77\ \sigma^2\nu(1 + \ln(n)).\end{aligned}$$
4. *[Reduced Strong Convexity] Assumption [RSC] holds with $\Upsilon = 1$ and $\rho = (\kappa_\Sigma\underline{r})^{-1}$.*

Proof of each point of the proposition is deferred to the appendix.

Next, in Lemma 5.5.1 we state the condition $\mathbf{Q}(\lambda, \psi)$ that provides insights on why the RSC assumption holds in the specific GLR setup we explore.

Lemma 5.5.1 *Let $\lambda > 0$ and $0 < \psi \leq 1$, and suppose that for all subsets $I \subset \{1, ..., n\}$ of cardinality smaller than s the following property is verified:*

$$\forall z \in \mathbf{R}^n \quad \|z_I\|_1 \leq \sqrt{\frac{s}{\lambda}}\|z\|_\Sigma + \frac{1}{2}(1-\psi)\|z\|_1 \qquad \mathbf{Q}(\lambda, \psi)$$

where z_I is obtained by zeroing all its components with indices $i \notin I$.

If objective function f satisfies the quadratic minoration condition*, i.e., for some $\mu > 0$,*

$$f(x) - f(x^*) \geq \frac{1}{2}\mu\|x - x^*\|_\Sigma^2, \tag{5.5.19}$$

and that $\widehat{x}$ is an admissible solution to (5.5.12) satisfying, with probability at least $1 - \epsilon$,

$$\Psi_\kappa(\widehat{x}) \leq \Psi_\kappa(x^*) + \Lambda.$$

Then, with probability at least $1 - \epsilon$,

$$\|\widehat{x} - x^*\|_1 \leq \frac{s\kappa}{\lambda\mu\psi} + \frac{\Lambda}{\kappa\psi}. \tag{5.5.20}$$

This condition is reminiscent of the family of conditions presented in [83] and appears as a generalization of both the Restricted Eigenvalue property [75] and the Compatibility Condition [77]. Notably, it represents a more relaxed condition compared to the latter two, as condition $\mathbf{Q}(\lambda, \psi)$ addresses the distribution of the random designs rather than being limited to scenarios with a fixed design matrix. This constraint limits applications of such conditions in addressing recovery problems within online settings, characterized by iterative changes in the design matrix. Thus, condition $\mathbf{Q}(\lambda, \psi)$ offers a more flexible approach.

Non-Euclidean Stochastic Gradient Extrapolation

In this section, we consider the following distance-generating function of the ℓ_1-ball of $\mathbf{R}^n$ (cf. [37, Section 5.7.1])

$$\omega(x) = \frac{c}{p}\|x\|_p^p, \quad p = \begin{cases} 2, & n = 2 \\ 1 + \frac{1}{\ln(n)}, & n \geq 3, \end{cases} \qquad c = \begin{cases} 2, & n = 2, \\ e\ln(n), & n \geq 3. \end{cases} \tag{5.5.21}$$

It immediately follows that ω is strongly convex with modulus 1 w.r.t. the norm $\|\cdot\|_1$ on its unit ball, and that $\Omega \leq e\ln(n)$. This d.-g.f. is chosen as the main tool of our local proximal setup. We have now at our disposal a sparse recovery algorithm, the CSGE-SR Algorithm (Alg. 9), that is specifically tailored to solve the sparse GLR problem of this section.

We can now conclude that running the CSGE-SR Algorithm for K stages, under the framework described in this section, produces an estimate $\widetilde{x}_K$ such that estimation error $\|\widetilde{x}_K - x^*\|_1 \leq \epsilon$ with probability at least $1 - \delta$ for any $\epsilon \in (0, R_0)$, when the total number of stages $K \asymp \ln(R_0/\epsilon)$. This entails a corresponding iteration complexity of order

$$\mathcal{O}\left(\sqrt{\frac{s\bar{r}\nu\ln(n)}{\underline{r}\kappa_\Sigma}}\ln\left(\frac{R_0}{\epsilon}\right)\right)$$

In terms of sample complexity, the total number of samples used by the CSGE-SR method to achieve this optimal iteration complexity is bounded by

$$\mathcal{O}\left(\sqrt{\frac{s\overline{r}\nu\ln(n)}{\underline{r}\kappa_\Sigma}}\ln\left(\frac{R_0}{\epsilon}\right)+\frac{s\overline{r}^2\nu\ln(n)\left(\sqrt{\varkappa\ln(n)}+1\right)^2\left(\ln(n)+\ln\left(\frac{1}{\delta}\right)^2\right)}{\underline{r}\kappa_\Sigma}\ln\left(\frac{R_0}{\epsilon}\right)+\frac{\sigma^2s^2\nu\ln(n)^2\ln(\frac{n}{\delta})}{\underline{r}^2\kappa_\Sigma^2\epsilon^2}\right),$$

when the dimension $n > 2$.

5.6 Numerical Experiments

In this section we present a numerical comparison of different sparse recovery algorithms in a high-dimensional estimation problem. We consider the sparse generalized linear regression model with random design. In this setting we are looking to recover the s-sparse optimum $x^* \in \mathbf{R}^n$ from i.i.d observations

$$\eta_i = u(\phi_i^T x^*) + \sigma\zeta_i,\ i = 1, 2, \ldots, N.$$

We use the same activation function $u(\cdot)$ used in the preceding chapter, i.e.,

$$u_\alpha(x) = x\mathbf{1}\{|x| \le \alpha\} + \operatorname{sign}(x)[\alpha^{-1}(|x|^\alpha - 1) + 1]\mathbf{1}\{|x| > \alpha\},\quad \alpha > 0,\ x \in \mathbf{R}.$$

In this study we explore three different types of activation functions by varying the parameter α which controls the non-linearity of the latter function. We consider the linear link function $u_1(\cdot)$, and two non-linear activation functions $u_{1/2}(\cdot)$ and $u_{1/10}(\cdot)$. The s non-vanishing components of the ground truth x^* are sampled from the s-dimensional standard Gaussian distribution. The regressors ϕ_i are independent and identically distributed realizations of the random variable ϕ, where $\phi \sim \mathcal{N}(0, \Sigma)$, and Σ is a diagonal matrix with entries $0 < \Sigma_{1,1} \le \cdots \le \Sigma_{n,n}$. The additive observation noise ζ follows the centered Gaussian distribution with variance σ^2. Due to memory constraints, we generate the pairs of observations (η_i, ϕ_i) on the fly at each oracle calls.

In the first series of numerical simulations (Figure 5.1), we compare the CSGE-SR algorithm against three other multistage routines. The first two algorithms against which we compare uses hard-thresholding steps to enforce sparsity of the estimators. These algorithms are SMD-SR [59] and SGE-SR [2]. An other algorithm that is compared is the CSMD-SR [1]. For the two algorithms using hard-thresholding steps, the proximal mapping contains the distance generating function of the form $\omega(x) = c_1(n)\|x\|_p^2$, with $p = 1 + 1/\ln(n)$ which allows the proximal mapping to be computed in a closed form solution. The composite methods are both similar in the idea, they are based on the same non-Euclidean composite proximal operator which contains a d.-g.f. of the form $\omega(x) = c_2(n)\|x\|_p^p$, with $p = 1 + 1/\ln(n)$, a penalization $h(x) = \kappa\|x\|_1$ and with the constraint of being inside the ℓ_1-ball. For these simulations we set the dimension $n = 500\,000$, we fix the maximal number of calls to the stochastic oracle (estimation sample size) to $N = 250\,000$, and sparsity level $s = 250$; unless stated otherwise the condition number is set to 1.

In the second series of experiments we study the impact of the condition number on the convergence of the CSGE-SR algorithm. For this setting we set the dimension $n = 100\,000$, we fix the maximal number of calls to the stochastic oracle to $N = 200\,000$, and sparsity level of the minizer is set to $s = 50$.

For each simulation, we run every algorithm to obtain 50 trajectories, and then present the median of these trajectories along with the first and last deciles.

From the series of experiments we have conducted, several observations emerge. First, from the first row of Figure 5.1, we can observe that all algorithms exhibit a regime with a fast linear rate

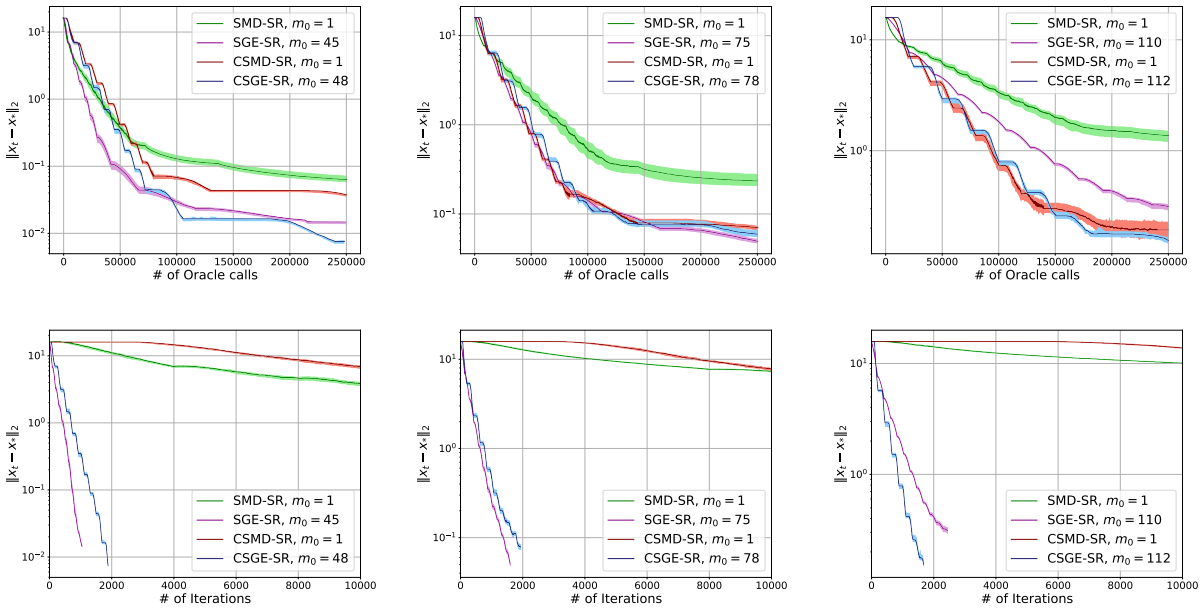


Figure 5.1: Estimation error $\|x_t - x^*\|_2$ against the number of stochastic oracle calls (top row) and against the number of algorithms iterations (bottom row) for SMD-SR, SGE-SR, CSMD-SR and CSGE-SR algorithms. In the left, middle, and right columns of the plot we show the results for the linear activation function u_1, and the nonlinear activation functions $u_{1/2}$ and $u_{1/10}$, respectively. The noise level is set to $\sigma = 0.1$. The legend specifies the value m_0 of the batch size of the preliminary phase of the algorithm for every routines.

of convergence and when they enter the regime where the noise dominates, they face a sublinear decay. The CSGE-SR algorithm consistently ranks among the best in terms of the lowest estimation error reached for a fixed number of samples and this across all three scenarios. The second row compares the performance of the four algorithms when they are compared in terms of the number of algorithm iterations, in other words, this represents the number of times each algorithm computes the non-Euclidean proximal operator. This metric is crucial since calculating a proximal operator can be computationally intensive. A distinct separation is apparent between algorithms for sparse recovery that utilize mini-batch accelerated methods (non-Euclidean SGE and CSGE) and those based on non-accelerated variants (non-Euclidean SMD and CSMD). The SGE-SR and CSGE-SR algorithms use mini-batch approximation, which allows to reduce the variance of the stochastic gradient estimates, consequently helping to reduce the number of times each algorithms computes a proximal operator. Their optimal iteration and sample complexities make CSGE-SR and SGE-SR algorithms viable solutions for distributed sparse recovery problems, where it is crucial to reduce the number of communication rounds between the servers and the clients. The CSGE-SR also have the advantage that it can be made adaptive to the sparsity level and the constant of quadratic growth both at the same time using Lepski's procedure, which is impossible for its non-composite counterpart.

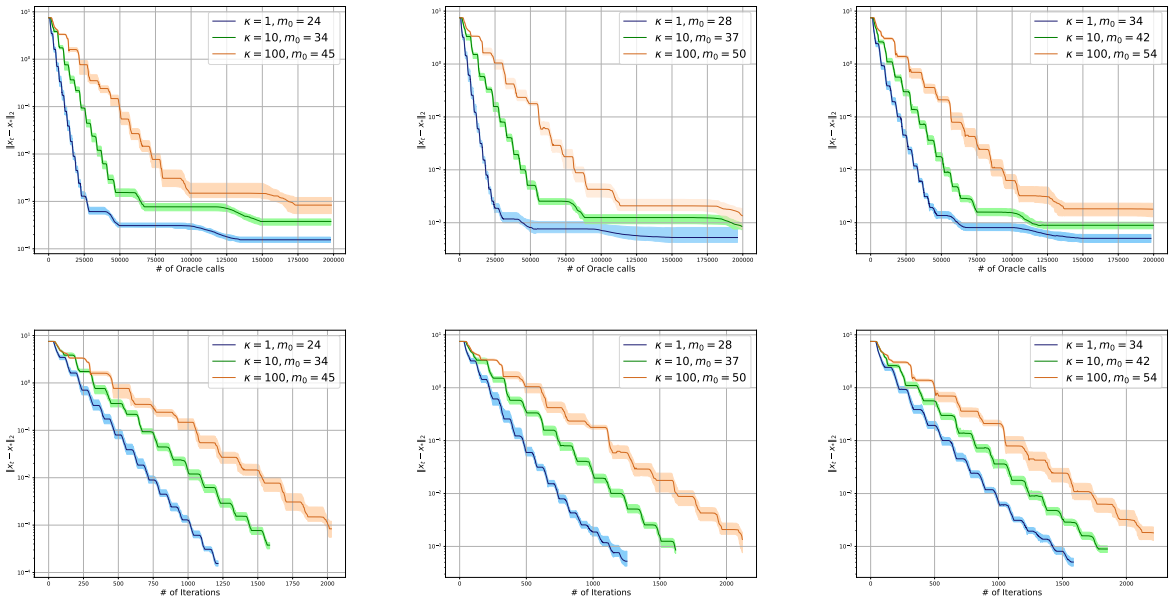


Figure 5.2: Estimation error $\|x_t - x^*\|_2$ against the number of stochastic oracle calls (top row) and against the number of algorithms iterations (bottom row) for the CSGE-SR algorithms. In the left, middle, and right columns of the plot we show the results for the linear activation function u_1, and the nonlinear activation functions $u_{1/2}$ and $u_{1/10}$, respectively. The noise level is set to $\sigma = 0.001$. The legend specifies the value m_0 of the batch size of the preliminary phase of the algorithm for each routine and κ denotes the condition number.

5.7 Appendix

5.7.1 Proof of Lemma 5.2.1

For notational simplicity, put

$$\delta_t := G_t(x_t) - g(x_t), \quad \delta_{t,i} := \mathcal{G}_t(x_t, \xi_{t,i}) - g(x_t).$$

For $y \in E$, put

$$\pi(y) := \sup_{\|z\| \le 1} [\langle y, z\rangle - \omega(z)],$$

$\eta_j = \sum_{i=1}^{j} \delta_{t,i}$ for $j \in [m_t]$, and $\eta_0 = 0$. Observe that for all $\beta > 0$, we have

$$\|\eta_{m_t}\|_* = \sup_{\|z\| \le 1} \langle \eta_{m_t}, z\rangle \le \sup_{\|z\| \le 1} \beta\omega(z) + \beta\pi(\eta_{m_t}/\beta) \le \frac{\beta\Omega}{2} + \beta\pi(\eta_{m_t}/\beta). \tag{5.7.1}$$

On the other hand, due to the strong convexity of ω, π is smooth with 1-Lipschitz continuous gradient, and besides this, as one can easily see, π is Lipschitz continuous with $\|\nabla\pi\| \le 1$. Let us denote

$$\pi_\beta(y) := \beta\pi(y/\beta), \quad \Delta_j := \pi_\beta(\eta_j) - \pi_\beta(\eta_{j-1}).$$

By the Lipschitz continuity of π, $|\Delta_j| \le \|\delta_{t,j}\|_*$, thus

$$\mathbf{E}[\exp(|\Delta_j|/\sigma_t)|\eta_{j-1}] \le \exp(1)$$

(here $\mathbf{E}[\cdot|\eta_{j-1}]$ stands for conditional expectation given η_{j-1}). Furthermore, we have

$$\pi_\beta(\eta_j) \le \pi_\beta(\eta_{j-1}) + \langle \nabla\pi_\beta(\eta_{j-1}), \delta_{t,j}\rangle + \tfrac{1}{2\beta}\|\delta_{t,j}\|_*^2.$$

Combining it with the fact that

$$\mathbf{E}[\|\delta_{t,j}\|_*^2] \le 1.12\ \sigma_t^2, \tag{5.7.2}$$

we get[2]

$$\mathbf{E}[\Delta_j|\eta_{j-1}] \le 0.56\sigma_t^2/\beta.$$

Next, using the bound $x^2e^{-x/2} \le 16e^{-2},\ \forall x \ge 0$, we obtain for $\lambda \in [0, \frac{1}{2\sigma_t}]$,

$$\mathbf{E}\left[\Delta_j^2 e^{\lambda(\Delta_j)_+}|\eta_{j-1}\right] \le \sigma_t^2\mathbf{E}\left[\tfrac{\Delta_j^2}{\sigma_t^2}e^{\frac{|\Delta_j|}{2\sigma_t}}|\eta_{j-1}\right] \le 16e^{-2}\sigma_t^2\mathbf{E}\left[e^{\frac{|\Delta_j|}{\sigma_t}}|\eta_{j-1}\right] \le 6\sigma_t^2,$$

which implies that for $\lambda \in [0, \frac{1}{2\sigma_t}]$,

$$\begin{aligned}\mathbf{E}[e^{\lambda\Delta_j}|\eta_{j-1}] &\overset{(i)}{\le} 1 + \mathbf{E}[\lambda\Delta_j|\eta_{j-1}] + \tfrac{\lambda^2}{2}\mathbf{E}[\Delta_j^2 e^{\lambda(\Delta_j)_+}|\eta_{j-1}] \le 1 + 0.56\frac{\lambda\sigma_t^2}{\beta} + 3\lambda^2\sigma_t^2 \\ &\le \exp\left(0.56\frac{\lambda\sigma_t^2}{\beta} + 3\lambda^2\sigma_t^2\right),\end{aligned} \tag{5.7.3}$$

where (i) follows from the fact that

$$e^x = \sum_{n=0}^{\infty}\frac{x^n}{n!} = 1 + x + \frac{x^2}{2}\left[\sum_{n=2}^{\infty}\frac{2x^{n-2}}{n!}\right] \le 1 + x + \frac{x^2}{2}\left[\sum_{n=2}^{\infty}\frac{2(x^{n-2})_+}{n!}\right] \le 1 + x + \tfrac{1}{2}x^2e^{(x)_+}.$$

When taking the product of inequalities in (5.7.3) from $j = 1$ to m_t and taking subsequent expectations, we get

$$\mathbf{E}[e^{\lambda\pi_\beta(\eta_{m_t})}] \le \exp\left(0.56\frac{\lambda m_t\sigma_t^2}{\beta} + 3\lambda^2 m_t\sigma_t^2\right).$$

Finally, due to Ineq. (5.7.1), we have $\mathbf{E}[e^{\lambda\|\eta_{m_t}\|_*}] \le e^{\beta\Omega\lambda/2}\mathbf{E}[e^{\lambda\pi_\beta(\eta_{m_t})}]$. Taking $\beta^2 = 1.12m_t\sigma_t^2$, we obtain that

$$\mathbf{E}\left\{\exp(\lambda\|\eta_{m_t}\|_*)\right\} \le \exp\left(2\lambda\sigma_t\sqrt{0.28\Omega m_t} + 3\lambda^2\sigma_t^2 m_t\right),$$

which, together with $m_t[G_t(x_t) - g(x_t)] = \eta_{m_t}$, implies (5.2.1).

To prove the bound in (5.2.2), we take $\lambda = \frac{1}{3\bar\sigma_t}$ in (5.2.1), arriving at

$$\mathbf{E}_{\lceil t-1\rceil}\left\{\exp\left(\frac{\|G_t(x_t) - g(x_t)\|_*}{3\bar\sigma_t}\right)\right\} \le \exp\left(\frac{1.06}{3} + \frac{1}{3\Omega}\right) \le \exp(1)$$

which completes the proof. □

[2]Indeed, the optimal distribution in the optimization problem $\max_{P_\xi}\left\{\mathbf{E}[\xi^2] : \mathbf{E}[e^\xi] \le \exp(1), \xi > 0\right\}$ has a 2-point support, and straightforward computation leads to the solution $\mathrm{Prob}\{\xi = a\} = p, \mathrm{Prob}\{\xi = 0\} = 1-p$ with $p = \frac{\exp(1)-1}{\exp(a)-1}$ where a is the maximizer of $a^2\frac{\exp(1)-1}{\exp(a)-1}$, yielding the optimal value of 1.1128... ≤ 1.12.

5.7.2 Proof of Lemma 5.2.2

Let $a \wedge b := \min\{a, b\}$ for $a, b \in \mathbf{R}$. Define

$$\tau(\eta, s, u) = (k+1) \wedge \min\{t \in \mathbf{Z}_+ : \mathfrak{Y}_t \geq \eta,\ \widehat{\mathfrak{s}}_t \leq s,\ \mathfrak{u}_t \leq u\}.$$

Note that $\mathfrak{u}_t$ and $\mathfrak{s}_t$ are $\mathcal{F}_{t-1}$-measurable, so $\tau(\eta, s, u)$ is a Markov stopping time. We denote

$$p(\eta, s, u) = \text{Prob}\{\exists\ t \leq k,\ \text{s.t. } \mathfrak{Y}_t \geq \eta,\ \widehat{\mathfrak{s}}_t \leq s,\ \mathfrak{u}_t \leq u\}.$$

By Ineq. (5.2.3), we have that $\mathbf{E}_{[t-1]}[\exp(\lambda(\mathfrak{y}_t - \mathfrak{r}_t) - \frac{1}{2}\lambda^2\mathfrak{s}_t^2)] \leq 1$ when $\lambda \in [0, (v\mathfrak{s}_t)^{-1}]$. Therefore, when $\lambda \in [0, (v\mathfrak{s}_t)^{-1}]$ for all $t = 0, ..., k$, $\mathfrak{R}_t := \exp(\lambda\mathfrak{Y}_t - \frac{1}{2}\lambda^2\mathfrak{u}_t^2)$ is a nonnegative supermartingale with $\mathbf{E}[\mathfrak{R}_t] \leq 1$. Therefore, we have that for $\lambda \leq (vs)^{-1}$,

$$\begin{aligned}
1 &\overset{(i)}{\geq} \mathbf{E}[\mathfrak{R}_{\tau(\eta,s,u)}\mathbf{1}\{\tau(\eta, s, u) \leq k\}] \\
&= \mathbf{E}\left[\sum_{t=0}^{k} \mathfrak{R}_{\tau(\eta,s,u)}\mathbf{1}\{\tau(\eta, s, u) = t\}\right] \\
&\overset{(ii)}{\geq} \exp(\lambda\eta - \tfrac{1}{2}\lambda^2u^2) \cdot \sum_{t=0}^{k} \text{Prob}\{\tau(\eta, s, u) = t\} \\
&= \exp(\lambda\eta - \tfrac{1}{2}\lambda^2u^2) \cdot p(\eta, s, u)
\end{aligned}$$

where $\mathbf{1}\{\cdot\}$ is the 0-1 indicator function, step (i) follows from that $\mathfrak{R}_t$ is a supermartingale and $\tau(\eta, s, u)$ is a stopping time, and step (ii) follows from the definition of $\mathfrak{R}_t$ and the condition $\mathfrak{Y}_t \geq \eta, \mathfrak{u}_t \leq u$ when $\mathbf{1}\{\tau(\eta, s, u) = t\} = 1$. Therefore, we have

$$p(\eta, s, u) \leq \exp(-\lambda\eta + \tfrac{1}{2}\lambda^2u^2),\ \forall\lambda \leq (vs)^{-1}. \tag{5.7.4}$$

By minimizing the right hand side of Ineq. (5.7.4) over $\lambda \leq (vs)^{-1}$, we obtain

$$p(\eta, s, u) \leq \begin{cases} e^{-\frac{\eta^2}{2u^2}}, & \eta \leq \frac{u^2}{vs} \\ e^{-\frac{\eta}{vs} + \frac{u^2}{2v^2s^2}} \overset{(i)}{\leq} e^{-\frac{\eta}{2vs}}, & \eta > \frac{u^2}{vs} \end{cases}, \tag{5.7.5}$$

where (i) follows from that $-\frac{\eta}{vs} + \frac{u^2}{2v^2s^2} < -\frac{\eta}{2vs}$ when $\eta > \frac{u^2}{vs}$. Then for any $y \geq 0$, by setting $\eta = u\sqrt{2y} + 2vsy$, we arrive at

$$p(\eta, s, u) = \text{Prob}\{\exists\ t \leq k,\ \text{s.t. } \mathfrak{Y}_t \geq u\sqrt{2y} + 2vsy,\ \widehat{\mathfrak{s}}_t \leq s,\ \mathfrak{u}_t \leq u\} \leq e^{-y}, \quad \forall y \geq 0. \tag{5.7.6}$$

This inequality is a Bernstein's type inequality, while the lower bound on $\mathfrak{Y}_t$ depends on a few constants, e.g., u and s. However, for the purpose of algorithmic analysis, we want a "data-driven" version of the above inequality, where the lower bound of $\mathfrak{Y}_t$ depends explicitly on the random variables $\widehat{\mathfrak{s}}_t$ and $\mathfrak{u}_t$ rather than the upper bound constants. To obtain such a result, we will utilize a sequence of constants $\{s_m\}$ and $\{u_m\}$, which are used to upper and lower bound $\widehat{\mathfrak{s}}_t$ and $\mathfrak{u}_t$, respectively. And a "data-driven" bound will be obtained by taking union bound of probability. Specifically, given positive reals $y, \bar{u}$ and $\bar{s}$, we consider positive vectors $\mathbf{u} \in \mathbf{R}^{L+1}$ and $\mathbf{s} \in \mathbf{R}^{M+1}$, where $u_l = 2^{-l}\bar{u},\ l = 0, ..., L-1$ and $s_m = 2^{-m}\bar{s},\ m = 0, ..., M-1$. Let $s_M = \underline{s}$ and $u_L = \underline{u} = \sqrt{2y}v\underline{s}$. Observe that

$$L = \lceil \log_2(\bar{u}/\underline{u}) \rceil \quad \text{and} \quad M = \lceil \log_2(\bar{s}/\underline{s}) \rceil.$$

Then by Ineq. (5.7.6) with $\eta = 2yv\underline{s},\ s = \underline{s}$ and $u = \underline{u}$, we have

$$\text{Prob}\{\mathfrak{Y}_t \geq 4v\underline{s}y,\ \widehat{\mathfrak{s}}_t \leq \underline{s},\ \mathfrak{u}_{t-1} \leq \underline{u} \text{ for some } t \leq k\} \leq e^{-y} \quad \forall y \geq 0.$$

On the other hand, observe that by Ineq. (5.7.6),

$$\begin{aligned}
\Delta p(y,m,l) &:= \text{Prob}\{\mathfrak{Y}_t \geq 2\mathfrak{u}_t\sqrt{2y} + 4v\widehat{\mathfrak{s}}_t y,\ s_m < \widehat{\mathfrak{s}}_t \leq s_{m-1},\ u_l < \mathfrak{u}_t \leq u_{l-1} \text{ for some } t \leq k\} \\
&\leq \text{Prob}\{\mathfrak{Y}_t \geq u_{l-1}\sqrt{2y} + 2vs_{m-1}y,\ s_m < \widehat{\mathfrak{s}}_t \leq s_{m-1},\ u_l < \mathfrak{u}_t \leq u_{l-1} \text{ for some } t \leq k\} \\
&\leq \text{Prob}\{\mathfrak{Y}_t \geq u_{l-1}\sqrt{2y} + 2vs_{m-1}y,\ \widehat{\mathfrak{s}}_t \leq s_{m-1},\ \mathfrak{u}_t \leq u_{l-1} \text{ for some } t \leq k\} \\
&\leq e^{-y}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\text{Prob}\{\mathfrak{Y}_t \geq 2\mathfrak{u}_t\sqrt{2y} + 4v(\widehat{\mathfrak{s}}_t + \underline{s})y,\ \widehat{\mathfrak{s}}_t \leq \bar{s},\ \mathfrak{u}_t \leq \bar{u} \text{ for some } t \leq k\} \\
&\leq \sum_{m=1}^{M}\sum_{l=1}^{L} \Delta p(y,m,l) + \text{Prob}\{\mathfrak{Y}_t \geq 4v\underline{s}y,\ \widehat{\mathfrak{s}}_t \leq \underline{s},\ \mathfrak{u}_{t-1} \leq \underline{u} \text{ for some } t \leq k\} \\
&\leq (ML+1)e^{-y},
\end{aligned}$$

which completes the proof. □

5.7.3 Proof of Lemma 5.3.1

The proof follows the same structure as the proof of Lemma 5.2.1. Specifically, let us denote $\widehat{\chi}_t := \frac{\chi_t(x^*)}{(2\theta_{t+1}+\theta_t)R_X}$. Then by the fact that $\chi_t(x^*) \leq (2\theta_{t+1}+\theta_t)\|\delta_t\|_* R_X$, we have $\widehat{\chi}_t \leq \|\delta_t\|_*$. Moreover, we have

$$\begin{aligned}
-\widehat{\chi}_t &= \frac{1}{(2\theta_{t+1}+\theta_t)R_X}\max\left\{p_t\langle\delta_t, z_t - x^*\rangle - \frac{5q_t}{2}\|\delta_t\|_*^2, -(2\theta_{t+1}+\theta_t)\|\delta_t\|_* R_X\right\} \\
&\leq \frac{1}{(2\theta_{t+1}+\theta_t)R_X} p_t\langle\delta_t, z_t - x^*\rangle \leq \frac{p_t\|\delta_t\|_* R_X}{(2\theta_{t+1}+\theta_t)R_X} \leq \|\delta_t\|_*,
\end{aligned}$$

where the last inequality follows from the definition of p_t. As a result, we obtain $|\widehat{\chi}_t| \leq \|\delta_t\|_*$. Then by Lemma 5.2.1,

$$\mathbf{E}_{\lceil t-1\rceil}\left\{\exp\left(\frac{|\widehat{\chi}_t|}{3\bar{\sigma}_t}\right)\right\} \leq \exp(1). \tag{5.7.7}$$

On the other hand, by using the fact that for any $x \in \mathbf{R}$, $e^x \leq 1 + x + \frac{x^2\exp\{(x)_+\}}{2}$, we obtain

$$\mathbf{E}_{\lceil t-1\rceil}\left\{e^{\lambda\widehat{\chi}_t}\right\} \leq 1 + \lambda\mathbf{E}_{\lceil t-1\rceil}[\widehat{\chi}_t] + \frac{\lambda^2}{2}\mathbf{E}_{\lceil t-1\rceil}[\widehat{\chi}_t^2 e^{\lambda(\widehat{\chi}_t)_+}], \tag{5.7.8}$$

For $\lambda \in [0, \frac{1}{6\bar{\sigma}_t}]$, we can upper bound $\mathbf{E}_{\lceil t-1\rceil}[\widehat{\chi}_t^2 e^{\lambda(\widehat{\chi}_t)_+}]$ as

$$\mathbf{E}_{\lceil t-1\rceil}[\widehat{\chi}_t^2 e^{\lambda(\widehat{\chi}_t)_+}] \overset{(i)}{\leq} 9\bar{\sigma}_t^2\mathbf{E}_{\lceil t-1\rceil}\left[\frac{\widehat{\chi}_t^2}{9\bar{\sigma}_t^2}\exp\left\{\frac{|\widehat{\chi}_t|}{6\bar{\sigma}_t}\right\}\right] \overset{(ii)}{\leq} 144e^{-2}\bar{\sigma}_t^2\mathbf{E}\left[\exp\left\{\frac{|\widehat{\chi}_t|}{3\bar{\sigma}_t}\right\}|\eta_{j-1}\right] \overset{(iii)}{\leq} 54\bar{\sigma}_t^2,$$

where step (i) follows from the range $\lambda \in [0, \frac{1}{6\bar{\sigma}_t}]$, step (ii) follows from the fact that $x^2 e^{x/2} \leq 16e^{-2+x}$ for all $x \geq 0$ and where $\eta_j = \sum_{i=1}^{j} \delta_{t,i}$ for $j \in [m_t]$, finally step (iii) follows from Ineq. (5.7.7). Substituting the above inequality into Ineq. (5.7.8), we obtain

$$\mathbf{E}_{\lceil t-1 \rceil}\left\{e^{\lambda \widehat{\chi}_t}\right\} \leq 1 + \lambda \mathbf{E}_{\lceil t-1 \rceil}\left[\widehat{\chi}_t\right] + 27\lambda^2 \bar{\sigma}_t^2 \leq \exp(\lambda \mathbf{E}_{\lceil t-1 \rceil}[\widehat{\chi}_t] + 27\lambda^2 \bar{\sigma}_t^2), \ \forall \lambda \in [0, \tfrac{1}{6\bar{\sigma}_t}].$$

Reusing the definition $\widehat{\chi}_t := \frac{\chi_t(x^*)}{(2\theta_{t+1}+\theta_t)R_X}$ and replacing λ with $\lambda(2\theta_{t+1}+\theta_t)R_X$, we have

$$\mathbf{E}_{\lceil t-1 \rceil}\left\{e^{\lambda \chi_t(x^*)}\right\} \leq \exp\left(\lambda \mathbf{E}_{\lceil t-1 \rceil}[\chi_t(x^*)] + 27\lambda^2 (2\theta_{t+1}+\theta_t)^2 R_X^2 \bar{\sigma}_t^2\right), \ \forall \lambda \in [0, \tfrac{1}{6\bar{\sigma}_t(2\theta_{t+1}+\theta_t)R_X}]. \tag{5.7.9}$$

At last, we upper bound $\mathbf{E}_{\lceil t-1 \rceil}[\chi_t(x^*)]$ as

$$\mathbf{E}_{\lceil t-1 \rceil}[\chi_t(x^*)] \leq \mathbf{E}_{\lceil t-1 \rceil}[p_t \langle \delta_t, x^* - z_t \rangle] + \frac{5q_t}{2} \mathbf{E}_{\lceil t-1 \rceil}\left[\|\delta_t\|_*^2\right] \leq \frac{1.12 \cdot 5}{2} \, q_t \bar{\sigma}_t^2 = 2.8 q_t \bar{\sigma}_t^2. \tag{5.7.10}$$

The last inequality is obtained by combining result from Ineqs. (4.9.1), (4.9.2) and (5.7.2). Finally, by invoking the relationships in (5.7.9) and (5.7.10) we prove the lemma. □

5.7.4 Proof of Theorem 5.3.1 and Corollary 5.3.1

Proof of Theorem 5.3.1 First, by fixing $x = x^*$, utilizing the convexity of f, i.e., $\langle g_t, x - x_t \rangle \leq f(x) - f(x_t)$, in Ineq. (5.3.4) and rearranging the terms, we obtain

$$\begin{aligned}\theta_k(1+\tau_k)[f(x_k) - f(x^*)] \leq \theta_1 \eta_1 V(x_0, x^*) + \sum_{t=0}^{k-2} \left(\theta_{t+2}\tau_{t+2} - \theta_{t+1}(1+\tau_{t+1})\right)[f(x_{t+1}) - f(x^*)] \\ + \sum_{t=0}^{k-1} \chi_t(x^*).\end{aligned} \tag{5.7.11}$$

Recall the sub-exponential tail of $\chi_t(x^*)$ characterized in Lemma 5.3.1, and notice that

$$\frac{\Omega \sigma_*^2}{m_t} \leq \bar{\sigma}_t \leq \frac{\Omega}{m_t}\left(\frac{\mathcal{L}L\|x_t - x^*\|^2}{2} + \sigma_*^2\right) \leq \frac{\Omega}{m_t}\left(\frac{\mathcal{L}LR_X^2}{2} + \sigma_*^2\right).$$

Then we use Lemma 5.2.2 to bound the term $\sum_{t=0}^{k-1} \chi_t(x^*)$ with high-probability. More specifically, we take

$$\begin{aligned}\bar{s} &:= \max_{0 \leq t \leq k-1} \sqrt{54\varsigma_t \left(\frac{\mathcal{L}LR_X^2}{2} + \sigma_*^2\right)}, \\ \underline{s} &:= \max_{0 \leq t \leq k-1} \sqrt{54\varsigma_t \sigma_*^2}, \\ \bar{u} &:= \sqrt{\sum_{t=0}^{k-1} 54\varsigma_t \left(\frac{\mathcal{L}LR_X^2}{2} + \sigma_*^2\right)},\end{aligned}$$

in Lemma 5.2.2. Then we have $\max_{0 \leq t \leq k-1} \sqrt{54(2\theta_{t+1}+\theta_t)^2 R_X^2 \bar{\sigma}_t^2} \leq \bar{s}$ and $\sqrt{\sum_{t=0}^{k-1} 54(2\theta_{t+1}+\theta_t)^2 R_X^2 \bar{\sigma}_t^2} \leq \bar{u}$. Then by Ineq. (5.2.4) of Lemma 5.2.2 and the definition of $\widehat{\delta}$ in (5.5.9), we have with probability

at least $1-\delta$,

$$\begin{aligned}\sum_{t=0}^{k-1}\chi_t(x^*) &\le \sum_{t=0}^{k-1} 2.8\ q_t\bar{\sigma}_t^2 + 2\sqrt{\sum_{t=0}^{k-1}54(2\theta_{t+1}+\theta_t)^2R_X^2\bar{\sigma}_t^2\widehat{\delta}} \\ &\quad + \frac{8}{\sqrt{6}}\big(\max_{0\le t\le k-1}\sqrt{54(2\theta_{t+1}+\theta_t)^2R_X^2\bar{\sigma}_t^2} + \max_{0\le t\le k-1}\sqrt{54\varsigma_t\sigma_*^2}\big)\widehat{\delta} \\ &\le \sum_{t=0}^{k-1}\big[\epsilon_t\mathcal{L}[f(x_t)-f(x^*)-\langle g(x^*),x_t-x^*\rangle]+\epsilon_t\sigma_*^2\big] \\ &\quad + 35\widehat{\delta}\sqrt{\sum_{t=0}^{k-1}\varsigma_t\mathcal{L}[f(x_t)-f(x^*)-\langle g(x^*),x_t-x^*\rangle]} \\ &\quad + 6\sqrt{\sum_{t=0}^{k-1}6\varsigma_t\sigma_*^2\widehat{\delta}} + 48\widehat{\delta}\max_{0\le t\le k-1}\sqrt{\varsigma_t\sigma_*^2}, \end{aligned} \tag{5.7.12}$$

where the last inequality follows from the definition of ϵ_t in (5.3.8), the definition of ς_t in (5.5.9), the definition of $\bar{\sigma}_t$ in Lemma 5.3.1, and the fact $\max_{t\le k} y_t \le \sum_{t\le k} y_t$ for $y_t \ge 0$. By combining Ineqs. (5.7.11) and (5.7.12), we have

$$\begin{aligned}&\theta_k(1+\tau_k)[f(x_k)-f(x^*)] \\ &\le \theta_1\eta_1 V(x_0,x^*) + \sum_{t=1}^{k-1}\big(\theta_{t+1}\tau_{t+1}-\theta_t(1+\tau_t)+\epsilon_t\mathcal{L}\big)[f(x_t)-f(x^*)] \\ &\quad + \epsilon_0\mathcal{L}[f(x_0)-f(x^*)-\langle g(x^*),x_0-x^*\rangle] + \sum_{t=0}^{k-1}\epsilon_t\sigma_*^2 \\ &\quad + 35\widehat{\delta}\sqrt{\sum_{t=0}^{k-1}\varsigma_t\mathcal{L}[f(x_t)-f(x^*)-\langle g(x^*),x_t-x^*\rangle]} + 6\sqrt{\sum_{t=0}^{k-1}6\varsigma_t\sigma_*^2\widehat{\delta}} + 48\widehat{\delta}\max_{0\le t\le k-1}\sqrt{\varsigma_t\sigma_*^2}\end{aligned} \tag{5.7.13}$$

Notice that we assume $\theta_{t+1}\tau_{t+1}-\theta_t(1+\tau_t)+\epsilon_t\mathcal{L}<0$. Then we have the following inequality.

$$\begin{aligned}&\sum_{t=1}^{k-1}\big(\theta_{t+1}\tau_{t+1}-\theta_t(1+\tau_t)+\epsilon_t\mathcal{L}\big)[f(x_t)-f(x^*)] + 35\widehat{\delta}\sqrt{\sum_{t=0}^{k-1}\varsigma_t\mathcal{L}[f(x_t)-f(x^*)-\langle g(x^*),x_t-x^*\rangle]} \\ &\overset{(i)}{\le} \sum_{t=1}^{k-1}-\big(\theta_t(1+\tau_t)-\theta_{t+1}\tau_{t+1}-\epsilon_t\mathcal{L}\big)[f(x_t)-f(x^*)] + 35\widehat{\delta}\sqrt{\sum_{t=1}^{k-1}\varsigma_t\mathcal{L}[f(x_t)-f(x^*)]} \\ &\qquad + 35\widehat{\delta}\sqrt{\frac{\mathcal{L}\varsigma_0 L\|x_0-x^*\|^2}{2}} \\ &\overset{(ii)}{\le} \frac{(35\widehat{\delta})^2\big[\sum_{t=1}^{k-1}\varsigma_t\mathcal{L}[f(x_t)-f(x^*)]\big]}{4\big[\sum_{t=1}^{k-1}-\big(\theta_t(1+\tau_t)-\theta_{t+1}\tau_{t+1}-\epsilon_t\mathcal{L}\big)[f(x_t)-f(x^*)]\big]} + 35\widehat{\delta}\sqrt{\frac{\mathcal{L}\varsigma_0 L\|x_0-x^*\|^2}{2}} \\ &\le \frac{307\widehat{\delta}^2\big[\sum_{t=1}^{k-1}\varsigma_t\mathcal{L}[f(x_t)-f(x^*)]\big]}{\big[\sum_{t=1}^{k-1}-\big(\theta_t(1+\tau_t)-\theta_{t+1}\tau_{t+1}-\epsilon_t\mathcal{L}\big)[f(x_t)-f(x^*)]\big]} + 35\widehat{\delta}\sqrt{\frac{\mathcal{L}\varsigma_0 L\|x_0-x^*\|^2}{2}},\end{aligned} \tag{5.7.14}$$

where step (i) follows from the optimality condition of f, i.e., $\langle g(x^*),x_t-x^*\rangle \ge 0$, and the smoothness of f; step (ii) follows from Young's inequality. By substituting Ineq. (5.7.14) in to Ineq. (5.7.13), we

have

$$\begin{aligned}
&\theta_k(1+\tau_k)[f(x_k)-f(x^*)]\\
&\le \theta_1\eta_1 V(x_0,x^*)+\frac{\mathcal{L}\epsilon_0 L}{2}\|x_0-x^*\|^2+35\widehat{\delta}\sqrt{\frac{\mathcal{L}\varsigma_0 L\|x_0-x^*\|^2}{2}}+6\sqrt{\sum_{t=0}^{k-1}6\varsigma_t\sigma_*^2\widehat{\delta}}+48\widehat{\delta}\max_{0\le t\le k-1}\sqrt{\varsigma_t\sigma_*^2}\\
&\quad+\frac{307\widehat{\delta}^2\sum_{t=1}^{k-1}\varsigma_t\mathcal{L}[f(x_t)-f(x^*)]}{\sum_{t=1}^{k-1}[\theta_t(1+\tau_t)-\theta_{t+1}\tau_{t+1}-\epsilon_t\mathcal{L}][f(x_t)-f(x^*)]}+\sum_{t=0}^{k-1}\epsilon_t\sigma_*^2,
\end{aligned}$$

which completes the proof. □

Proof of Corollary 5.3.1 We first check that the stepsize conditions (4.9.16a)-(4.9.16c) and (5.3.7) hold. It is easy to see that condition (4.9.16a) holds. Now let us check that Ineqs. (5.3.2) - (4.9.16c) hold. We have

$$\begin{aligned}
\eta_1\eta_2 &\ge \frac{1}{2}\cdot(24L)^2\ge 25\alpha_2 L^2,\\
\eta_t\tau_{t-1} &\ge \frac{24L}{t}\left(\frac{t-2}{2}-\frac{t-1}{24}\right)\ge\frac{24L}{t}\frac{11t-23}{24}\ge\frac{10L}{3}\ge 5L\alpha_t,\ t=3,...,k\\
\eta_k\tau_k &\ge \frac{24L}{k}\left(\frac{k-1}{2}-\frac{k}{24}\right)\ge\frac{L(11k-12)}{k}\ge L.
\end{aligned}$$

To check Ineq. (5.3.7), we have that for $t\ge 1$,

$$q_t=\frac{\theta_{t+1}(1+\alpha_{t+1}^2)}{\eta_{t+1}}+\frac{\theta_{t+2}\alpha_{t+2}^2}{\eta_{t+2}}+\frac{\theta_{t+2}^2\alpha_{t+2}^2}{\theta_{t+1}\eta_{t+1}}=\frac{3(t+1)^2+t^2}{24L}\le\frac{4(t+1)^2}{24L},$$

then considering the batch size $m_t\ge\lceil\frac{216\Omega\mathcal{L}(t+2)}{L}\rceil$, we obtain

$$\epsilon_t=\frac{2.8\ \Omega q_t}{m_t}\le\frac{11.2\ \Omega(t+1)^2}{24Lm_t}\le\frac{t+1}{24\mathcal{L}}.$$

Given the inequality above, we have that for $t\ge 2$,

$$\theta_t\tau_t+\mathcal{L}\epsilon_{t-1}=t\left(\frac{t-1}{2}-\frac{t}{24}\right)+\mathcal{L}\epsilon_{t-1}\le\frac{t(t-1)}{2}-\frac{(t-1)t}{24}.$$

Combining the above inequality with the fact that $\theta_{t-1}(1+\tau_{t-1})=\frac{t(t-1)}{2}-\frac{(t-1)^2}{24}$, we arrive at

$$\theta_{t-1}(1+\tau_{t-1})-(\theta_t\tau_t+\mathcal{L}\epsilon_{t-1})\ge\frac{t(t-1)}{24}-\frac{(t-1)^2}{24}=\frac{t-1}{24},\qquad t\ge 3.$$

For the case when $t=2$, we have

$$\theta_2\tau_2+\mathcal{L}\epsilon_1\le 1-\frac{1}{12}<1=\theta_1(1+\tau_1),$$

which indicates that condition (5.3.7) holds.

On the other hand, we have $m_t\ge\lceil\frac{216\mathcal{L}(t+2)(\widehat{\delta}^2+\Omega)}{L}\rceil$, thus

$$\varsigma_t=\frac{(2\theta_{t+1}+\theta_t)^2R_X^2\Omega}{m_t}=\frac{(3t+2)^2R_X^2\Omega}{m_t}\le\begin{cases}\min\{\frac{tR_X^2L}{24\mathcal{L}},\frac{t\Omega R_X^2L}{24\mathcal{L}\widehat{\delta}^2}\}, & t\ge 1\\ \min\{\frac{R_X^2L}{108\mathcal{L}},\frac{\Omega R_X^2L}{108\mathcal{L}\widehat{\delta}^2}\}, & t=0\end{cases}.$$

Thus we have

$$\frac{307\widehat{\delta}^2\sum_{t=1}^{k-1}\varsigma_t\mathcal{L}[f(x_t)-f(x^*)]}{\sum_{t=1}^{k-1}[\theta_t(1+\tau_t)-\theta_{t+1}\tau_{t+1}-\epsilon_t\mathcal{L}][f(x_t)-f(x^*)]}\leq 307\Omega LR_X^2. \tag{5.7.15}$$

Moreover, invoking that $m_t\geq\lceil\frac{5(k+1)^3(\widehat{\delta}+\Omega)\sigma_*^2}{L^2\Omega R_X^2}\rceil$ we have

$$6\sqrt{\sum_{t=0}^{k-1}6\varsigma_t\sigma_*^2\widehat{\delta}}\leq 20L\Omega R_X^2,$$

$$48\widehat{\delta}\cdot\max_{0\leq t\leq k-1}\sqrt{\varsigma_t\sigma_*^2}\leq\frac{22L\Omega R_X^2\sqrt{\widehat{\delta}}}{\sqrt{k}}\leq 22L\Omega R_X^2. \tag{5.7.16}$$

Meanwhile, we have

$$\epsilon_t=\frac{2.8\ \Omega q_t}{m_t}\leq\frac{11.2\Omega(t+1)^2}{24Lm_t}\leq\frac{11.2(t+1)^2L\Omega R_X^2}{120(k+1)^3\sigma_*^2},$$

thus $\sum_{t=0}^{k-1}\epsilon_t\sigma_*^2\leq\frac{1}{24}L\Omega R_X^2$. By arranging the terms, we obtain the desired result. □

5.7.5 Proof of Corollary 5.3.2

First of all, we show a modified version of Corollary 5.3.1 without the assumption $k\geq\widehat{\delta}$, but with a modified batch-size policy. Notice that if we modify the batch size in Corollary 5.3.1 to

$$m_t\geq\max\left\{1,\left\lceil\frac{216\mathcal{L}(t+2)(\widehat{\delta}^2+\Omega)}{L}\right\rceil,\left\lceil\frac{5(k+1)^3(\widehat{\delta}^2+\Omega)\sigma_*^2}{L^2\Omega R_X^2}\right\rceil\right\},\quad t\geq 0,$$

then all the results in the proof of Corollary 5.3.1 holds except for Ineq. (5.7.16). Instead, we have the following alternative inequality

$$48\widehat{\delta}\cdot\max_{0\leq t\leq k-1}\sqrt{\varsigma_t\sigma_*^2}\leq 22L\Omega R_X^2, \tag{5.7.17}$$

due to the condition $m_t\geq\lceil\frac{5(k+1)^3(\widehat{\delta}^2+\Omega)\sigma_*^2}{L^2\Omega R_X^2}\rceil$. As a result, we have that with probability $1-\delta$,

$$f(\bar{x}_k)-f(x^*)\leq\frac{797\Omega LR_X^2}{k(k+1)}. \tag{5.7.18}$$

Now let us utilize the result above to prove Corollary 5.3.2. Specifically, we need to show that for each stage, we can halve the distance to optimal solution $\|\widetilde{x}_k-x^*\|^2$ with probability at least $1-\delta/K$. Then by the uniform bound of probability, we will obtain the desired result. We prove it by induction.

First, for $k=1$, by Ineq. (5.7.18) we have with probability at least $1-\delta/K$,

$$f(\widetilde{x}_1)-f(x^*)\leq\frac{797\Omega LR_0^2}{\overline{N}(\overline{N}+1)}\overset{(i)}{\leq}\frac{\mu}{4}R_0^2=\frac{\mu}{2}R_1^2.$$

where step (i) follows from the definition of $\overline{N}$ in (5.5.14). Invoking the relationship $\|x-x^*\|^2\leq\frac{2(f(x)-f(x^*))}{\mu}$, we have that with probability at least $1-\delta/K$,

$$\|\widetilde{x}_1-x^*\|^2\leq R_1^2,$$

completing the proof of the base case.

Now assume that for $k \leq K$, with probability at least $1-(k-1)\delta/K$,

$$f(\widetilde{x}_{k-1}) - f(x^*) \leq \tfrac{1}{2}\mu R_{k-1}^2 \quad \text{and} \quad \|\widetilde{x}_{k-1} - x^*\|^2 \leq R_{k-1}^2.$$

So with probability at least $1-(k-1)\delta/K$, x^* is in X_s. Then by Corollary 5.3.1, we have with probability at least $1-k\delta/K$,

$$f(\widetilde{x}_k) - f(x^*) \leq \frac{797\Omega L R_{k-1}^2}{\overline{N}(\overline{N}+1)} \leq \frac{\mu}{4}R_{k-1}^2 = \frac{\mu}{2}R_k^2.$$

Consequently, we have with probability at least $1-k\delta/K$

$$\|\widetilde{x}_k - x^*\|^2 \leq \frac{2[f(\widetilde{x}_k) - f(x^*)]}{\mu} \leq R_k^2,$$

which completes the proof. □

5.7.6 Proof of Proposition 5.4.1, Corollary 5.4.1 and Corollary 5.4.2

Proof of Proposition 5.4.1 First, notice that Ineqs. (4.9.19) and (4.9.20) in the proof of Proposition 4.9.2 holds. By taking the telescope sum from $t = t_0$ to k, and using the condition (4.9.16a), we obtain

$$\begin{aligned}
&\sum_{t=t_0}^{k} \theta_t \left\{\tau_t f(x_t) - \langle G_t, x - x_t\rangle - \tau_t f(x_{t-1})\right\} \\
&\leq \theta_{t_0}\eta_{t_0} V(z_{t_0-1}, x) - \theta_k \eta_k V(z_k, x) + \theta_k \langle g_k - g_{k-1}, z_k - x\rangle \\
&\quad - \theta_{t_0}\alpha_{t_0}\langle g_{t_0-1} - g_{t_0-2}, z_{t_0-1} - x\rangle + \widetilde{\Delta}_{t_0,k} \\
&\overset{(i)}{\leq} \theta_{t_0}\eta_{t_0} V(z_{t_0-1}, x) + \frac{\theta_k}{2\eta_k}\|g_k - g_{k-1}\|_*^2 + \frac{\theta_{t_0}\alpha_{t_0}^2}{2\eta_{t_0}}\|g_{t_0-1} - g_{t_0-2}\|_*^2 + \frac{\theta_{t_0}\eta_{t_0}}{2}\|z_{t_0-1} - x\|^2 + \widetilde{\Delta}_{t_0,k},
\end{aligned} \tag{5.7.19}$$

where step (i) follows from Cauchy-Schwarz inequality, Ineq. (4.1.11) and Young's inequality, and $\widetilde{\Delta}_{t_0,k}$ is defined as

$$\begin{aligned}
\widetilde{\Delta}_{t_0,k} &:= \sum_{t=t_0}^{k} \theta_t \Big[\alpha_t \langle g_{t-1} - g_{t-2}, z_t - z_{t-1} \rangle - \frac{\tau_t}{2L} \|g_t - g_{t-1}\|_*^2 \\
&\quad - \langle \delta_{t-1} + \alpha_t(\delta_{t-1} - \delta_{t-2}), z_t - x \rangle - \eta_t V(z_{t-1}, z_t) \Big] + \sum_{t=t_0}^{k} \theta_t \langle \delta_t, x_t - x \rangle \\
&\overset{(i)}{\leq} \sum_{t=t_0}^{k} \theta_t \Big[\alpha_t \|g_{t-1} - g_{t-2}\|_* \|z_t - z_{t-1}\| - \frac{\tau_t}{2L} \|g_t - g_{t-1}\|_*^2 - \frac{\eta_t}{10} \|z_{t-1} - z_t\|^2 \\
&\quad - \langle \delta_{t-1} + \alpha_t(\delta_{t-1} - \delta_{t-2}), z_t - x \rangle - \frac{2\eta_t}{5} \|z_{t-1} - z_t\|^2 \Big] + \sum_{t=t_0}^{k} \theta_t \langle \delta_t, x_t - x \rangle \\
&\overset{(ii)}{\leq} \frac{5\theta_{t_0}\alpha_{t_0}^2}{2\eta_{t_0}} \|g_{t_0-1} - g_{t_0-2}\|_*^2 - \frac{\theta_k \tau_k}{2L} \|g_k - g_{k-1}\|_*^2 \\
&\quad + \underbrace{\sum_{t=t_0}^{k} \theta_t \left[-\langle \delta_{t-1} + \alpha_t(\delta_{t-1} - \delta_{t-2}), z_t - x \rangle - \frac{2\eta_t}{5} \|z_t - z_{t-1}\|^2 + \theta_t \langle \delta_t, x_t - x \rangle \right]}_{=: \widetilde{\Delta}_{t_0,k}^{(2)}}.
\end{aligned}$$

where step (i) follows from Cauchy-Schwarz inequality and Ineq. (4.1.11), and step (ii) follows from Young's inequality and condition (5.3.2). Substituting the above inequality into Ineq. (5.7.19) and rearranging the terms, we have

$$\begin{aligned}
&\sum_{t=t_0}^{k} \theta_t \langle G_t, x_t - x \rangle + \theta_k \tau_k [f(x_k) - f(x^*)] \\
&\leq \theta_{t_0}\eta_{t_0} V(z_{t_0-1}, x) + \theta_{t_0}\tau_{t_0}[f(x_{t_0-1}) - f(x^*)] + \sum_{t=t_0}^{k-1} (\theta_{t+1}\tau_{t+1} - \theta_t \tau_t)[f(x_t) - f(x^*)] \\
&\quad + \frac{3\theta_{t_0}\alpha_{t_0}^2}{\eta_{t_0}} \|g_{t_0-1} - g_{t_0-2}\|_*^2 + \left(\frac{\theta_k}{2\eta_k} - \frac{\theta_k \tau_k}{2L} \right) \|g_k - g_{k-1}\|_*^2 + \frac{\theta_{t_0}\eta_{t_0}}{2} \|z_{t_0-1} - x\|^2 + \widetilde{\Delta}_{t_0,k}^{(2)} \\
&\overset{(i)}{\leq} \theta_{t_0}\eta_{t_0} V(z_{t_0-1}, x) + \theta_{t_0}\tau_{t_0}[f(x_{t_0-1}) - f(x^*)] + \sum_{t=t_0}^{k-1} (\theta_{t+1}\tau_{t+1} - \theta_t \tau_t)[f(x_t) - f(x^*)] \\
&\quad + \frac{3\theta_{t_0}\alpha_{t_0}^2 L^2}{\eta_{t_0}(1 + \tau_{t_0-1})^2} \|z_{t_0-1} - x_{t_0-2}\|_*^2 + \frac{\theta_{t_0}\eta_{t_0}}{2} \|z_{t_0-1} - x\|^2 + \widetilde{\Delta}_{t_0,k}^{(2)} \\
&\overset{(ii)}{\leq} \theta_{t_0}\eta_{t_0} V(z_{t_0-1}, x) + \theta_{t_0}\tau_{t_0}[f(x_{t_0-1}) - f(x^*)] + \sum_{t=t_0}^{k-1} (\theta_{t+1}\tau_{t+1} - \theta_t \tau_t)[f(x_t) - f(x^*)] \\
&\quad + \left(\frac{\theta_{t_0}\eta_{t_0}}{2} + \frac{3\theta_{t_0}\eta_{t_0}}{25} \right) R_X^2 + \widetilde{\Delta}_{t_0,k}^{(2)},
\end{aligned} \tag{5.7.20}$$

where step (i) follows from condition (4.9.16c) and the fact that

$$\|g_{t_0-1} - g_{t_0-2}\|^2 \leq L^2 \|x_{t_0-1} - x_{t_0-2}\|^2 = L^2 \left\| \frac{z_{t_0-1} - x_{t_0-2}}{1 + \tau_{t_0-1}} \right\|^2,$$

and step (ii) follows by using condition (5.3.2) to obtain

$$\frac{3\theta_{t_0}\alpha_{t_0}^2 L^2}{\eta_{t_0}(1+\tau_{t_0-1})^2}\|z_{t_0-1}-x_{t_0-2}\|_*^2 \le \frac{3\theta_{t_0}\alpha_{t_0}^2\eta_{t_0}^2\tau_{t_0-1}^2 R_X^2}{25\alpha_{t_0}^2\eta_{t_0}(1+\tau_{t_0-1})^2} \le \frac{3\theta_{t_0}\eta_{t_0}R_X^2}{25}.$$

Next, we work on upper bounding $\widetilde{\Delta}_{t_0,k}^{(2)}$.

$$\begin{aligned}
\widetilde{\Delta}_{t_0,k}^{(2)} \le & -\sum_{t=t_0+1}^{k}\theta_t[\langle\delta_{t-1},z_t-z_{t-1}\rangle+\langle\delta_{t-1},z_{t-1}-x\rangle]-\theta_{t_0}\langle\delta_{t_0-1},z_{t_0}-x\rangle\\
& -\sum_{t=t_0+1}^{k}\theta_t\alpha_t[\langle\delta_{t-1},z_t-z_{t-1}\rangle+\langle\delta_{t-1},z_{t-1}-x\rangle]-\theta_{t_0}\alpha_{t_0}\langle\delta_{t_0-1},z_{t_0}-x\rangle\\
& +\sum_{t=t_0+2}^{k}\theta_t\alpha_t[\langle\delta_{t-2},z_t-z_{t-1}\rangle+\langle\delta_{t-2},z_{t-1}-z_{t-2}\rangle+\langle\delta_{t-2},z_{t-2}-x\rangle]+\theta_{t_0}\alpha_{t_0}\langle\delta_{t_0-2},z_{t_0}-x\rangle\\
& +\theta_{t_0+1}\alpha_{t_0+1}\langle\delta_{t_0-1},z_{t_0+1}-x\rangle-\frac{2}{5}\sum_{t=t_0}^{k}\theta_t\eta_t\|z_{t-1}-z_t\|^2+\sum_{t=t_0}^{k}\theta_t\langle\delta_t,x_t-x\rangle.
\end{aligned}$$

By splitting $\|z_{t-1}-z_t\|^2$ into 4 copies, and using $\theta_t\alpha_t=\theta_{t-1}$ and Young's inequality, i.e., $\langle\delta,z\rangle - a_t\|z\|^2 \le \|\delta\|_*^2/(4a_t)$ for $a_t>0$, we have

$$\begin{aligned}
\widetilde{\Delta}_{t_0,k}^{(2)} \le & \sum_{t=t_0+1}^{k}\left[\frac{5\theta_t}{2\eta_t}\|\delta_{t-1}\|_*^2-\theta_t\langle\delta_{t-1},z_{t-1}-x\rangle\right]+\frac{5\theta_{t_0}}{2\eta_{t_0}}\|\delta_{t_0-1}\|_*^2+\frac{\theta_{t_0}\eta_{t_0}}{10}\|z_{t_0}-x\|^2\\
& +\sum_{t=t_0+1}^{k}\left[\frac{5\theta_t\alpha_t^2}{2\eta_t}\|\delta_{t-1}\|_*^2-\theta_{t-1}\langle\delta_{t-1},z_{t-1}-x\rangle\right]+\frac{5\theta_{t_0}\alpha_{t_0}^2}{2\eta_{t_0}}\|\delta_{t_0-1}\|_*^2+\frac{\theta_{t_0}\eta_{t_0}}{10}\|z_{t_0}-x\|^2\\
& +\sum_{t=t_0+2}^{k}\left[\frac{5\theta_t\alpha_t^2}{2\eta_t}\|\delta_{t-2}\|_*^2+\frac{5\theta_{t-1}}{2\eta_{t-1}}\|\delta_{t-2}\|_*^2+\theta_{t-1}\langle\delta_{t-2},z_{t-2}-x\rangle\right]+\frac{5\theta_{t_0}\alpha_{t_0}^2}{2\eta_{t_0}}\|\delta_{t_0-2}\|_*^2+\frac{\theta_{t_0}\eta_{t_0}}{10}\|z_{t_0}-x\|^2\\
& +\frac{5\theta_{t_0}}{2\eta_{t_0}}\|\delta_{t_0-1}\|_*^2+\frac{\theta_{t_0}\eta_{t_0}}{10}\|z_{t_0+1}-x\|^2+\sum_{t=t_0}^{k}\theta_t\langle\delta_t,x_t-x\rangle\\
\le & \sum_{t=t_0-2}^{k-2}\frac{5}{2}\left[\frac{\theta_{t+1}(2+\alpha_{t+1}^2)}{\eta_{t+1}}+\frac{\theta_{t+2}\alpha_{t+2}^2}{\eta_{t+2}}\right]\|\delta_{t-1}\|_*^2+\frac{5\theta_k(1+\alpha_k^2)}{2\eta_k}\|\delta_{k-1}\|_*^2+\sum_{t=t_0}^{k-1}\theta_t\langle\delta_t,x_t-z_t\rangle\\
& -\theta_k\langle\delta_{k-1},z_{k-1}-x\rangle+\theta_k\langle\delta_k,x_k-x\rangle+\frac{3\theta_{t_0}\eta_{t_0}}{10}\|z_{t_0}-x\|^2+\frac{\theta_{t_0}\eta_{t_0}}{10}\|z_{t_0+1}-x\|^2\\
\le & \sum_{t=t_0-2}^{k}\frac{5}{2}\left[\frac{\theta_{t+1}(2+\alpha_{t+1}^2)}{\eta_{t+1}}+\frac{\theta_{t+2}\alpha_{t+2}^2}{\eta_{t+2}}\right]\|\delta_{t-1}\|_*^2+\sum_{t=t_0}^{k-1}\theta_t\langle\delta_t,x_t-z_t\rangle\\
& +\frac{3\theta_{t_0}\eta_{t_0}}{10}\|z_{t_0}-x\|^2+\frac{\theta_{t_0}\eta_{t_0}}{10}\|z_{t_0+1}-x\|^2+\frac{\theta_k\eta_k}{10}\|z_{k-1}-x\|^2+\frac{\theta_k\eta_k}{10}\|x_k-x\|^2\\
\le & \sum_{t=t_0-2}^{k}\frac{5}{2}\left[\frac{\theta_{t+1}(2+\alpha_{t+1}^2)}{\eta_{t+1}}+\frac{\theta_{t+2}\alpha_{t+2}^2}{\eta_{t+2}}\right]\|\delta_{t-1}\|_*^2+\sum_{t=t_0}^{k-1}\theta_t\langle\delta_t,x_t-z_t\rangle+\frac{2\theta_{t_0}\eta_{t_0}R_X^2}{5}+\frac{\theta_k\eta_k R_X^2}{5}.
\end{aligned} \tag{5.7.21}$$

Alternatively, we can bound $\widetilde{\Delta}^{(2)}_{t_0,k}$ directly by using Cauchy-Schwarz inequality,

$$\begin{aligned}\widetilde{\Delta}^{(2)}_{t_0,k} &\leq \sum_{t=t_0}^{k} \theta_t[(1+\alpha_t)\|\delta_{t-1}\|_* + \alpha_t\|\delta_{t-2}\|_*]R_X + \sum_{t=t_0}^{k}\theta_t\|\delta_t\|_* R_X \\ &\leq \sum_{t=t_0-2}^{k}(2\theta_t + 2\theta_{t+1})\|\delta_t\|_* R_X. \end{aligned} \tag{5.7.22}$$

The result in (5.4.1) follows by combining (5.7.20), (5.7.21) and (5.7.22) and rearranging terms. □

Proof of Corollary 5.4.1 Take $t_0 = \lceil \frac{k}{2} \rceil$. We first provide a high-probability bound for $\sum_{t=t_0-2}^{k} \widetilde{\chi}_t$. Let us define

$$\widetilde{\varsigma}_t := \frac{(2\theta_t + 2\theta_{t+1})^2 R_X^2 \Omega}{m_t}.$$

Then following the same arguments as in Ineq. (5.7.12), we can obtain that with probability at least $1 - \delta/(\lceil k/2 \rceil + 1)$,

$$\begin{aligned}\sum_{t=t_0-2}^{k}\chi_t(x^*) \leq \sum_{t=t_0-2}^{k}&\left[\epsilon_t\mathcal{L}[f(x_t) - f(x^*) - \langle g(x^*), x_t - x^*\rangle] + \epsilon_t\sigma_*^2\right] \\ &+ 35\widetilde{\delta}\sqrt{\sum_{t=t_0-2}^{k}\widetilde{\varsigma}_t\mathcal{L}[f(x_t) - f(x^*) - \langle g(x^*), x_t - x^*\rangle]} \\ &+ 6\sqrt{\sum_{t=t_0-2}^{k} 6\widetilde{\varsigma}_t\sigma_*^2\widetilde{\delta}} + 48\widetilde{\delta}\max_{t_0-2\leq t\leq k}\sqrt{\widetilde{\varsigma}_t\sigma_*^2}. \end{aligned} \tag{5.7.23}$$

By combining Ineqs. (5.4.1) and (5.7.23), utilizing the fact that $V(x,y) \leq \Omega\|x-y\|^2 \leq \Omega R_X^2$, and rearranging the terms, we obtain that with probability at least $1 - \delta/(\lceil k/2 \rceil + 2)$,

$$\begin{aligned}&\sum_{t=t_0}^{k}\theta_t\langle G_t, x_t - x\rangle + (\theta_k\tau_k - \epsilon_k\mathcal{L})[f(x_k) - f(x^*)] \\ &\leq \theta_{t_0}\eta_{t_0}\Omega R_X^2 + \theta_{t_0}\eta_{t_0}R_X^2 + \frac{\theta_k\eta_k}{5}R_X^2 + (\theta_{t_0}\tau_{t_0} + \epsilon_{t_0-1}\mathcal{L})[f(x_{t_0-1}) - f(x^*)] \\ &\quad + \sum_{t=t_0}^{k-1}(\theta_{t+1}\tau_{t+1} - \theta_t\tau_t + \epsilon_t\mathcal{L})[f(x_t) - f(x^*)] + \epsilon_{t_0-2}\mathcal{L}[f(x_{t_0-2}) - f(x^*)] \\ &\quad + \sum_{t=t_0-2}^{k}\epsilon_t\sigma_*^2 + 35\widetilde{\delta}\sqrt{\sum_{t=t_0-2}^{k}\widetilde{\varsigma}_t\mathcal{L}[f(x_t) - f(x^*) - \langle g(x^*), x_t - x^*\rangle]} \\ &\quad + 6\sqrt{\sum_{t=t_0-2}^{k} 6\widetilde{\varsigma}_t\sigma_*^2\widetilde{\delta}} + 48\widetilde{\delta}\max_{t_0-2\leq t\leq k}\sqrt{\widetilde{\varsigma}_t\sigma_*^2}. \end{aligned}$$

Meanwhile, noticing the condition (5.3.7), we have

$$\sum_{t=t_0}^{k} \theta_t \langle G_t, x_t - x\rangle + (\theta_k\tau_k - \epsilon_k\mathcal{L})[f(x_k) - f(x^*)] \tag{5.7.24}$$

$$\leq \theta_{t_0}\eta_{t_0}\Omega R_X^2 + \theta_{t_0}\eta_{t_0}R_X^2 + \frac{\theta_k\eta_k}{5}R_X^2 + (\theta_{t_0}\tau_{t_0} + \epsilon_{t_0-1}\mathcal{L})[f(x_{t_0-1}) - f(x^*)] \tag{5.7.25}$$

$$+ \sum_{t=t_0}^{k-1} \theta_t[f(x_t) - f(x^*)] + \epsilon_{t_0-2}\mathcal{L}[f(x_{t_0-2}) - f(x^*)] \tag{5.7.26}$$

$$+ \sum_{t=t_0-2}^{k} \epsilon_t\sigma_*^2 + 35\widetilde{\delta}\sqrt{\sum_{t=t_0-2}^{k} \widetilde{\varsigma}_t\mathcal{L}[f(x_t) - f(x^*)]}$$

$$+ 6\sqrt{\sum_{t=t_0-2}^{k} 6\widetilde{\varsigma}_t\sigma_*^2\widetilde{\delta}} + 48\widetilde{\delta}\max_{t_0-2\leq t\leq k}\sqrt{\widetilde{\varsigma}_t\sigma_*^2}. \tag{5.7.27}$$

On the other hand, utilizing Corollary 5.3.1 and the uniform bound of probability, we have that with probability at least $1 - \lceil k/2\rceil\delta/(\lceil k/2\rceil + 2)$

$$f(x_t) - f(x^*) \leq \frac{797\Omega L R_X^2}{t(t+1)}, \quad \forall t \in [t_0 - 2, k-1]$$

Next, noticing that the mini-batch size m_t in this corollary is greater or equal to the one in Corollary 4.4.1, thus (4.9.16a)-(4.9.16c) and (5.3.7) hold. Meanwhile, we have $\epsilon_t \leq \frac{t+1}{24\mathcal{L}}$. Therefore, we have $\theta_k\tau_k - \epsilon_k\mathcal{L} \geq 0$ and

$$(\theta_{t_0}\tau_{t_0} + \epsilon_{t_0-1}\mathcal{L})[f(x_{t_0-1}) - f(x^*)] + \sum_{t=t_0}^{k-1} \theta_t[f(x_t) - f(x^*)] + \epsilon_{t_0-2}\mathcal{L}[f(x_{t_0-2}) - f(x^*)]$$

$$\leq \left[t_0\left(\frac{t_0-1}{2} - \frac{t_0}{24}\right) + \frac{t_0}{24}\right]\frac{797\Omega L R_X^2}{(t_0-1)t_0} + \sum_{t=t_0}^{k-1}\frac{797\Omega L R_X^2}{t+1} + \frac{797\Omega L R_X^2}{24(t_0-2)}$$

$$\leq \frac{797\Omega L R_X^2}{2} + \ln 2 \cdot 797\Omega L R_X^2 \leq 951\Omega L R_X^2. \tag{5.7.28}$$

On the other hand, we have $m_t \geq \lceil\frac{216\mathcal{L}(t+2)(\widetilde{\delta}^2+\Omega)}{L}\rceil$, thus for $t \geq t_0 - 2$,

$$\widetilde{\varsigma}_t = \frac{(2\theta_{t+1} + 2\theta_t)^2 R_X^2\Omega}{m_t} = \frac{(4t+2)^2 R_X^2\Omega}{m_t} \leq \min\left\{\frac{tR_X^2 L}{32\mathcal{L}}, \frac{t\Omega R_X^2 L}{32\mathcal{L}\widetilde{\delta}^2}\right\}.$$

Consequently, we have

$$35\widetilde{\delta}\sqrt{\sum_{t=t_0-2}^{k} \widetilde{\varsigma}_t\mathcal{L}[f(x_t) - f(x^*)]} \leq 35\widetilde{\delta}\sqrt{\sum_{t=t_0-2}^{k} \frac{t\Omega R_X^2 L}{32\widetilde{\delta}^2}\cdot\frac{797\Omega L R_X^2}{t(t+1)}} \leq 146\Omega L R_X^2. \tag{5.7.29}$$

Moreover, invoking that $m_t \geq \lceil\frac{5(k+1)^3(\widetilde{\delta}+\Omega)\sigma_*^2}{L^2\Omega R_X^2}\rceil$ we have

$$6\sqrt{\sum_{t=t_0-2}^{k} 6\widetilde{\varsigma}_t\sigma_*^2\widetilde{\delta}} \leq 27L\Omega R_X^2,$$

$$48\widetilde{\delta}\cdot\max_{0\leq t\leq k-1}\sqrt{\widetilde{\varsigma}_t\sigma_*^2} \leq \frac{30L\Omega R_X^2\sqrt{\widetilde{\delta}}}{\sqrt{k}} \leq 30L\Omega R_X^2. \tag{5.7.30}$$

At last, we have

$$\epsilon_t = \frac{2.8\Omega q_t}{m_t} \leq \frac{11.2\Omega(t+1)^2}{24Lm_t} \leq \frac{(t+1)^2 L\Omega R_X^2}{12(k+1)^3\sigma_*^2},$$

thus

$$\sum_{t=t_0-2}^{k} \epsilon_t \sigma_*^2 \leq \frac{1}{36} L\Omega R_X^2. \tag{5.7.31}$$

Finally, by substituting the bounds in Ineqs. (5.7.28), (5.7.29), (5.7.30), and (5.7.31) into Ineq. (5.7.24), we obtain

$$\sum_{t=t_0}^{k} \theta_t \langle G_t, x_t - x\rangle \leq \left(24\Omega + 24 + \frac{24}{5} + 951\Omega + 146\Omega + 27\Omega + 30\Omega + \frac{\Omega}{36}\right) LR_X^2 \leq 1207 L\Omega R_X^2,$$

and we complete the proof. □

To prove Corollary 5.4.2 we will make use of the following technical lemma.

Lemma 5.7.1 *Let Assumption (SEN) hold, then for* $\lambda \in \left[0, \frac{m_t}{2\sigma_t R_X}\right]$,

$$\mathbf{E}_{\lceil t-1\rceil}\left[\exp(\lambda\langle g_t - G_t, x_t - x^*\rangle)\right] \leq \exp\left(\frac{3\lambda^2\sigma_t^2 R_X^2}{m_t}\right) \tag{5.7.32}$$

almost surely. Consequently, for all $m_t \geq 1$,

$$\mathbf{E}_{\lceil t-1\rceil}\left\{\exp\left(\frac{\langle g_t - G_t, x_t - x^*\rangle}{2\sigma_t R_X/\sqrt{m_t}}\right)\right\} \leq \exp(1), \tag{5.7.33}$$

almost surely.

The proof follows from similar argument to Lemma 5.2.1.

Proof of Corollary 5.4.2 First, using Lemma 5.2.2 and Lemma 5.7.1, we have with probability $1-\delta$,

$$\begin{aligned}
\sum_{t=\lceil k/2\rceil}^{k} t\langle \delta_t, x^* - x_t\rangle &\leq 2\sqrt{\widehat{\delta}\sum_{t=\lceil k/2\rceil}^{k}\frac{12t^2\sigma_t^2R_X^2}{m_t}} + \frac{8\widehat{\delta}}{\sqrt{6}}\left(\max_{\lceil k/2\rceil\leq t\leq k}\sqrt{\frac{6t^2\sigma_t^2R_X^2}{m_t}} + \max_{\lceil k/2\rceil\leq t\leq k}\sqrt{\frac{6t^2\sigma_*^2R_X^2}{m_t}}\right)\\
&\leq (8+4\sqrt{3})\widehat{\delta}\sqrt{\sum_{t=\lceil k/2\rceil}^{k}\frac{t^2\mathcal{L}(f(x_t)-f(x_*))R_X^2}{m_t}} + 4\sqrt{3\widehat{\delta}\sum_{t=\lceil k/2\rceil}^{k}\frac{t^2\sigma_*^2R_X^2}{m_t}}\\
&\quad + 16\widehat{\delta}\cdot\max_{\lceil k/2\rceil\leq t\leq k}\sqrt{\frac{t^2\sigma_t^2R_X^2}{m_t}}\\
&\leq (8+4\sqrt{3})\widehat{\delta}\sqrt{\sum_{t=\lceil k/2\rceil}^{k}\frac{t^2\mathcal{L}\langle g_t, x_t - x^*\rangle R_X^2}{m_t}} + 4\sqrt{3\widehat{\delta}\sum_{t=\lceil k/2\rceil}^{k}\frac{t^2\sigma_*^2R_X^2}{m_t}}\\
&\quad + 16\widehat{\delta}\cdot\max_{\lceil k/2\rceil\leq t\leq k}\sqrt{\frac{t^2\sigma_*^2R_X^2}{m_t}}.
\end{aligned}$$

Therefore, we have with probability $1-\delta$

$$\begin{aligned}
&\sum_{t=\lceil k/2\rceil}^{k} \frac{t}{2}\langle g_t, x_t - x^*\rangle \\
&= \sum_{t=\lceil k/2\rceil}^{k} t\langle G_t, x_t - x^*\rangle + \sum_{t=\lceil k/2\rceil}^{k} t\langle \delta_t, x^* - x_t\rangle - \sum_{t=\lceil k/2\rceil}^{k} \frac{t}{2}\langle g_t, x_t - x^*\rangle \\
&\le \sum_{t=\lceil k/2\rceil}^{k} t\langle G_t, x_t - x^*\rangle + (8+4\sqrt{3})\widehat{\delta}\sqrt{\sum_{t=\lceil k/2\rceil}^{k} \frac{t^2\mathcal{L}\langle g_t, x_t - x^*\rangle R_X^2}{m_t}} - \sum_{t=\lceil k/2\rceil}^{k} \frac{t}{2}\langle g_t, x_t - x^*\rangle \\
&\quad + 4\sqrt{3\widehat{\delta}\sum_{t=\lceil k/2\rceil}^{k} \frac{t^2\sigma_*^2 R_X^2}{m_t}} + 16\widehat{\delta}\cdot \max_{\lceil k/2\rceil \le t\le k} \sqrt{\frac{t^2\sigma_*^2 R_X^2}{m_t}} \\
&\overset{(i)}{\le} \sum_{t=\lceil k/2\rceil}^{k} t\langle G_t, x_t - x^*\rangle + \frac{112\widehat{\delta}^2 \sum_{t=\lceil k/2\rceil}^{k} t^2\mathcal{L}R_X^2\langle g_t, x_t - x^*\rangle/m_t}{\sum_{t=\lceil k/2\rceil}^{k} t\langle g_t, x_t - x^*\rangle} \\
&\quad + 4\sqrt{3\widehat{\delta}\sum_{t=\lceil k/2\rceil}^{k} \frac{t^2\sigma_*^2 R_X^2}{m_t}} + 16\widehat{\delta}\cdot \max_{\lceil k/2\rceil \le t\le k} \sqrt{\frac{t^2\sigma_*^2 R_X^2}{m_t}},
\end{aligned}$$

where step (i) follows from Young's inequality. By utilizing the batch size defined in (5.4.4), we have

$$\begin{aligned}
\sum_{t=\lceil k/2\rceil}^{k} \frac{t}{2}\langle g_t, x_t - x^*\rangle &\le \sum_{t=\lceil k/2\rceil}^{k} t\langle G_t, x_t - x^*\rangle + \frac{0.52\sum_{t=\lceil k/2\rceil}^{k} tLR_X^2\langle g_t, x_t - x^*\rangle}{\sum_{t=\lceil k/2\rceil}^{k} t\langle g_t, x_t - x^*\rangle} + \sqrt{\frac{14L^2R_X^2}{5}} + 16\sqrt{\frac{\widehat{\delta}L^2R_X^4}{5k}} \\
&\le \sum_{t=\lceil k/2\rceil}^{k} t\langle G_t, x_t - x^*\rangle + 2.2LR_X^2 + 35.78LR_X^2\sqrt{\frac{\widehat{\delta}}{k}}.
\end{aligned}$$

Finally, we conclude by noting that for $k \ge 1$ we have

$$\sum_{t=\lceil \frac{k}{2}\rceil}^{k} t = \frac{1}{2}\left[\left(\left\lceil\frac{k}{2}\right\rceil + k\right)\left(k - \left\lceil\frac{k}{2}\right\rceil + 1\right)\right] \ge \frac{3k^2}{8}.$$

□

5.7.7 Proof of Proposition 5.5.1, Theorem 5.5.1 and Corollary 5.5.1

Proof of Proposition 5.5.1 Let us start by writing the first order optimality condition of (5.5.4b) :

$$\forall x \in X,\ \langle \widetilde{G}_t + h_t' + \eta_t \nabla_x V(z_{t-1}, x)(z_t), x - z_t\rangle \ge 0,$$

where $h_t' \in \partial h(z_t)$. This condition together with convexity of h and the three point Lemma can be rewritten as follows

$$\langle \widetilde{G}_t, z_t - x\rangle \le h(x) - h(z_t) + \eta_t\left[V(z_{t-1}, x) - V(z_t, x) - V(z_{t-1}, z_t)\right]. \tag{5.7.34}$$

Here and for the rest of the section, $\tau_t = \frac{1-\beta_t}{\beta_t}$, such that (5.5.4c) becomes

$$x_t = \frac{z_t + \tau_t x_{t-1}}{1+\tau_t}. \tag{5.7.35}$$

Smoothness of f together with (5.7.35) writes

$$\begin{aligned}\tau_t f(x_t) - \langle g_t, x - x_t\rangle + (1+\tau_t)h(x_t) &= \tau_t[f(x_t) - \langle g_t, x_t - x_{t-1}\rangle] + \langle g_t, z_t - x\rangle + (1+\tau_t)h(x_t).\\ &\le \tau_t[f(x_{t-1}) - \tfrac{1}{2L}\|g_t - g_{t-1}\|_*^2] + \langle g_t, z_t - x\rangle + (1+\tau_t)h(x_t)\end{aligned}$$

Now by combining result (5.7.34) and the above equation results in

$$\begin{aligned}\tau_t f(x_t) - &\langle g_t, x - x_t\rangle + (1+\tau_t)h(x_t) - \tau_t f(x_{t-1})\\ &\le -\tfrac{\tau_t}{2L}\|g_t - g_{t-1}\|_*^2 + \langle g_t - \widetilde{G}_t, z_t - x\rangle + (1+\tau_t)h(x_t)\\ &\qquad + h(x) - h(z_t) + \eta_t\left[V(z_{t-1}, x) - V(z_t, x) - V(z_{t-1}, z_t)\right].\end{aligned} \tag{5.7.36}$$

Using convexity of h and again (5.7.35) we have

$$\begin{aligned}\tau_t \widetilde{\Psi}(x_t) - &\langle g_t, x - x_t\rangle + h(x_t) - h(x) - \tau_t \widetilde{\Psi}(x_{t-1})\\ &\le -\tfrac{\tau_t}{2L}\|g_t - g_{t-1}\|_*^2 + \langle g_t - \widetilde{G}_t, z_t - x\rangle + \eta_t\left[V(z_{t-1}, x) - V(z_t, x) - V(z_{t-1}, z_t)\right].\end{aligned} \tag{5.7.37}$$

Recall that f is convex, which means that

$$\forall x \in X, \quad \langle g_t, x - x_t\rangle \le f(x) - f(x_t).$$

Therefore the L.H.S. of (5.7.37) writes $\forall x \in X$

$$\begin{aligned}(1+\tau_t)\widetilde{\Psi}(x_t) - &\widetilde{\Psi}(x) - \tau_t \widetilde{\Psi}(x_{t-1})\\ &\le -\frac{\tau_t}{2L}\|g_t - g_{t-1}\|_*^2 + \langle g_t - \widetilde{G}_t, z_t - x\rangle + \eta_t\left[V(z_{t-1}, x) - V(z_t, x) - V(z_{t-1}, z_t)\right].\end{aligned} \tag{5.7.38}$$

Note that

$$\begin{aligned}\langle g_t - \widetilde{G}_t, z_t - x\rangle &= \langle g_t - g_{t-1} - \alpha_t(g_{t-1} - g_{t-2}), z_t - x\rangle - \langle \delta_{t-1} + \alpha_t(\delta_{t-1} - \delta_{t-2}), z_t - x\rangle\\ &= \langle g_t - g_{t-1}, z_t - x\rangle - \alpha_t\langle g_{t-1} - g_{t-2}, z_{t-1} - x\rangle\\ &\quad + \alpha_t\langle g_{t-1} - g_{t-2}, z_t - z_{t-1}\rangle - \langle \delta_{t-1} + \alpha_t(\delta_{t-1} - \delta_{t-2}), z_t - x\rangle.\end{aligned}$$

We now, similarly as what has been done in the proof of Theorem 4.4.1, take the θ_t-weighted sum of equation (5.7.38) for $t = 1, \dots, k$ and by noting that $x_0 = z_0$, and using (5.5.5a), we obtain

$$\begin{aligned}&\sum_{t=1}^{k} \theta_t \left\{(1+\tau_t)\widetilde{\Psi}(x_t) - \widetilde{\Psi}(x) - \tau_t \widetilde{\Psi}(x_{t-1})\right\}\\ &\le \theta_1\eta_1 V(x_0, x) - \theta_k\eta_k V(z_k, x) + \theta_k\langle g_k - g_{k-1}, z_k - x\rangle + \Delta_k,\end{aligned} \tag{5.7.39}$$

where

$$\begin{aligned}\Delta_k := \sum_{t=1}^{k} \theta_t\big[&\alpha_t\langle g_{t-1} - g_{t-2}, z_t - z_{t-1}\rangle - \frac{\tau_t}{2L}\|g_t - g_{t-1}\|_*^2\\ &- \langle \delta_{t-1} + \alpha_t(\delta_{t-1} - \delta_{t-2}), z_t - x\rangle - \eta_t V(z_{t-1}, z_t)\big].\end{aligned}$$

Now we analyze the R.H.S of (5.7.39) in the exact same way as what have done in the proof of Theorem 4.4.1 and we can show that $\forall x \in X$,

$$\sum_{t=1}^{k} \theta_t \left\{ (1+\tau_t)\widetilde{\Psi}(x_t) - \widetilde{\Psi}(x) - \tau_t \widetilde{\Psi}(x_{t-1}) \right\} \leq \theta_1 \eta_1 V(x_0, x) + \sum_{t=0}^{k-1} [p_t \langle \delta_t, x - z_t \rangle + \frac{5q_t}{2} \|\delta_t\|_*^2],$$

where $p_t := \theta_t \mathbf{1}\{t \leq k-2\} + (\theta_t + \theta_{t+1})\mathbf{1}\{t = k-1\}$ and $\delta_t = G_t - g(x_t)$. □

Proof of Theorem 5.5.1 First, by fixing $x = x^*$ and rearranging the terms in Ineq. (5.5.6), we obtain

$$\theta_k(1+\tau_k)[\widetilde{\Psi}(x_k) - \widetilde{\Psi}(x^*)] \leq \theta_1 \eta_1 V(x_0, x^*) + \sum_{t=1}^{k-1} \big(\theta_{t+1}\tau_{t+1} - \theta_t(1+\tau_t)\big)[\widetilde{\Psi}(x_t) - \widetilde{\Psi}(x^*)] + \sum_{t=0}^{k-1} \chi_t(x^*). \tag{5.7.40}$$

Recall the sub-exponential tail of $\chi_t(x^*)$ characterized in Lemma 5.3.1, and notice that

$$\frac{\Omega\sigma_*^2}{m_t} \leq \bar{\sigma}_t \leq \frac{\Omega}{m_t}\left(\frac{\mathcal{L}L\|x_t - x^*\|^2}{2} + \sigma_*^2\right) \leq \frac{\Omega}{m_t}\left(\frac{\mathcal{L}LR_X^2}{2} + \sigma_*^2\right).$$

Then we use Lemma 5.2.2 to bound the term $\sum_{t=0}^{k-1} \chi_t(x^*)$ with high-probability. More specifically, we take

$$\begin{aligned}
\bar{s} &:= \max_{0\leq t\leq k-1} \sqrt{54\varsigma_t\left(\frac{\mathcal{L}LR_X^2}{2} + \sigma_*^2\right)}, \\
\underline{s} &:= \max_{0\leq t\leq k-1} \sqrt{54\varsigma_t\sigma_*^2}, \\
\bar{u} &:= \sqrt{\sum_{t=0}^{k-1} 54\varsigma_t\left(\frac{\mathcal{L}LR_X^2}{2} + \sigma_*^2\right)}.
\end{aligned}$$

Then we have $\max_{0\leq t\leq k-1}\sqrt{54(2\theta_{t+1}+\theta_t)^2R_X^2\bar{\sigma}_t^2} \leq \bar{s}$ and $\sqrt{\sum_{t=0}^{k-1} 54(2\theta_{t+1}+\theta_t)^2R_X^2\bar{\sigma}_t^2} \leq \bar{u}$. Then by Ineq. (5.2.4) of Lemma 5.2.2 and the definition of $\widehat{\delta}$ in (5.5.9), we have with probability at least $1-\delta$,

$$\begin{aligned}
\sum_{t=0}^{k-1} \chi_t(x^*) &\leq \sum_{t=0}^{k-1} 2.8q_t\bar{\sigma}_t^2 + 2\sqrt{\sum_{t=0}^{k-1} 54p_t^2R_X^2\bar{\sigma}_t^2\widehat{\delta}} \\
&\quad + \frac{8}{\sqrt{6}}\big(\max_{0\leq t\leq k-1}\sqrt{54p_t^2R_X^2\bar{\sigma}_t^2} + \max_{0\leq t\leq k-1}\sqrt{54\varsigma_t\sigma_*^2}\big)\widehat{\delta} \\
&\leq \sum_{t=0}^{k-1}\left[\epsilon_t\mathcal{L}[f(x_t) - f(x^*) - \langle g(x^*), x_t - x^*\rangle] + \epsilon_t\sigma_*^2\right] \\
&\quad + 35\widehat{\delta}\sqrt{\sum_{t=0}^{k-1}\varsigma_t\mathcal{L}[f(x_t) - f(x^*) - \langle g(x^*), x_t - x^*\rangle]} \\
&\quad + 6\sqrt{\sum_{t=0}^{k-1} 6\varsigma_t\sigma_*^2\widehat{\delta}} + 48\widehat{\delta}\max_{0\leq t\leq k-1}\sqrt{\varsigma_t\sigma_*^2},
\end{aligned} \tag{5.7.41}$$

where the last inequality follows from the definition of ϵ_t in (5.5.9), the definition of ς_t in (5.5.9), the definition of $\bar{\sigma}_t$ in Lemma 5.3.1, and the fact $\max_{t\leq k} y_t \leq \sum_{t\leq k} y_t$ for $y_t \geq 0$. By combining Ineqs. (5.7.40), (5.7.41) and recalling that $\Psi = \frac{1}{2}f + h$, we have

$$\begin{aligned}
&\theta_k(1+\tau_k)[\widetilde{\Psi}(x_k) - \widetilde{\Psi}(x^*)] + \sum_{t=1}^{k-1}\big(\theta_t(1+\tau_t) - \theta_{t+1}\tau_{t+1}\big)[\Psi(x_t) - \Psi(x^*)] \\
&\leq \theta_1\eta_1 V(x_0, x^*) + \sum_{t=1}^{k-1}\left(\frac{1}{2}(\theta_{t+1}\tau_{t+1} - \theta_t(1+\tau_t)) + \epsilon_t\mathcal{L}\right)[f(x_t) - f(x^*)] \\
&\quad + 35\widehat{\delta}\sqrt{\sum_{t=0}^{k-1}\varsigma_t\mathcal{L}[f(x_t) - f(x^*) - \langle g(x^*), x_t - x^*\rangle]} + \epsilon_0\mathcal{L}[f(x_0) - f(x^*) - \langle g(x^*), x_0 - x^*\rangle] \\
&\quad + \sum_{t=0}^{k-1}\epsilon_t\sigma_*^2 + 6\sqrt{\sum_{t=0}^{k-1}6\varsigma_t\sigma_*^2\widehat{\delta}} + 48\widehat{\delta}\max_{0\leq t\leq k-1}\sqrt{\varsigma_t\sigma_*^2}
\end{aligned} \tag{5.7.42}$$

Notice that due to f being convex and x^* being a minimizer of function f, we have the inequality

$$\widetilde{\Psi}(x_k) - \widetilde{\Psi}(x^*) \geq \Psi(x_k) - \Psi(x^*).$$

Recall that the output of the CSGE algorithm is defined as follows

$$\widehat{x}_k = \Gamma_k^{-1}\left(\theta_k(1+\tau_k)x_k + \sum_{t=1}^{k-1}[\theta_t(1+\tau_t) - \theta_{t+1}\tau_{t+1}]x_t\right),$$

where $\Gamma_k = \theta_k(1+\tau_k) + \sum_{t=1}^{k-1}\theta_t(1+\tau_t) - \theta_{t+1}\tau_{t+1}$. By combining the latter definition of the algorithm's output together with the convexity of Ψ and by observing that for all $t \geq 1$, $\theta_t(1+\tau_t) > \theta_{t+1}\tau_{t+1}$ (a consequence of inequality (5.5.8)), we obtain

$$\begin{aligned}
&\Gamma_k^{-1}\left(\theta_k(1+\tau_k)[\Psi(x_k) - \Psi(x^*)] + \sum_{t=1}^{k-1}\big(\theta_t(1+\tau_t) - \theta_{t+1}\tau_{t+1}\big)[\Psi(x_t) - \Psi(x^*)]\right) \\
&\qquad\qquad \geq \Psi(\widehat{x}_k) - \Psi(x^*).
\end{aligned} \tag{5.7.43}$$

Let us focus on the R.H.S of Ineq.(5.7.42), after using the optimality condition of f, i.e., $\langle g(x^*), x_t - x^*\rangle \geq 0$ and Young's inequality this upper bound holds true for the following two terms

$$\begin{aligned}
&-\sum_{t=1}^{k-1}\big(\tfrac{1}{2}(\theta_t(1+\tau_t) - \theta_{t+1}\tau_{t+1}) - \epsilon_t\mathcal{L}\big)[f(x_t) - f(x^*)] + 35\widehat{\delta}\sqrt{\sum_{t=1}^{k-1}\varsigma_t\mathcal{L}[f(x_t) - f(x^*) - \langle g(x^*), x_t - x^*\rangle]} \\
&\qquad \leq \frac{(35\widehat{\delta})^2\sum_{t=1}^{k-1}\varsigma_t\mathcal{L}[f(x_t) - f(x^*)]}{2\sum_{t=1}^{k-1}\big(\theta_t(1+\tau_t) - \theta_{t+1}\tau_{t+1} - 2\epsilon_t\mathcal{L}\big)[f(x_t) - f(x^*)]}.
\end{aligned}$$

Combining the precedent inequality with result in Ineq. (5.7.43) yields to the following bound

$$\begin{aligned}
\Psi(\widehat{x}_k) - \Psi(x^*) \leq \Gamma_k^{-1}\Bigg[&\theta_1\eta_1 V(x_0, x^*) + \epsilon_0\mathcal{L}[f(x_0) - f(x^*) - \langle g(x^*), x_0 - x^*\rangle] \\
&+ 35\widehat{\delta}\sqrt{\varsigma_0\mathcal{L}[f(x_0) - f(x^*) - \langle g(x^*), x_0 - x^*\rangle]} + \frac{(35\widehat{\delta})^2\sum_{t=1}^{k-1}\varsigma_t\mathcal{L}[f(x_t) - f(x^*)]}{2\sum_{t=1}^{k-1}\big(\theta_t(1+\tau_t) - \theta_{t+1}\tau_{t+1} - 2\epsilon_t\mathcal{L}\big)[f(x_t) - f(x^*)]} \\
&+ \sum_{t=0}^{k-1}\epsilon_t\sigma_*^2 + 6\sqrt{\textstyle\sum_{t=0}^{k-1}6\varsigma_t\sigma_*^2\widehat{\delta}} + 48\widehat{\delta}\max_{0\leq t\leq k-1}\sqrt{\varsigma_t\sigma_*^2}\Bigg].
\end{aligned}$$

Note that the smoothness assumption verified by f provides this upper bound

$$f(x_0) - f(x^*) - \langle g(x^*), x_0 - x^* \rangle \leq \tfrac{1}{2} L \|x_0 - x^*\|.$$

We then have that following result holds true

$$\Psi(\widehat{x}_k) - \Psi(x^*) \leq \Gamma_k^{-1} \Bigg[\theta_1 \eta_1 V(x_0, x^*) + \tfrac{1}{2}\epsilon_0 \mathcal{L} L \|x_0 - x^*\|^2 + 35\widehat{\delta} \sqrt{\frac{\varsigma_0 \mathcal{L} L \|x_0 - x^*\|^2}{2}} + 6\sqrt{\textstyle\sum_{t=0}^{k-1} 6\varsigma_t \sigma_*^2 \widehat{\delta}}$$
$$+ \frac{(35\widehat{\delta})^2 \sum_{t=1}^{k-1} \varsigma_t \mathcal{L}[f(x_t) - f(x^*)]}{2\sum_{t=1}^{k-1} \big(\theta_t(1+\tau_t) - \theta_{t+1}\tau_{t+1} - 2\epsilon_t \mathcal{L}\big)[f(x_t) - f(x^*)]} + \sum_{t=0}^{k-1} \epsilon_t \sigma_*^2 + 48\widehat{\delta} \max_{0 \leq t \leq k-1} \sqrt{\varsigma_t \sigma_*^2} \Bigg]. \tag{5.7.44}$$

Finally, we finish by recalling that $\|x_0 - x^*\|^2 \leq R_X^2$ and $V(x, y) \leq \frac{\Omega}{2}\|x - y\|^2 \leq \frac{\Omega R_X^2}{2}$ which leads to result in (5.5.10). This concludes the proof of the theorem. □

Proof of Corollary 5.5.1 It is easy to verify that the provided parameters verify the conditions (5.5.5a) - (5.5.5c). Now we will check that (5.5.8) is also verified. First note that for $t \geq 1$:

$$q_t = \frac{\theta_{t+1}(1 + \alpha_{t+1}^2)}{\eta_{t+1}} + \frac{\theta_{t+2}\alpha_{t+2}^2}{\eta_{t+2}} + \frac{\theta_{t+2}^2 \alpha_{t+2}^2}{\theta_{t+1}\eta_{t+1}} = \frac{3(t+1)^2 + t^2}{24L} \leq \frac{4(t+1)^2}{24L},$$

observe also that the choice of the batch size implies $m_t \geq \lceil \frac{216\mathcal{L}(t+2)(\widehat{\delta}^2 + \Omega)}{L} \rceil \geq \frac{216\mathcal{L}(t+2)\Omega}{L}$ this leads to

$$\epsilon_t = \frac{2.8\Omega q_t}{m_t} \leq \frac{11.2(t+1)^2}{24 L m_t} \leq \frac{t+1}{24\mathcal{L}}.$$

Combining the above inequality with the parameter value for θ_t and τ_t, we arrive at

$$\theta_t(1+\tau_t) - (\theta_{t+1}\tau_{t+1} + 2\mathcal{L}\epsilon_t) \geq \frac{t(t+1)}{24} - \frac{t^2}{24} = \frac{t}{24} > 0, \qquad t \geq 2.$$

For the case when $t = 1$, we have

$$\theta_1(1+\tau_1) - \theta_2\tau_2 - 2\mathcal{L}\epsilon_1 \geq \frac{7}{12} > \frac{1}{24}.$$

We have shown that $\forall t \geq 1,\ \theta_t(1+\tau_t) - (\theta_{t+1}\tau_{t+1} + 2\mathcal{L}\epsilon_t) \geq \frac{t}{24} > 0$, which writes equivalently as (5.5.8). This proves that the latter condition holds. On the other hand, we have $m_t \geq \lceil \frac{216\mathcal{L}(t+2)(\widehat{\delta}^2 + \Omega)}{L} \rceil$, thus

$$\varsigma_t = \frac{(2\theta_{t+1} + \theta_t)^2 R_X^2 \Omega}{m_t} = \frac{(3t+2)^2 R_X^2 \Omega}{m_t} \leq \begin{cases} \min\{\frac{t R_X^2 L}{24\mathcal{L}}, \frac{t\Omega R_X^2 L}{24\mathcal{L}\widehat{\delta}^2}\}, & t \geq 1 \\ \min\{\frac{R_X^2 L}{108\mathcal{L}}, \frac{\Omega R_X^2 L}{108\mathcal{L}\widehat{\delta}^2}\}, & t = 0 \end{cases}.$$

Thus we have

$$\frac{(35\widehat{\delta})^2 \sum_{t=1}^{k-1} \varsigma_t \mathcal{L}[f(x_t) - f(x^*)]}{2\sum_{t=1}^{k-1}[\theta_t(1+\tau_t) - \theta_{t+1}\tau_{t+1} - 2\epsilon_t\mathcal{L}][f(x_t) - f(x^*)]} \leq 613\Omega L R_X^2.$$

Moreover, invoking that $m_t \geq \lceil \frac{5(k+1)^3(\widehat{\delta}+\Omega)\sigma_*^2}{36L^2\Omega R_X^2} \rceil$ and $\widehat{\delta} \leq k$ we have

$$6\sqrt{\sum_{t=0}^{k-1} 6\varsigma_t\sigma_*^2\widehat{\delta}} \leq 69L\Omega R_X^2,$$

$$48\widehat{\delta} \cdot \max_{0\leq t\leq k-1} \sqrt{\varsigma_t\sigma_*^2} \leq \frac{387L\Omega R_X^2\sqrt{\widehat{\delta}}}{\sqrt{k}} \leq 387L\Omega R_X^2.$$

Meanwhile, we have

$$\epsilon_t = \frac{2.8\Omega q_t}{m_t} \leq \frac{11.2\Omega(t+1)^2}{24Lm_t} \leq \frac{404(t+1)^2L\Omega R_X^2}{120(k+1)^3\sigma_*^2},$$

thus $\sum_{t=0}^{k-1}\epsilon_t\sigma_*^2 \leq L\Omega R_X^2$. With the proposed set of parameter we can also show

$$35\widehat{\delta}\sqrt{\frac{\varsigma_0\mathcal{L}LR_X^2}{2}} \leq 3\Omega LR_X^2$$
$$\frac{\theta_1\eta_1\Omega + \epsilon_0\mathcal{L}L}{2}R_X^2 \leq 13\Omega LR_X^2.$$

We now show that for $k \geq 2$, $\Gamma_k \geq \frac{11}{24}k(k+1)$. Observe that with the prescribed choice of parameter we have

$$\begin{aligned}
\Gamma_k &= \theta_k(1+\tau_k) + \sum_{t=1}^{k-1}(\theta_t(1+\tau_t) - \theta_{t+1}\tau_{t+1}) \\
&= \frac{1}{24}k(12+11k) + \sum_{t=2}^{k-1}(\theta_t(1+\tau_t) - \theta_{t+1}\tau_{t+1}) + \theta_1 - \theta_2\tau_2 \\
&= \frac{1}{24}k(12+11k) + \frac{1}{24}\sum_{t=2}^{k-1}(2t+1) + \frac{1}{6} \\
&\geq \frac{11}{24}k(k+1).
\end{aligned}$$

Finally, by arranging the terms, we obtain the result in (5.5.11). □

5.7.8 Proof of Theorem 5.5.2

We begin by proving the second inequality first. Consider the sequence defined, for any positive integer k such that $k \geq 1$, by

$$R_k := R_{k-1}/\sqrt{2} \text{ and } R_0 \geq \|\widetilde{x}_0 - x^*\|,$$

where $\widetilde{x}_0$ is an initialization point of Algorithm 9. To simplify the analysis we can choose the intial radius as $R_0 = R_X$. The proof is carried out by induction, by using simultaneously Corollary 5.5.1 and Lemma 5.7.54 in order to show that at the end of each stage $k \in \{1, 2, \ldots, K\}$ we have $\|\widetilde{x}_k - x^*\| \leq R_k$ with probability at least $1 - k\delta/K$. For the first stage $k = 1$, observe that when applying Algorithm 8 for $\overline{N}$ iterations to Problem 5.5.12 with κ_1 we have, according to Corollary 5.5.1, with probability at least $1 - \delta/K$:

$$\Psi_{\kappa_1}(\widetilde{x}_1) - \Psi_{\kappa_1}(x^*) \leq \frac{1811\ \Omega LR_0^2}{\overline{N}(\overline{N}+1)} := \Lambda_1$$

A direct application of Lemma 5.7.54 gives us that the following bound also holds with probability at least $1-\delta/K$

$$\|\widetilde{x}_1 - x^*\| \leq \frac{\upsilon}{\kappa_1 \Upsilon} + \kappa_1 \rho s \Upsilon^{-1} = \frac{1811\ \Omega L R_0^2}{\kappa_1 \Upsilon \overline{N}(\overline{N}+1)} + \kappa_1 \rho s \Upsilon^{-1}.$$

After plugging into the last inequality the value of $\kappa_1 = R_0 \sqrt{\frac{1811 \Omega L}{\rho s \overline{N}(\overline{N}+1)}}$ we obtain

$$\|\widetilde{x}_1 - x^*\| \leq \frac{2R_0}{\Upsilon} \sqrt{\frac{1811 \rho s \Omega L}{\overline{N}(\overline{N}+1)}}.$$

Now observe that $\overline{N} \geq \frac{121}{\Upsilon}\sqrt{\rho s \Omega L}$, therefore we have

$$\|\widetilde{x}_1 - x^*\| \leq \frac{2R_0}{\Upsilon} \sqrt{\frac{1811 \rho s \Omega L}{\overline{N}(\overline{N}+1)}} \leq \frac{2R_0\sqrt{1811 \rho s \Omega L}}{\overline{N}\Upsilon} \leq \frac{2R_0\sqrt{1811}}{121} < \frac{R_0}{\sqrt{2}} = R_1.$$

We have shown that with probability at least $1-\delta/K$ we have $\|\widetilde{x}_1 - x^*\| \leq R_0/\sqrt{2}$.

Now let us assume that for some $k \in \{1, 2, \ldots, K\}$, we have run the multistage procedure up to stage $k-1$. The estimate provided by Algorithm 9 satisfies with probability at least $1-(k-1)\delta/K$:

$$\|\widetilde{x}_{k-1} - x^*\| \leq R_{k-1} = 2^{-\frac{k-1}{2}} R_0,$$

in other words, with at least the same probability, we have that $x^* \in X_k$.
We now launch Algorithm 9 for the k-th stage. Again, Corollary 5.5.1 gives us with probability at least $1-\delta/K$ that the k-th stage estimate $\widetilde{x}_k$ verifies

$$\Psi_{\kappa_k}(\widetilde{x}_k) - \Psi_{\kappa_k}(x^*) \leq \frac{1811\ \Omega L R_{k-1}^2}{\overline{N}(\overline{N}+1)} =: \Lambda_k.$$

We can now use the RSC assumption to obtain that with probability at least $1-\delta/K$ the following holds true

$$\|\widetilde{x}_k - x^*\| \leq \frac{1811\ \Omega L R_{k-1}^2}{\kappa_k \Upsilon \overline{N}(\overline{N}+1)} + \frac{\rho \kappa_k s}{\Upsilon}.$$

Recall that $\kappa_k = R_{k-1}\sqrt{\frac{1811 \Omega L}{\rho s \overline{N}(\overline{N}+1)}}$ and $\overline{N} \geq \frac{121}{\Upsilon}\sqrt{\rho s \Omega L}$, this, together with the last inequality results in

$$\|\widetilde{x}_k - x^*\| \leq \frac{R_{k-1}}{\Upsilon} \sqrt{\frac{1811 \rho s \Omega L}{\overline{N}(\overline{N}+1)}} \leq \frac{2R_{k-1}\sqrt{1811}}{121} < \frac{R_{k-1}}{\sqrt{2}} = R_k.$$

We have proved that with probability at least $1-(k-1)\delta/K$, $x^* \in X_k$ and with probability at least $1-\delta/K$, $x^* \in X_{k+1}$ therefore the union bound gives us that $x^* \in X_k \cap X_{k+1}$ with probability at least $1-(k-1)\delta/K - \delta/K = 1-k\delta/K$.

Now we move onto the proof of the first inequality. Let consider that we have run the CSGE-SR Algorithm for K stages. Then according to the result provided in the first part of the proof, we

have shown that with probability greater or equal than $1-\delta$ that $x^* \in X_{K+1}$ and together with Corollary 5.5.1 we have

$$\Psi_{\kappa_K}(\widetilde{x}_K) - \Psi_{\kappa_K}(x^*) \leq \frac{1811\ \Omega L R_K^2}{\overline{N}(\overline{N}+1)}.$$

We then have that

$$\begin{aligned} f(\widetilde{x}_K) - f(x^*) &\leq \frac{1811\ \Omega L R_K^2}{\overline{N}(\overline{N}+1)} + \kappa_K\|x_*\| - \kappa_K\|\widetilde{x}_K\| \\ &\leq \frac{1811\ \Omega L R_K^2}{\overline{N}(\overline{N}+1)} + \kappa_K\|\widetilde{x}_K - x^*\| \\ &\leq \frac{1811}{121^2}\frac{\Upsilon^2 R_0^2}{\rho s}\cdot 2^{-K} + \frac{\sqrt{3622}}{121}\frac{\Upsilon R_0^2}{\rho s}\cdot 2^{-K} \\ &\leq \frac{\Upsilon(\Upsilon+1)}{\rho s}R_0^2\cdot 2^{-K-1}. \end{aligned}$$

Where in the third inequality we have replaced κ_K and $\overline{N}$ by their prescribed values as appearing in Algorithm 9. This concludes the proof of the theorem. □

5.7.9 Proof of Proposition 5.5.2

1^o. We start by proving that the objective function f has L-Lipschitz continuous derivatives. Recall that the gradient of f writes : $g(x) = \mathbf{E}[\phi\big(u(\phi^T x) - \eta\big)]$ and by considering the setup described in Section 5.5.4 we have

$$\begin{aligned} \|g(x) - g(x')\|_\infty &= \sup_{\|z\|_1\leq 1}\langle g(x) - g(x'), z\rangle = \sup_{\|z\|_1\leq 1}\mathbf{E}\{\phi^T z[u(\phi^T x) - u(\phi^T x')]\} \\ &\leq \sup_{\|z\|_1\leq 1}\bar{r}\mathbf{E}\{|\phi^T z||\phi^T(x-x')|\} \leq \bar{r}\sup_{\|z\|_1\leq 1}\sqrt{\mathbf{E}\{(\phi^T z)^2\}}\cdot\|x-x'\|_\Sigma \\ &\leq \bar{r}\sqrt{\nu}\|x-x^*\|_\Sigma \leq \bar{r}\nu\|x-x'\|_1. \end{aligned} \tag{5.7.45}$$

We can conclude that the objective function has L-Lipschitz continuous derivatives w.r.t $\|\cdot\|_1$ with $L = \bar{r}\nu$.

2^o. We now proceed to the verification of the quadratic growth condition (5.5.18),

$$\begin{aligned} f(x) - f(x^*) &= \int_0^1 g(x^* + t(x-x^*))^T(x-x^*)dt \\ &= \int_0^1 \mathbf{E}\{\phi[u(\phi^T x^* + t(x-x^*)) - u(\phi^T x^*)]\}^T(x-x^*)dt \\ \text{[by (5.5.16)]} \quad &\geq \int_0^1 \underline{r}\mathbf{E}\{[\phi^T(x-x^*)]^2\}tdt = \frac{\underline{r}}{2}\|x-x^*\|_\Sigma^2 \geq \frac{\underline{r}\kappa_\Sigma}{2}\|x-x^*\|_2^2. \end{aligned} \tag{5.7.46}$$

Therefore, condition (5.5.19) holds with $\mu = \underline{r}$ which is independent of problem dimension n.

3^o. To prove the third point of the proposition we show that the stochastic noise $\mathcal{G}(x,(\phi,\eta))-g(x)$ verifies

$$\mathbf{E}\left[\exp\left(\frac{\|\mathcal{G}(x,(\phi,\eta))-g(x)\|_\infty}{\sigma(x)}\right)\right]\le\exp(1) \tag{5.7.47}$$

where

$$\sigma(x):=2.32\sqrt{\nu(1+\ln n)}[\sigma+\bar r\|x-x^*\|_\Sigma\sqrt{\varkappa}]+\bar r\sqrt{\nu}\|x-x^*\|_\Sigma. \tag{5.7.48}$$

Using the value of the stochastic estimate of the gradient of the problem objective together with (5.7.45), we have

$$\left\|\mathbf{E}\left[\phi(u(\phi^Tx)-u(\phi^Tx^*))\right]\right\|_\infty=\|g(x)\|_\infty\le\bar r\sqrt{\nu}\|x-x^*\|_\Sigma.$$

Observe also that

$$\begin{aligned}\|\mathcal{G}(x,(\phi,\eta))-g(x)\|_\infty&=\left\|\phi\big(u(\phi^Tx)-u(\phi^Tx^*)\big)+\sigma\phi\zeta+[g(x^*)-g(x)]\right\|_\infty\\ &[\text{by (5.7.45)}]\le\|\phi\big(u(\phi^Tx)-u(\phi^Tx^*)\big)\|_\infty+\sigma\|\phi\zeta\|_\infty+\bar r\sqrt{\nu}\|x-x^*\|_\Sigma\\ &[\text{by (5.5.16)}]\quad\le\bar r\|\phi\|_\infty\left|\phi^Tx-\phi^Tx^*\right|+\sigma\|\phi\|_\infty|\zeta|+\bar r\sqrt{\nu}\|x-x^*\|_\Sigma.\end{aligned} \tag{5.7.49}$$

- Let us show first that

$$\mathbf{E}\left[\exp\left(\frac{\|\phi\|_\infty|\zeta|}{2.32\sqrt{\nu(1+\ln n)}}\right)\right]\le\exp(1). \tag{5.7.50}$$

Note that for $\zeta\sim\mathcal{SG}(0,1)$ and $s<\frac12$, $\mathbf{E}\left[e^{\frac{\zeta^2}{2.32}}\right]\le\exp(1)$.[3] Moreover,

$$\mathbf{E}\left[\exp\left(\frac{\|\phi\|_\infty^2}{2.32\nu}\right)\right]\le\sum_{i=1}^n\mathbf{E}\left[\exp\left(\frac{[\phi]_i^2}{2.32\nu}\right)\right]\le n\max_i\mathbf{E}\left[\exp\left(\frac{[\phi]_i^2}{2.32\nu}\right)\right]\le n\exp(1).$$

By convexity of the exponential function,

$$\mathbf{E}\left[\exp\left(\frac{\|\phi\|_\infty^2}{2.32\nu(1+\ln n)}\right)\right]\le\left(\mathbf{E}\left[\exp\left(\frac{\|\phi\|_\infty^2}{2.32\nu}\right)\right]\right)^{\frac{1}{1+\ln n}}\le\exp(1). \tag{5.7.51}$$

As a consequence, one has

$$\begin{aligned}\mathbf{E}\left[\exp\left(\frac{\|\phi\|_\infty|\zeta|}{2.32\sqrt{\nu(1+\ln n)}}\right)\right]&\le\mathbf{E}\left[\exp\left(\frac{\|\phi\|_\infty^2}{4.64\nu(1+\ln n)}+\frac{\zeta^2}{4.64}\right)\right]\\ \mathbf{E}\left[\exp\left(\frac{\|\phi\|_\infty^2}{4.64\nu(1+\ln n)}\right)\right]\mathbf{E}\left[\exp\left(\frac{\zeta^2}{4.64}\right)\right]&\le\exp(\tfrac12)\exp(\tfrac12)=\exp(1).\end{aligned}$$

[3] For the sake of completeness, here is one-line proof of this standard bound:

$$\mathbf{E}_\zeta\left[e^{\frac{\zeta^2}{2.32}}\right]=\mathbf{E}_\zeta\left[\mathbf{E}_{\eta\sim\mathcal{N}(0,1)}\left[e^{\frac{\zeta\eta}{\sqrt{1.16}}}\right]\right]=\mathbf{E}_{\eta\sim\mathcal{N}(0,1)}\left[\mathbf{E}_\zeta\left[e^{\frac{\zeta\eta}{\sqrt{1.16}}}\right]\right]\le\mathbf{E}_{\eta\sim\mathcal{N}(0,1)}\left[e^{\frac{\eta^2}{2.32}}\right]=\left(1-\tfrac{1}{1.16}\right)^{-1/2}\le\exp(1).$$

- Let us show next that

$$e(x) := \mathbf{E}\left[\exp\left(\frac{\|\phi[u(\phi^T x) - u(\phi^T x^*)]\|_\infty}{2.32\sqrt{\varkappa\nu(1+\ln n)}\bar{r}\|x-x^*\|_\Sigma}\right)\right] \le \exp(1).$$

Let us put

$$\alpha = \frac{\bar{r}\|x-x^*\|_\Sigma\sqrt{\varkappa}}{\sqrt{\nu(1+\ln n)}}.$$

One has

$$\|\phi[u(\phi^T x) - u(\phi^T x^*)]\|_\infty \le \bar{r}\|\phi\|_\infty|\phi^T(x-x^*)| \le \tfrac{\alpha}{2}\|\phi\|_\infty^2 + \tfrac{\bar{r}^2}{2\alpha}(\phi^T(x-x^*))^2.$$

Now observe that r.v. $\phi^T(x-x^*)$ is sub-Gaussian with zero mean and sub-Gaussianity parameter $\|x-x^*\|_S^2 \le \varkappa\|x-x^*\|_\Sigma^2$. Together with (5.7.51) this implies that

$$\begin{aligned} e(x) &\le \mathbf{E}\left[\exp\left(\frac{\|\phi\|_\infty^2}{4.64\nu(1+\ln n)} + \frac{(\phi^T(x-x^*))^2}{4.64\varkappa\bar{r}^2\|x-x^*\|_\Sigma^2}\right)\right] \\ &\le \mathbf{E}^{1/2}\left[\exp\left(\frac{\|\phi\|_\infty^2}{2.32\nu(1+\ln n)}\right)\right] + \mathbf{E}^{1/2}\left[\exp\left(\frac{(\phi^T(x-x^*))^2}{2.32\varkappa\bar{r}^2\|x-x^*\|_\Sigma^2}\right)\right] \\ &\le \exp(\tfrac{1}{2})\exp(\tfrac{1}{2}) = \exp(1). \end{aligned} \tag{5.7.52}$$

- We now set

$$\mu_1 = 2.32\sigma\nu(1+\ln n), \quad \mu_2 = 2.32\sqrt{\varkappa\nu(1+\ln n)}\bar{r}\|x-x^*\|_\Sigma.$$

From (5.7.51) and (5.7.52), we conclude that

$$\begin{aligned} &\mathbf{E}\left[\exp\left(\frac{\|\phi[u(\phi^T x) - u(\phi^T x^*)] + \sigma\phi\zeta\|_\infty}{2.32\sqrt{\nu(1+\ln n)}[\sigma + \bar{r}\|x-x^*\|_\Sigma\sqrt{\varkappa}]}\right)\right] \\ &\le \frac{\mu_1}{\mu_1+\mu_2}\mathbf{E}\left[\exp\left(\mu_1^{-1}\sigma\|\phi\|_\infty|\zeta|\right)\right] + \frac{\mu_2}{\mu_1+\mu_2}\mathbf{E}\left[\exp\left(\mu_2^{-1}\|\phi[u(\phi^T x) - u(\phi^T x^*)]\|_\infty\right)\right] \le \exp(1). \end{aligned}$$

Together with (5.7.49), the latter bound implies (5.7.47), (5.7.48).

4°. In the setting described in Section 5.5.4, Σ is a positive definite matrix, such that $\Sigma \succeq \kappa_\Sigma I$ with $\kappa_\Sigma > 0$, and condition $\mathbf{Q}(\lambda,\psi)$ is satisfied with $\lambda = \kappa_\Sigma$ and $\psi = 1$. Because quadratic minoration condition 5.5.19 of Lemma 5.5.1 for f is verified with $\mu \ge \underline{r}$ due to (5.5.18), we conclude that Assumption [RSC] holds with $\Upsilon = 1$ and $\rho = (\kappa_\Sigma\underline{r})^{-1}$.[4]

5.7.10 Condition $\mathbf{Q}(\lambda,\psi)$

We say that a positive semidefinite mapping $\Sigma : E \to E$ satisfies condition $\mathbf{Q}(\lambda,\psi)$ for given $s \in \mathbf{Z}_+$ if for some $\psi, \lambda > 0$ and all $P \in \mathcal{P}_s$ and $z \in E$

$$\|Pz\| \le \sqrt{s/\lambda}\|z\|_\Sigma + \|\overline{P}z\| - \psi\|z\|. \tag{5.7.53}$$

[4]We refer to Section 5.7.10 and Lemma 5.7.2 for the proof of Lemma 5.5.1.

Lemma 5.7.2 *Suppose that x^* is an optimal solution to 5.1.1 such that for some $P \in \mathcal{P}_s$, $\|(I-P)x_*\| \le \Delta$, and that condition $\boldsymbol{Q}(\lambda,\psi)$ is satisfied. Furthermore, assume that objective f of 5.1.1 satisfies the following growth condition*

$$f(x) - f(x^*) \ge \mu\big(\|x - x^*\|_\Sigma\big)$$

where $\mu(\cdot)$ is monotone increasing and convex. Then a feasible solution $\widehat{x} \in X$ to (5.5.12) such that

$$\mathrm{Prob}\left\{\Psi_\kappa(\widehat{x}) - \Psi_k(x^*) \le \upsilon\right\} \ge 1-\epsilon.$$

satisfies, with probability at least $1-\epsilon$,

$$\|\widehat{x} - x^*\| \le \frac{\frac{1}{2}\mu^*\left(\kappa\sqrt{s/\lambda}\right) + \upsilon}{\kappa\psi} + \frac{2\Delta}{\psi} \tag{5.7.54}$$

where $\mu^* : \mathbf{R}_+ \to \mathbf{R}_+$ is conjugate to $\mu(\cdot)$, $\mu^*(t) = \sup_{u\ge 0}[tu - \mu(u)]$.

Proof. When setting $z = \widehat{x} - x^*$ we have the following inequality

$$\begin{aligned}\|\widehat{x}\| = \|x^* + z\| &= \|Px^* + (I-P)x^* + z\| \\ &\ge \|Px^* + z\| - \|(I-P)x^*\| \\ &\ge \|Px^*\| + \|\overline{P}z\| - \|Pz\| - \Delta\end{aligned}$$

where we used the relation

$$\|Px^* + z\| \ge \|Px^*\| + \|\overline{P}z\| - \|Pz\|,$$

(Cf Lemma 3.1 of [120] applied to $w = Px^*$, here $\overline{P}$ is such that $P\overline{P} = 0$). When using condition $\mathbf{Q}(\lambda,\psi)$ we obtain

$$\|\widehat{x}\| \ge \|Px^*\| - \sqrt{s/\lambda}\|z\|_\Sigma + \psi\|z\| - \Delta,$$

such that $\Psi_\kappa(\widehat{x}) \le \Psi_\kappa(x^*) + \upsilon$ implies

$$\begin{aligned}\kappa\left(\|Px^*\| + \psi\|z\| - \Delta\right) &\le \tfrac{1}{2}(f(x^*) - f(\widehat{x})) + \kappa\sqrt{s/\lambda}\|z\|_\Sigma + \kappa\|x^*\| + \upsilon \\ &\le -\tfrac{1}{2}\mu\left(\|z\|_\Sigma\right) + \kappa\sqrt{s/\lambda}\|z\|_\Sigma + \kappa\|x^*\| + \upsilon \\ &\le \tfrac{1}{2}\mu^*\left(2\kappa\sqrt{s/\lambda}\right) + \kappa\|x^*\| + \upsilon.\end{aligned}$$

Due to $\|x_*\| - \|Px_*\| \le \|(I-P)x_*\| \le \Delta$ we then finally obtain

$$\|\widehat{x} - x^*\| \le \frac{\frac{1}{2}\mu^*\left(2\kappa\sqrt{s/\lambda}\right) + \upsilon}{\kappa\psi} + \frac{2\Delta}{\psi}.$$

■

Proof of Lemma 5.5.1 is a particular case of the proof of the preceding lemma. We obtain Lemma 5.5.1 by considering $\mu(x) = \frac{\mu}{2}x^2$ which leads to $\mu^*(t) = \frac{1}{2\mu}t^2$ and $\Delta = 0$.

Appendix

This appendix is devoted to provide additional information that can facilitate the reading of this thesis. We detail below useful assumptions used to provide the theoretical guarantees of each chapters of the manuscript. We also explain how to derive and implement the composite proximal operator, used in the CSMD and CSGE algorithms, in a time complexity of $\mathcal{O}(n)$.

5.8 Regularity in Convex Optimization.

We provide below the definition of several assumptions used in this manuscript. We consider an objective function g we aim at minimizing. Let E be a Euclidean space. A common assumption in the optimization literature is the convexity assumption. For clarity and consistency in our discussion, we provide here the specific definition of convexity that used in the manuscript.

Definition 5.8.1 *(Convex set) A set $X \subset E$ is said to be convex if $\lambda x + (1-\lambda)y \in X$, for any $x, y \in X$ and any $\lambda \in [0,1]$.*

Now we can define the notion of convexity of a function g, it will have many important implications in the analysis of the optimization algorithms.

Definition 5.8.2 *(Convex function) Let $X \subset E$ be a convex set. A function $g : X \to \mathbf{R}$ is said to be convex on X if it verifies $\forall x, y \in X$ and $\forall \lambda \in [0,1]$,*

$$g\left(\lambda x + (1-\lambda)y\right) \leq \lambda g(x) + (1-\lambda)g(y). \tag{5.8.1}$$

Function g is said to be stricly convex if the inequality is strict for $x \neq y$.

When g is differentiable, there exists an equivalent definition.

Definition 5.8.3 *Let $X \subset E$ be a convex set. A continuously differentiable function $g : X \to \mathbf{R}$ is said to be convex if it verifies $\forall x, y \in X$,*

$$g(x) \geq g(y) + \langle \nabla g(y), x - y \rangle. \tag{5.8.2}$$

An important implication is that whenever there is $x \in \mathbf{R}^n$ which is a local minimizer it is also a global minimizer. Convexity can be interprated geometrically as a linear lower-bound approximation of g at every point. Differentiability is a central concept in convex optimization. We introduce the notion of *subgradients* which generalizes the notion of differentiability for convex functions.

Definition 5.8.4 *(Subgradients) Let $g : X \to \mathbf{R}$ be a convex function. We say that c is a subgradient of g at point $x \in X$ if for all $y \in X$*

$$f(y) \geq f(x) + \langle c, y - x \rangle. \tag{5.8.3}$$

The set of all *subgradients* of g at a given point x is called the *subdifferential* of g at x and is denoted $\partial g(x)$. If g is differentiable at point x, then the subdifferential reduces to the singleton $\partial g(x) = \{\nabla g(x)\}$. In the case of the absolute value function $f : x \mapsto |x|$, we have for $x \neq 0$ that $\partial f(x) = \{-1, 1\}$ and for $x = 0$, the subdifferential of the absolute value function is the segment $\partial f(0) = [-1, 1]$.

Following the work of [34, 126, 156], the notion of convexity can be augmented by introducing the concept of *uniform convexity*.

Definition 5.8.5 *(Uniformly convex function) Let $X \subset E$ be a convex set and $p \geq 2$. A function $g : X \to \mathbf{R}$ is said to be (μ, p)-uniformly convex if there exists $\mu > 0$ such that $\forall x, y \in X$ and $\forall \lambda \in [0, 1]$*

$$g\left(\lambda x + (1-\lambda)y\right) \leq \lambda g(x) + (1-\lambda)g(y) - \frac{\mu\lambda(1-\lambda)}{p}\Big[(1-\lambda)^{p-1} + \lambda^p\Big]\|x-y\|^p. \tag{5.8.4}$$

Equivalently, the definition of the uniform convexity also holds for a continuously differentiable function g.

Definition 5.8.6 *Let $X \subset E$ be a convex set, $\mu > 0$, and $p \geq 2$. A continuously differentiable function $g : X \to \mathbf{R}$ is said to be (μ, p)-uniformly convex if there exists $\alpha > 0$ such that $\forall x, y \in X$,*

$$g(x) - g(y) - \langle \nabla g(y), x - y\rangle \geq \frac{\mu}{p}\|x-y\|^p. \tag{5.8.5}$$

A special case that has received significant attention is the $(\mu, 2)$-uniform convexity, also known as *μ-strong convexity*. This assumption imposes a quadratic lower bound on the objective function. Additionally, it assures the uniqueness of the optimal solution x^*, an important property for many optimization problems. Throughout this manuscript, we use a more relaxed variant of the uniform convexity assumption, a growth condition that provides a lower-bound on the suboptimality gap.

Definition 5.8.7 *Let $X \subset E$ be a convex set, $\mu > 0$, and $p \geq 2$. A continuously differentiable function $g : X \to \mathbf{R}$ is said to verify the (μ, p)-growth condition w.r.t. $\|\cdot\|$, if it verifies, $\forall x \in X$ and x^* the minimizer of g, the following condition :*

$$g(x) - g(x^*) \geq \frac{\mu}{p}\|x - x^*\|^p. \tag{5.8.6}$$

This definition relaxes the conditions of Definition 5.8.6, translating the notion of (μ, p)-uniform convexity to the neighborhood of the optimum x^*. Note that the particular case $p = 2$ is known in the optimization literature as the *quadratic growth condition*.

Given a norm $\|\cdot\|$ on E, we define its associated conjugate norm as $\|z\|_* := \sup\{\langle z, x\rangle : \|x\| \leq 1\}$. For instance, for $x \in \mathbf{R}^n$, the conjugate norm of the Euclidean norm $\|x\|_2 = \sqrt{x^T x}$ is itself. More generally, the dual norm of the ℓ_p-norm ($p \geq 1$) is the ℓ_q-norm with q such that $\frac{1}{p} + \frac{1}{q} = 1$.

We now turn our attention to another type of regularity assumptions in optimization: the smoothness condition.

Definition 5.8.8 *Let $X \subset E$ be a convex set, let also be $\|\cdot\|$ and $\|\cdot\|_*$ a given norm and its conjugate. A continuously differentiable function $g : X \to \mathbf{R}$ is said to be L-smooth if it verifies $\forall x, y \in X$,*

$$\|\nabla g(x) - \nabla g(y)\|_*^2 \leq L\|x-y\|^2. \tag{5.8.7}$$

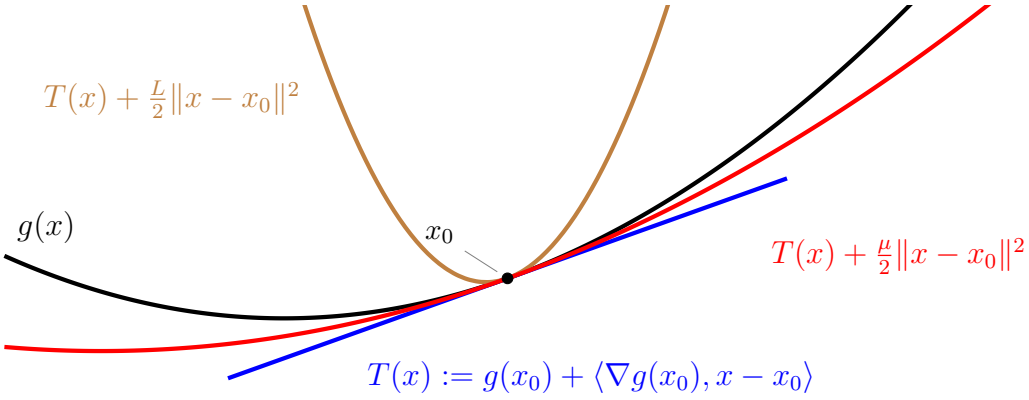


Figure 5.3: Geometric interpretation of the quadratic upper bound and quadratic lower bound provided respectively by L-smoothness (in brown) and μ-strong-convexity (in red) assumption of a function g at a point x_0.

Equivalently, the latter condition can be rewritten as

$$g(x) - g(y) - \langle \nabla g(y), x - y \rangle \leq \frac{L}{2}\|x - y\|^2. \tag{5.8.8}$$

Figure 5.3 provides a geometric interpretation of the notions of convexity and smoothness that have been previously discussed.

Remark 5.8.1 *Note that, for a twice continuously differentiable function g at a point x, the notion of L-smoothness and μ-strong convexity with respect to the Euclidean norm can also be expressed in terms of a relation on the Hessian of g denoted as $\nabla^2 g$. We have*

$$\mu I_n \preccurlyeq \nabla^2 g(x) \preccurlyeq L I_n,$$

where the relation $A \preccurlyeq B$ means that $B - A$ is a positive semi-definite matrix and I_n is the $n \times n$ identity matrix. This immediately introduces the quantity $\kappa = L/\mu$ often referred to as the condition number of g.

In the following development, we will introduce some important tools that will be used throughout the chapters of this thesis. We begin by defining the Fenchel-Legendre transform.

Definition 5.8.9 *(Fenchel-Legendre transform) The Fenchel-Legendre transform of a function $g : X \to \mathbf{R}$ denoted g^* is a function defined as*

$$g^*(w) = \sup_{x \in X} \{\langle w, x \rangle - g(x)\}. \tag{5.8.9}$$

This transform, also known as *Fenchel conjugate* has several useful properties. To name a few, the biconjugate of a function g denoted g^{**} (the convex conjugate of the convex conjugate of g) is the largest lower semi-continuous convex function that lower bounds g, i.e., $g^{**} \leq g$. Furthermore, if g is at least strictly-convex, then $\nabla g(\nabla g^*(w)) = w$ for $w \in X$ and $\nabla g^*(w) = \operatorname{argmax}_{x \in X}\{\langle w, x \rangle - g(x)\}$. Below we present some examples of *Legendre-Fenchel* transforms.

- Let $X = \mathbf{R}^n$, define for $A \in \mathcal{S}_n^{++}(\mathbf{R})$, $b \in \mathbf{R}^n$ and $c \in \mathbf{R}$, the quadratic function $g(x) = \frac{1}{2}x^T Ax + b^T x + c$. Then its convex conjugate is defined by $g^*(w) = \frac{1}{2}(w - b)^T A^{-1}(w - b) - c$.

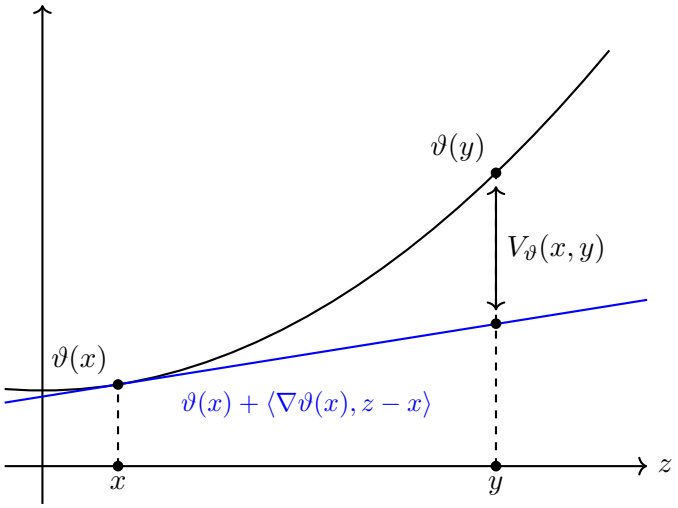


Figure 5.4: Geometric interpretation of the Bregman divergence.

- When $X = \mathcal{S}_n^{++}(\mathbf{R})$ and $g(Z) = -\ln\det Z$ we have $g^*(W) = -\ln\det(-W) - n$.

We will now formally introduce a widely used tool in the optimization literature, namely the *Bregman Divergence*. The Bregman Divergence is a measure used in optimization and information theory to quantify the difference between two points with respect to a strictly convex function ϑ. This concept is essential in the development of our methods and will be frequently be used throughout this manuscript.

Definition 5.8.10 *(Bregman Divergence) Let X be a closed convex set and $\vartheta : X \to \mathbf{R}$ a continuously differentiable stricly convex function. The Bregman divergence V_ϑ is defined according to*

$$\forall x, y \in X, \quad V_\vartheta(x,y) = \vartheta(y) - \vartheta(x) - \langle \nabla\vartheta(x), y - x\rangle; \tag{5.8.10}$$

The latter divergence can be interpreted as the difference between the *distance generating function* ϑ evaluated at a point y and the first-order Taylor approximation of ϑ at a point x evaluated at the point y. The Bregman divergence extends the interesting properties of the squared ℓ_2-norm to non-Euclidean spaces via the distance generating function ϑ. We mention below some examples.

- Squared Euclidean distance. We set $X = \mathbf{R}^n$ and $\vartheta(x) = \frac{1}{2}x^T x$, then the Bregman divergence is : $V_\vartheta(x,y) = \frac{1}{2}\|x - y\|_2^2$.
- Logistic loss divergence. We set $X = [0,1]^n$ and $\vartheta(x) = \sum_{i=1}^n (x_i \ln(x_i) + (1 - x_i)\ln(1 - x_i))$, the associated Bregman divergence is then $V_\vartheta(x,y) = \sum_{i=1}^n (y_i \ln\frac{y_i}{x_i} + (1 - y_i)\ln\frac{1-y_i}{1-x_i})$
- Matrix Entropy. We set $X = \mathcal{S}_n^{++}(\mathbf{R})$ and $\vartheta(Z) = \mathrm{tr}(Z\ln Z)$ for Z having the following eigenvalue decomposition $Z = \sum_{\lambda\in \mathrm{Sp(Z)}} \lambda q_\lambda q_\lambda^T$ such that $Zq_\lambda = \lambda q_\lambda$, we denote $\ln Z = \sum_{\lambda\in \mathrm{Sp(Z)}} \ln(\lambda) q_\lambda q_\lambda^T$. Then the Bregman divergence is written $V_\vartheta(Z,Y) = \mathrm{tr}(Y\ln Y - Y\ln Z - Y + Z)$.

One important property that will be used during the entire manuscript is the 3 points identity. This can be seen as the generalization to non-Euclidean spaces of the identity verified by the squared Euclidean distance

$$\frac{1}{2}\|x - z\|_2^2 = \frac{1}{2}\|x - y\|_2^2 + \frac{1}{2}\|y - z\|_2^2 + (x - y)^T(y - z).$$

When using a non-Euclidean distance generating function $\vartheta(\cdot)$, this can be generalized to the following relation

$$V_\vartheta(x,z) = V_\vartheta(x,y) + V_\vartheta(y,z) + \langle \nabla\vartheta(x) - \nabla\vartheta(y), y - z\rangle. \tag{5.8.11}$$

5.9 Composite proximal operator

In this section of the appendix, we propose a procedure aiming at computing the composite proximal operator that is the building block of our Composite Stochastic Mirror Descent method (algorithm 1) and Composite Stochastic Gradient Extrapolation method (algorithm 8). Recall that the main idea of our multistage routine is to solve a sequence of subproblems of the form (2.2.4) via either the CSMD or the CSGE algorithm. To that matter we have to compute a composite proximal operator of the form

$$\mathrm{Prox}_{\kappa\|\cdot\|_1,y}(\eta) := \operatorname*{argmin}_{\|z\|_1\leq 1} \left\{ \langle \eta, z\rangle + \kappa\|z+y\|_1 + \chi\|z\|_p^p \right\}, \tag{5.9.1}$$

where $\kappa, \chi > 0$, $p > 1$ and $y \in X$. To compute these types of proximal operator many methods are available. First, one can look for a closed form solution to the previous optimization problem, unfortunately in our setting, this option is not feasible. An other option is to solve the problem up to some precision with optimizers, the main drawback is that these optimizers scale poorly with the problem's dimension. For instance by using the CVX framework [157], based notably, on interior point methods [158], the time complexity is of order $\mathcal{O}(n^3)$ or $\mathcal{O}(n^{3.5})$ depending on the problem at hand, which in our high-dimensional setting is prohibitive. In what will follow, we will provide a procedure to solve this optimization problem, up to a prescribed precision, with a time complexity of order $\mathcal{O}(n)$.

We start by writing the Lagrangian of (5.9.1)

$$\mathcal{L}(z,\lambda) = \langle \eta, z\rangle + \kappa\|z+y\|_1 + \chi\|z\|_p^p + \lambda(\|z\|_1 - 1),$$

where $\lambda \geq 0$ is a Lagrangian multiplier. We then search at finding

$$\max_{\lambda\geq 0} \min_x \ \mathcal{L}(z,\lambda).$$

Observe that this function is separable, indeed we have $\mathcal{L}(z,\lambda) = \sum_{i=1}^n \mathcal{L}_i(z_i,\lambda)$, with

$$\forall i \in [n],\ \mathcal{L}_i(z_i,\lambda) := \eta_i z_i + \kappa|z_i + y_i| + \chi|z_i|^p + \lambda(|z_i| - \tfrac{1}{n}).$$

Instead of dealing with the full Lagrangian, we will concentrate on the $\mathcal{L}_i$'s instead. For the sake of readability, from now on, we will drop the indices of the variables.

The subdifferential of the above function is :

$$\partial_z\mathcal{L}(z,\lambda) := \eta + \kappa\ \mathrm{sign}(z+y) + \chi p\ \mathrm{sign}(z)|z|^{p-1} + \lambda\ \mathrm{sign}(z). \tag{5.9.2}$$

Satisfying the first order optimality condition boils down to finding $z^*(\lambda)$ for a fixed $\lambda \geq 0$ such that $0 \in \partial_z\mathcal{L}(z^*(\lambda),\lambda)$. We will now proceed the analysis by studying three cases, namely $y = 0$, $y > 0$ and $y < 0$.

Case $y = 0$: Here we distinguish three subcases, indeed observe that the first order optimality condition obtained with (5.9.2) rewrites

$$-\text{sign}\left(z^*(\lambda)\right)\left[(\kappa+\lambda)+\chi p|z^*(\lambda)|^{p-1}\right] \ni \eta.$$

- If $|\eta| < \kappa + \lambda$, we have $z^*(\lambda) = 0$.
- If $\eta > \kappa + \lambda$, then in particular, we have $\eta > 0$. For both the first-order optimality condition and the last inequality to hold, it is necessary that $z^*(\lambda) < 0$. This boils down to have $z^*(\lambda) = -\left(\frac{\eta-\kappa-\lambda}{\chi p}\right)^{\frac{1}{p-1}}$.
- If $\eta < -\kappa - \lambda$, now we have $\eta < 0$, and similarly for both the first-order optimality condition and the last inequality to hold, it is necessary that $z^*(\lambda) > 0$. This implies that $z^*(\lambda) = \left(-\frac{\eta+\kappa+\lambda}{\chi p}\right)^{\frac{1}{p-1}}$.

Case $y > 0$: In this case the first order optimality condition can be reformulated as follows

$$\eta + \kappa\ \text{sign}(z^*(\lambda)+y) + \chi p\ \text{sign}(z^*(\lambda))\ |z^*(\lambda)|^{p-1} + \lambda\ \text{sign}(z^*(\lambda)) \ni 0.$$

- If $|\eta - \chi p|y|^{p-1} - \lambda| \leq \kappa \implies z^*(\lambda) = -y$.
- If $\eta - \chi p|y|^{p-1} - \lambda - \kappa > 0$, then we have $z^*(\lambda) < -y$, which implies that $z^*(\lambda) = -\left(\frac{\eta-\kappa-\lambda}{\chi p}\right)^{\frac{1}{p-1}}$.
- Now if $\eta - \chi p|y|^{p-1} - \lambda + \kappa < 0$ we have $z^*(\lambda) + y > 0$. Contrary to the previous case, here we cannot directly conclude, we have to check the sign of $z^*(\lambda)$. This bring us to study three new subcases.
  - If $|\eta + \kappa| \leq \lambda$, then we immediately have that $z^*(\lambda) = 0$.
  - Otherwise, if $\eta + \kappa > \lambda$, in other words if we have $\eta + \kappa > 0$, then it implies that $z^*(\lambda) < 0$. By using the first order optimality condition we can show that we have $z^*(\lambda) = -\left(\frac{\eta+\kappa-\lambda}{\chi p}\right)^{\frac{1}{p-1}}$.
  - The last case to study is when $\eta + \kappa < -\lambda$ this implies that $\eta + \kappa < 0$ which ultimately gives that $z^*(\lambda) > 0$. We can then conclude that in this case $z^*(\lambda) = \left(-\frac{\eta+\kappa+\lambda}{\chi p}\right)^{\frac{1}{p-1}}$.

Case $y < 0$: The analysis of this case is very similar to the previous one. Here, the first order optimality condition remains as follows

$$\eta + \kappa\ \text{sign}(z^*(\lambda)+y) + \chi p\ \text{sign}(z^*(\lambda))\ |z^*(\lambda)|^{p-1} + \lambda\ \text{sign}(z^*(\lambda)) \ni 0.$$

- If $|\eta + \chi p|y|^{p-1} + \lambda| \leq \kappa \implies z^*(\lambda) + y = 0 \implies z^*(\lambda) = -y$.
- If $\eta + \chi p|y|^{p-1} + \lambda + \kappa < 0$, then we have $z^*(\lambda) > -y > 0$, which implies that $z^*(\lambda) = \left(-\frac{\eta+\kappa+\lambda}{\chi p}\right)^{\frac{1}{p-1}}$.
- Now if $\eta + \chi p|y|^{p-1} + \lambda - \kappa > 0$ we have $z^*(\lambda) < -y$. Here again we cannot directly conclude and we have to check the sign of $z^*(\lambda)$. Similarly to the previous case, this bring us to study three subcases.

- If $|\eta - \kappa| \leq \lambda$, then we immediately have that $z^*(\lambda) = 0$.
- Now, if $\eta - \kappa < -\lambda$, in other words if we have $\eta - \kappa < 0$, then we have $z^*(\lambda) > 0$. By using the first order optimality condition we can show that we have $z^*(\lambda) = \left(\frac{\kappa - \eta - \lambda}{\chi p}\right)^{\frac{1}{p-1}}$.
- Finally the last case to study is $\eta - \kappa > \lambda$, this implies that $\eta - \kappa > 0$ which ultimately gives that $z^*(\lambda) < 0$. In this case $z^*(\lambda) = -\left(\frac{\eta - \kappa - \lambda}{\chi p}\right)^{\frac{1}{p-1}}$.

After finding the components $z_i^*(\lambda)$ of the vector $z^*(\lambda)$ for a fixed non-negative λ, the next idea is to find $\lambda^* \in \operatorname{argmax}_{\lambda \geq 0} \mathcal{L}(z^*(\lambda), \lambda)$. To do so, we will use the complementary slackness condition. In our context the latter states that λ^* is such that $\lambda^*(\|z^*(\lambda^*)\|_1 - 1) = 0$. So either, $\lambda^* = 0$, meaning that the inequality constraints is already verified, or $\|z^*(\lambda^*)\|_1 - 1 = 0$. To summarize, our strategy to compute a composite proximal operator of the form (5.9.1) is to first check if $\|z^*(0)\|_1 - 1$ is equal to zero (up to a tolerance level), otherwise we iteratively search for the root of the function $\lambda \mapsto \|z^*(\lambda)\|_1 - 1$ by using the bissection method for a fixed number of iterations. It is easy to see that such a procedure comes with a total time complexity of order $\mathcal{O}(n)$.

List of Algorithms

1 Composite Stochastic Mirror Descent for Sparse Recovery (CSMD-SR) 51
2 Asymptotic phase of CSMD-SR with minibatch 52
3 Stochastic Accelerated Gradient Descent method (SAGD) 104
4 Stochastic Gradient Extrapolation method (SGE) 107
5 Multi-stage stochastic gradient extrapolation method 110
6 Stochastic Gradient Extrapolation method for Sparse Recovery (SGE-SR) 114
7 Shrinking multi-stage Stochastic Gradient Extrapolation method 140
8 Composite Stochastic Gradient Extrapolation method (CSGE) 145
9 Composite Stochastic Gradient Extrapolation method for Sparse Recovery (CSGE-SR) 148

List of Figures

1.1 Geometric interpretation of Stochastic Mirror Descent algorithm. 27
1.2 Visualization of the update of NAG algorithm. 30
1.3 *Left:* Sublinear convergence of an optimization method. *Right:* Faster convergence thanks to restart scheme. From d'Aspremont et al. [62]. 31
1.4 Sparse linear recovery problem. 36

2.1 Given the activation function $\mathfrak{r}$ in (2.3.5) and $\alpha = (0, 0.01, 0.1, 0.25, 1)$; left plot: mappings h; right plot: moduli of strong monotonicity of mappings H on $\{z : \|z\|_2 \leq r\}$ as function of r. 54
2.2 Comparison between CSMD-SR and baseline algorithms in Generalized Linear Regression problem: ℓ_1 error as a function of the number of oracle calls 56
2.3 Preliminary stages of the CSMD-SR and its variant with data recycling: linear regression experiment (left pane), GLR with activation $\mathfrak{r}_{1/10}(t)$ (right pane). 57
2.4 CSMD-SR and "vanilla" SMD in Linear Regression problem (activation function $\mathfrak{r}(t) = t$); ℓ_1 error as a function of the number of oracle calls 73
2.5 CSMD-SR and "vanilla" SMD in Generalized Linear Regression problem: activation function $\mathfrak{r}_{1/10}(t)$; ℓ_1 error as a function of the number of oracle calls 75

4.1 Left plot: variance of the stochastic oracle $\mathcal{G}_1(x, \xi)$ as function of x (solid line) and upper bound $4\sigma^2(0) + 3[f(x) - f(0)]$ (dashed line); right plot: variance of $\mathcal{G}_2(x, \xi)$ as function of x (solid line) and upper bound $2.3\sigma^2(0) + 3[f(x) - f(0)]$ (dashed line). . 103
4.2 Activation functions 115
4.3 Estimation error $\|x_t - x^*\|_2$ against the number of stochastic oracle calls for SGE-SR and SMD-SR algorithms. In the left, middle, and right columns of the plot we show results for the linear activation u_1, and nonlinear $u_{1/2}$ and $u_{1/10}$, respectively. Two figure rows correspond to two different noise levels, $\sigma = 0.1$ (the upper row) and $\sigma = 0.001$ (the bottom row). The legend specifies the value m_0 of the batch size of the preliminary phase of the algorithm for both routines. 116
4.4 Estimation error $\|x_t - x^*\|_2$ against the number of stochastic oracle calls for SGE-SR in Gaussian (light-tail) and Student t_5 (heavy-tail) regressor and noise generation setups. In the left, middle, and right columns of the plot we show results for the linear activation u_1, and nonlinear $u_{1/2}$ and $u_{1/10}$, respectively. Two figure rows correspond to two different noise levels, $\sigma = 0.1$ (the upper row) and $\sigma = 0.001$ (the bottom row). The legend specifies the value m_0 of the batch size of the preliminary phase of the algorithm for both routines. 117
4.5 Error $\|x_t - x^*\|_2$ of the SGE-SR algorithm for three values of the problem condition number. First row: algorithm error against the number of oracle calls; second row: error against the number of algorithm iterations. Figure columns correspond to the results for u_1, $u_{1/2}$, and $u_{1/10}$ activation functions and $\sigma = 0.001$. 118

4.6 SGE-SR compared to "vanilla" SMD-SR and its mini-batch variant. First row: error $\|x_t - x^*\|_2$ against the number of oracle calls; second row: the error against the number of algorithm iterations. Figure columns correspond to the results for u_1, $u_{1/2}$, and $u_{1/10}$ activation functions and $\sigma = 0.001$. 119

5.1 Estimation error $\|x_t - x^*\|_2$ against the number of stochastic oracle calls (top row) and against the number of algorithms iterations (bottom row) for SMD-SR, SGE-SR, CSMD-SR and CSGE-SR algorithms. In the left, middle, and right columns of the plot we show the results for the linear activation function u_1, and the nonlinear activation functions $u_{1/2}$ and $u_{1/10}$, respectively. The noise level is set to $\sigma = 0.1$. The legend specifies the value m_0 of the batch size of the preliminary phase of the algorithm for every routines. 152

5.2 Estimation error $\|x_t - x^*\|_2$ against the number of stochastic oracle calls (top row) and against the number of algorithms iterations (bottom row) for the CSGE-SR algorithms. In the left, middle, and right columns of the plot we show the results for the linear activation function u_1, and the nonlinear activation functions $u_{1/2}$ and $u_{1/10}$, respectively. The noise level is set to $\sigma = 0.001$. The legend specifies the value m_0 of the batch size of the preliminary phase of the algorithm for each routine and κ denotes the condition number. 153

5.3 Geometric interpretation of the quadratic upper bound and quadratic lower bound provided respectively by L-smoothness (in brown) and μ-strong-convexity (in red) assumption of a function g at a point x_0. 181

5.4 Geometric interpretation of the Bregman divergence. 182

Bibliography

1. Ilandarideva, S., Bekri, Y., Iouditski, A. & Perchet, V. *Stochastic Mirror Descent for Large-Scale Sparse Recovery* in *International Conference on Artificial Intelligence and Statistics* (2023), 5931–5957 (Cited on pages 15, 99, 111, 147, 151).
2. Ilandarideva, S., Juditsky, A., Lan, G. & Li, T. Accelerated stochastic approximation with state-dependent noise. *arXiv preprint arXiv:2307.01497* (2023) (Cited on pages 16, 151).
3. Xia, G.-S. *et al. DOTA: A large-scale dataset for object detection in aerial images* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), 3974–3983 (Cited on page 19).
4. Vaswani, A. *et al.* Attention is all you need. *Advances in neural information processing systems* **30** (2017) (Cited on page 19).
5. Hastie, T., Tibshirani, R., Friedman, J. H. & Friedman, J. H. *The elements of statistical learning: data mining, inference, and prediction* (Springer, 2009) (Cited on page 20).
6. Vapnik, V. *The nature of statistical learning theory* (Springer science & business media, 1999) (Cited on page 21).
7. Groetsch, C. The theory of Tikhonov regularization for Fredholm equations. *104p, Boston Pitman Publication* (1984) (Cited on page 21).
8. Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **58,** 267–288 (1996) (Cited on page 21).
9. Nelder, J. A. & Wedderburn, R. W. Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society* **135,** 370–384 (1972) (Cited on page 22).
10. McCullagh, P. *Generalized linear models* (Routledge, 2019) (Cited on page 22).
11. Hsu, D. & Sabato, S. Loss minimization and parameter estimation with heavy tails. *The Journal of Machine Learning Research* **17,** 543–582 (2016) (Cited on page 23).
12. Gurbuzbalaban, M., Simsekli, U. & Zhu, L. *The heavy-tail phenomenon in SGD* in *International Conference on Machine Learning* (2021), 3964–3975 (Cited on page 23).
13. Gorbunov, E., Danilova, M., Shibaev, I., Dvurechensky, P. & Gasnikov, A. Near-optimal high probability complexity bounds for non-smooth stochastic optimization with heavy-tailed noise. *arXiv preprint arXiv:2106.05958* (2021) (Cited on page 23).
14. Lou, Z., Zhu, W. & Wu, W. B. Beyond sub-gaussian noises: Sharp concentration analysis for stochastic gradient descent. *The Journal of Machine Learning Research* **23,** 2227–2248 (2022) (Cited on page 23).
15. Robbins, H. & Monro, S. A stochastic approximation method. *The annals of mathematical statistics,* 400–407 (1951) (Cited on page 24).

16. Duchi, J., Hazan, E. & Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research* **12** (2011) (Cited on page 25).

17. Ghadimi, S. & Lan, G. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization* **23,** 2341–2368 (2013) (Cited on page 25).

18. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014) (Cited on pages 25, 31).

19. Robinson, S. M. Analysis of sample-path optimization. *Mathematics of Operations Research* **21,** 513–528 (1996) (Cited on page 25).

20. Shapiro, A. & Wardi, Y. Convergence analysis of stochastic algorithms. *Mathematics of operations research* **21,** 615–628 (1996) (Cited on page 25).

21. Shapiro, A., Homem-de-Mello, T. & Kim, J. Conditioning of convex piecewise linear stochastic programs. *Mathematical Programming* **94,** 1–19 (2002) (Cited on page 25).

22. Kleywegt, A. J., Shapiro, A. & Homem-de-Mello, T. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on optimization* **12,** 479–502 (2002) (Cited on page 25).

23. Nesterov, Y. Gradient methods for minimizing composite functions. *Mathematical programming* **140,** 125–161 (2013) (Cited on pages 26, 30, 37, 47, 48).

24. Lan, G. An optimal method for stochastic composite optimization. *Mathematical Programming* **133,** 365–397 (2012) (Cited on pages 26, 30, 98, 99).

25. Ghadimi, S. & Lan, G. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization* **22,** 1469–1492 (2012) (Cited on pages 26, 27, 31, 98, 111).

26. Ghadimi, S. & Lan, G. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, II: shrinking procedures and optimal algorithms. *SIAM Journal on Optimization* **23,** 2061–2089 (2013) (Cited on pages 26, 31, 38, 46, 98).

27. Kulunchakov, A. & Mairal, J. *Estimate sequences for variance-reduced stochastic composite optimization* in *International Conference on Machine Learning* (2019), 3541–3550 (Cited on page 26).

28. Kulunchakov, A. & Mairal, J. A generic acceleration framework for stochastic composite optimization. *Advances in Neural Information Processing Systems* **32** (2019) (Cited on page 26).

29. Kulunchakov, A. & Mairal, J. Estimate sequences for stochastic composite optimization: Variance reduction, acceleration, and robustness to noise. *The Journal of Machine Learning Research* **21,** 6184–6235 (2020) (Cited on page 26).

30. Nemirovski, A. Efficient methods for large-scale convex optimization problems. *Ekonomika i Matematicheskie Metody* **15** (1979) (Cited on page 26).

31. Nemirovski, A. & Yudin, D. B. Problem complexity and method efficiency in optimization (1983) (Cited on pages 26, 29).

32. Bubeck, S. *et al.* Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning* **8,** 231–357 (2015) (Cited on page 26).

33. Nemirovski, A., Juditsky, A., Lan, G. & Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization* **19,** 1574–1609 (2009) (Cited on pages 27, 28, 48, 98).

34. Juditsky, A. & Nesterov, Y. Primal-dual subgradient methods for minimizing uniformly convex functions. *arXiv preprint arXiv:1401.1792* (2014) (Cited on pages 27, 180).

35. Teboulle, M. Entropic proximal mappings with applications to nonlinear programming. *Mathematics of Operations Research* **17,** 670–690 (1992) (Cited on page 27).

36. Kivinen, J. & Warmuth, M. K. Exponentiated gradient versus gradient descent for linear predictors. *information and computation* **132,** 1–63 (1997) (Cited on page 28).

37. Juditsky, A. & Nemirovski, A. First order methods for nonsmooth convex large-scale optimization, i: general purpose methods. *Optimization for Machine Learning* **30,** 121–148 (2011) (Cited on pages 28, 46, 48, 54, 111, 150).

38. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* http://www.deeplearningbook.org (MIT Press, 2016) (Cited on page 28).

39. Azizan, N., Lale, S. & Hassibi, B. Stochastic mirror descent on overparameterized nonlinear models. *IEEE Transactions on Neural Networks and Learning Systems* **33,** 7717–7727 (2021) (Cited on page 28).

40. Garcia, J. R. *et al. Fisher-Legendre (FishLeg) optimization of deep neural networks* in *The Eleventh International Conference on Learning Representations* (2022) (Cited on page 28).

41. Yang, L. *et al. Policy optimization with stochastic mirror descent* in *Proceedings of the AAAI Conference on Artificial Intelligence* **36** (2022), 8823–8831 (Cited on page 28).

42. Lan, G. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical programming* **198,** 1059–1106 (2023) (Cited on page 28).

43. Jin, Y. & Sidford, A. *Efficiently solving MDPs with stochastic mirror descent* in *International Conference on Machine Learning* (2020), 4890–4900 (Cited on page 28).

44. Yu, J., Aberdeen, D. & Schraudolph, N. Fast online policy gradient learning with SMD gain vector adaptation. *Advances in neural information processing systems* **18** (2005) (Cited on page 28).

45. Zinkevich, M. *Online convex programming and generalized infinitesimal gradient ascent* in *Proceedings of the 20th international conference on machine learning (icml-03)* (2003), 928–936 (Cited on page 28).

46. Srebro, N., Sridharan, K. & Tewari, A. On the universality of online mirror descent. *Advances in neural information processing systems* **24** (2011) (Cited on page 28).

47. Orabona, F., Crammer, K. & Cesa-Bianchi, N. A generalized online mirror descent with applications to classification and regression. *Machine Learning* **99,** 411–435 (2015) (Cited on page 28).

48. Juditsky, A. & Nesterov, Y. Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization. *Stochastic Systems* **4,** 44–80 (2014) (Cited on pages 28, 30, 38, 46, 77, 84, 111).

49. Lan, G. *First-order and stochastic optimization methods for machine learning* (Springer, 2020) (Cited on pages 28, 29).

50. Ruppert, D. *Efficient estimations from a slowly convergent Robbins-Monro process* tech. rep. (Cornell University Operations Research and Industrial Engineering, 1988) (Cited on page 29).

51. Polyak, B. T. & Juditsky, A. B. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization* **30,** 838–855 (1992) (Cited on pages 29, 98).

52. Polyak, B. T. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics* **4,** 1–17 (1964) (Cited on page 29).

53. Hestenes, M. R., Stiefel, E., *et al.* Methods of conjugate gradients for solving linear systems. *Journal of research of the National Bureau of Standards* **49,** 409–436 (1952) (Cited on page 29).

54. Nesterov, Y. E. *A method for solving the convex programming problem with convergence rate* $O(1/k^2)$ in *Dokl. akad. nauk Sssr* **269** (1983), 543–547 (Cited on pages 30, 98).

55. Tseng, P. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization* **2** (2008) (Cited on page 30).

56. Beck, A. & Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences* **2,** 183–202 (2009) (Cited on pages 30, 36).

57. Becker, S. R., Candès, E. J. & Grant, M. C. Templates for convex cone problems with applications to sparse signal recovery. *Mathematical programming computation* **3,** 165–218 (2011) (Cited on page 30).

58. O'donoghue, B. & Candes, E. Adaptive restart for accelerated gradient schemes. *Foundations of computational mathematics* **15,** 715–732 (2015) (Cited on page 30).

59. Juditsky, A., Kulunchakov, A. & Tsyntseus, H. Sparse recovery by reduced variance stochastic approximation. *Information and Inference: A Journal of the IMA* **12,** 851–896 (2023) (Cited on pages 30, 37–40, 56, 98, 99, 111, 114, 116, 134, 151).

60. Zeiler, M. D. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701* (2012) (Cited on page 31).

61. Dozat, T. Incorporating nesterov momentum into adam (2016) (Cited on page 31).

62. d'Aspremont, A., Scieur, D., Taylor, A., *et al.* Acceleration methods. *Foundations and Trends® in Optimization* **5,** 1–245 (2021) (Cited on page 31).

63. Donoho, D. L. Compressed sensing. *IEEE Transactions on information theory* **52,** 1289–1306 (2006) (Cited on page 32).

64. Candès, E. J. *et al. Compressive sampling* in *Proceedings of the international congress of mathematicians* **3** (2006), 1433–1452 (Cited on page 32).

65. Candes, E. J., Romberg, J. K. & Tao, T. 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,** 1207–1223 (2006) (Cited on pages 32, 46, 111).

66. Candes, E. J. & Tao, T. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE transactions on information theory* **52,** 5406–5425 (2006) (Cited on page 32).

67. Candes, E. & Tao, T. The Dantzig selector: Statistical estimation when p is much larger than n (2007) (Cited on pages 32, 34, 46, 111).

68. Candes, E. J. The restricted isometry property and its implications for compressed sensing. *Comptes rendus. Mathematique* **346,** 589–592 (2008) (Cited on page 32).

69. Natarajan, B. K. Sparse approximate solutions to linear systems. *SIAM journal on computing* **24,** 227–234 (1995) (Cited on page 33).

70. Mallat, S. G. & Zhang, Z. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on signal processing* **41,** 3397–3415 (1993) (Cited on page 33).

71. Chen, S. S., Donoho, D. L. & Saunders, M. A. Atomic decomposition by basis pursuit. *SIAM review* **43,** 129–159 (2001) (Cited on page 33).

72. Donoho, D. L. Neighborly polytopes and sparse solution of underdetermined linear equations. *preprint,* 107 (2004) (Cited on page 33).

73. Juditsky, A. & Nemirovski, A. On verifiable sufficient conditions for sparse signal recovery via ℓ_1-minimization. *Mathematical programming* **127,** 57–88 (2011) (Cited on page 33).

74. Candes, E. J. & Tao, T. Decoding by linear programming. *IEEE transactions on information theory* **51,** 4203–4215 (2005) (Cited on page 34).

75. Bickel, P. J., Ritov, Y. & Tsybakov, A. B. Simultaneous analysis of Lasso and Dantzig selector (2009) (Cited on pages 34, 46, 52, 53, 55, 111, 150).

76. Lounici, K. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators (2008) (Cited on page 34).

77. Van De Geer, S. A. & Bühlmann, P. On the conditions used to prove oracle results for the lasso (2009) (Cited on pages 34, 46, 53, 150).

78. Bunea, F., Tsybakov, A. & Wegkamp, M. Sparsity oracle inequalities for the Lasso (2007) (Cited on page 34).

79. Zhang, C.-H. & Huang, J. The sparsity and bias of the lasso selection in high-dimensional linear regression (2008) (Cited on page 34).

80. Wainwright, M. J. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint* (Cambridge University Press, 2019) (Cited on page 34).

81. Raskutti, G., Wainwright, M. J. & Yu, B. Restricted eigenvalue properties for correlated Gaussian designs. *The Journal of Machine Learning Research* **11,** 2241–2259 (2010) (Cited on pages 35, 37, 52, 53).

82. Van de Geer, S. A. & Bühlmann, P. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics* **3,** 1360–1392. https://doi.org/10.1214/09-EJS506 (2009) (Cited on page 35).

83. Juditsky, A. & Nemirovski, A. Accuracy Guarantees for ℓ_1-Recovery. *IEEE Transactions on Information Theory* **57,** 7818–7839 (2011) (Cited on pages 35, 52, 111, 150).

84. Blumensath, T. & Davies, M. E. Iterative hard thresholding for compressed sensing. *Applied and computational harmonic analysis* **27,** 265–274 (2009) (Cited on pages 35, 46, 113).

85. Jain, P., Tewari, A. & Kar, P. On iterative hard thresholding methods for high-dimensional m-estimation. *Advances in neural information processing systems* **27** (2014) (Cited on pages 35, 46, 113).

86. Liu, H. & Foygel Barber, R. Between hard and soft thresholding: optimal iterative thresholding algorithms. *Information and Inference: A Journal of the IMA* **9,** 899–933 (2020) (Cited on pages 35, 46, 113).

87. Srebro, N., Sridharan, K. & Tewari, A. Smoothness, low noise and fast rates. *Advances in neural information processing systems* **23** (2010) (Cited on pages 36, 46, 98, 111, 134).

88. Agarwal, A., Negahban, S. & Wainwright, M. J. Stochastic optimization and sparse statistical recovery: Optimal algorithms for high dimensions. *Advances in Neural Information Processing Systems* **25** (2012) (Cited on pages 37, 38, 46, 47, 111, 134).

89. Xiao, L. Dual averaging method for regularized stochastic learning and online optimization. *Advances in Neural Information Processing Systems* **22** (2009) (Cited on pages 37, 56).

90. Sedghi, H., Anandkumar, A. & Jonckheere, E. Multi-step stochastic ADMM in high dimensions: Applications to sparse optimization and matrix decomposition. *Advances in neural information processing systems* **27** (2014) (Cited on page 38).

91. Lepskii, O. On a problem of adaptive estimation in Gaussian white noise. *Theory of Probability & Its Applications* **35,** 454–466 (1991) (Cited on pages 39, 80).

92. Lepskii, O. Asymptotically minimax adaptive estimation. i: Upper bounds. optimally adaptive estimates. *Theory of Probability & Its Applications* **36,** 682–697 (1992) (Cited on page 39).

93. Lepskii, O. Asymptotically minimax adaptive estimation. II. Schemes without optimal adaptation: Adaptive estimators. *Theory of Probability & Its Applications* **37,** 433–448 (1993) (Cited on page 39).

94. Candes, E. J. & Plan, Y. Near-ideal model selection by ℓ_1 minimization. *Annals of Statistics* **37,** 2145–2177 (2009) (Cited on pages 46, 111).

95. Candes, E. J. & Plan, Y. A probabilistic and RIPless theory of compressed sensing. *IEEE transactions on information theory* **57,** 7235–7254 (2011) (Cited on pages 46, 111).

96. Meier, L., Van De Geer, S. & Bühlmann, P. The group lasso for logistic regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **70,** 53–71 (2008) (Cited on pages 46, 55).

97. Baes, M., Burgisser, M. & Nemirovski, A. A randomized mirror-prox method for solving structured large-scale matrix saddle-point problems. *SIAM Journal on Optimization* **23,** 934–962 (2013) (Cited on page 46).

98. Juditsky, A., Kılınç Karzan, F. & Nemirovski, A. Randomized first order algorithms with applications to ℓ_1-minimization. *Mathematical Programming* **142,** 269–310 (2013) (Cited on page 46).

99. Ghaoui, L. E., Viallon, V. & Rabbani, T. Safe feature elimination for the lasso and sparse supervised learning problems. *arXiv preprint arXiv:1009.4219* (2010) (Cited on page 46).

100. Kowalski, M., Weiss, P., Gramfort, A. & Anthoine, S. *Accelerating ISTA with an active set strategy* in *OPT 2011: 4th International Workshop on Optimization for Machine Learning* (2011) (Cited on page 46).

101. Mairal, J. *Sparse coding for machine learning, image processing and computer vision* PhD thesis (École normale supérieure de Cachan-ENS Cachan, 2010) (Cited on page 46).

102. Agarwal, A., Negahban, S. & Wainwright, M. J. Fast global convergence of gradient methods for high-dimensional statistical recovery. *The Annals of Statistics* **40,** 2452–2482. https://doi.org/10.1214/12-AOS1032 (2012) (Cited on pages 46, 55).

103. Barber, R. F. & Ha, W. Gradient descent with non-convex constraints: local concavity determines convergence. *Information and Inference: A Journal of the IMA* **7,** 755–806 (2018) (Cited on page 46).

104. Nemirovski, A. & Yudin, D. Complexity of problems and effectiveness of methods of optimization(Russian book). *Moscow, Izdatel'stvo Nauka, 1979. 384* (1979) (Cited on pages 46, 98).

105. Shalev-Shwartz, S. & Tewari, A. *Stochastic methods for ℓ_1 regularized loss minimization* in *Proceedings of the 26th Annual International Conference on Machine Learning* (2009), 929–936 (Cited on pages 46, 111, 134).

106. Gaillard, P. & Wintenberger, O. *Sparse accelerated exponential weights* in *Artificial Intelligence and Statistics* (2017), 75–82 (Cited on pages 46, 47, 111).

107. Juditsky, A. & Nemirovski, A. First order methods for nonsmooth convex large-scale optimization, ii: utilizing problems structure. *Optimization for Machine Learning* **30,** 149–183 (2011) (Cited on page 47).

108. Lei, Y. & Tang, K. Stochastic composite mirror descent: Optimal bounds with high probabilities. *Advances in Neural Information Processing Systems* **31** (2018) (Cited on page 47).

109. Nesterov, Y. & Nemirovski, A. On first-order algorithms for ℓ_1/nuclear norm minimization. *Acta Numerica* **22,** 509–575 (2013) (Cited on pages 48, 58, 101).

110. Dalalyan, A. & Thompson, P. Outlier-robust estimation of a sparse linear model using ℓ_1-penalized Huber's M-estimator. *Advances in neural information processing systems* **32** (2019) (Cited on page 52).

111. Cohen, A., Dahmen, W. & DeVore, R. Compressed sensing and best k-term approximation. *Journal of the American mathematical society* **22,** 211–231 (2009) (Cited on page 53).

112. Rauhut, H. Compressive sensing and structured random matrices. *Theoretical foundations and numerical methods for sparse recovery* **9,** 92 (2010) (Cited on page 53).

113. Bercu, B., Delyon, B. & Rio, E. *Concentration inequalities for sums and martingales* (Springer, 2015) (Cited on pages 55, 63, 134).

114. Chen, G. & Teboulle, M. Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM Journal on Optimization* **3,** 538–543 (1993) (Cited on page 58).

115. Juditsky, A. & Nemirovski, A. S. Large deviations of vector-valued martingales in 2-smooth normed spaces. *arXiv preprint arXiv:0809.0813* (2008) (Cited on pages 59, 64).

116. Laurent, B. & Massart, P. Adaptive estimation of a quadratic functional by model selection. *Annals of statistics,* 1302–1338 (2000) (Cited on page 62).

117. Birgé, L. & Massart, P. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli,* 329–375 (1998) (Cited on page 62).

118. Fan, X., Grama, I. & Liu, Q. Hoeffding's inequality for supermartingales. *Stochastic Processes and their Applications* **122,** 3545–3559 (2012) (Cited on pages 63, 134).

119. Nesterov, Y. Primal-dual subgradient methods for convex problems. *Mathematical programming* **120,** 221–259 (2009) (Cited on page 64).

120. Juditsky, A., Karzan, F. K. & Nemirovski, A. On a unified view of nullspace-type conditions for recoveries associated with general sparsity structures. *Linear Algebra and its Applications* **441,** 124–151 (2014) (Cited on pages 70, 72, 177).

121. Recht, B., Xu, W. & Hassibi, B. Null space conditions and thresholds for rank minimization. *Mathematical programming* **127,** 175–202 (2011) (Cited on page 72).

122. Nesterov, Y. Universal gradient methods for convex optimization problems. *Mathematical Programming* **152,** 381–404 (2015) (Cited on page 77).

123. Bach, F. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *The Journal of Machine Learning Research* **15,** 595–627 (2014) (Cited on page 77).

124. Roulet, V. & d'Aspremont, A. Sharpness, restart and acceleration. *Advances in Neural Information Processing Systems* **30** (2017) (Cited on pages 78, 84).

125. Polyak, B. T. *Existence theorems and convergence of minimizing sequences for extremal problems with constraints* in *Doklady Akademii Nauk* **166** (1966), 287–290 (Cited on page 84).

126. Azé, D. & Penot, J.-P. *Uniformly convex and uniformly smooth convex functions* in *Annales de la Faculté des sciences de Toulouse: Mathématiques* **4** (1995), 705–730 (Cited on pages 84, 180).

127. Vladimirov, A., Nesterov, Y. E. & Chekanov, Y. N. On uniformly convex functionals. *Vestnik Moskov. Univ. Ser. XV Vychisl. Mat. Kibernet* **3,** 12–23 (1978) (Cited on page 84).

128. Polyak, B. T. New stochastic approximation type procedures. *Automat. i Telemekh* **7,** 2 (1990) (Cited on page 98).

129. Lan, G. *First-order and stochastic optimization methods for machine learning* (Springer, 2020) (Cited on pages 98, 100, 107, 108, 120).

130. Yudin, D. & Nemirovski, A. Computational complexity of strictly convex programming. *Eknomika i Matematicheskie Metody* **3,** 550–569 (1977) (Cited on page 98).

131. Nemirovskii, A. & Yudin, D. Information-based complexity of Mathematical Programming. *Engineering Cybernetics* **1,** 76–100 (1983) (Cited on page 98).

132. Nemirovski, A. S. & Yudin, D. *Problem complexity and method efficiency in optimization* (John Wiley, XV, 1983) (Cited on pages 98, 100).

133. Nemirovskii, A. & Nesterov, Y. Optimal methods of smooth convex minimization. *USSR Computational Mathematics and Mathematical Physics* **25,** 21–30 (1985) (Cited on page 98).

134. Bietti, A. & Mairal, J. Stochastic optimization with variance reduction for infinite datasets with finite sum structure. *Advances in Neural Information Processing Systems* **30** (2017) (Cited on page 99).

135. Gower, R. M. *et al. SGD: General analysis and improved rates* in *International conference on machine learning* (2019), 5200–5209 (Cited on page 99).

136. Gower, R., Sebbouh, O. & Loizou, N. *Sgd for structured nonconvex functions: Learning rates, minibatching and interpolation* in *International Conference on Artificial Intelligence and Statistics* (2021), 1315–1323 (Cited on page 99).

137. Khaled, A. & Richtárik, P. Better Theory for SGD in the Nonconvex World. *Transactions on Machine Learning Research* (2022) (Cited on page 99).

138. Woodworth, B. E. & Srebro, N. An even more optimal stochastic optimization algorithm: minibatching and interpolation learning. *Advances in Neural Information Processing Systems* **34,** 7333–7345 (2021) (Cited on pages 99, 100, 102, 105, 106, 111).

139. Kotsalis, G., Lan, G. & Li, T. Simple and optimal methods for stochastic variational inequalities, I: Operator extrapolation. *SIAM Journal on Optimization* **32,** 2041–2073 (2022) (Cited on page 99).

140. Kotsalis, G., Lan, G. & Li, T. Simple and optimal methods for stochastic variational inequalities, ii: Markovian noise and policy evaluation in reinforcement learning. *SIAM Journal on Optimization* **32,** 1120–1155 (2022) (Cited on page 99).

141. Li, T., Lan, G. & Pananjady, A. Accelerated and instance-optimal policy evaluation with linear function approximation. *arXiv preprint arXiv:2112.13109* (2021) (Cited on page 99).

142. Li, T., Wu, F. & Lan, G. Stochastic first-order methods for average-reward markov decision processes. *arXiv preprint arXiv:2205.05800* (2022) (Cited on page 99).

143. Cotter, A., Shamir, O., Srebro, N. & Sridharan, K. Better mini-batch algorithms via accelerated gradient methods. *Advances in neural information processing systems* **24** (2011) (Cited on page 99).

144. Liu, C. & Belkin, M. Mass: an accelerated stochastic method for over-parametrized learning. *arXiv preprint arXiv:1810.13395* (2018) (Cited on page 99).

145. Lan, G. & Zhou, Y. Random gradient extrapolation for distributed and stochastic optimization. *SIAM Journal on Optimization* **28,** 2753–2782 (2018) (Cited on pages 100, 107, 108).

146. Lan, G. & Zhou, Y. An optimal randomized incremental gradient method. *Mathematical programming.* https://arxiv.org/abs/1507.02000 (2015) (Cited on page 108).

147. Candes, E. J. & Plan, Y. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory* **57,** 2342–2359 (2011) (Cited on page 111).

148. Candes, E. & Recht, B. Exact matrix completion via convex optimization. *Communications of the ACM* **55,** 111–119 (2012) (Cited on page 111).

149. Fazel, M., Candes, E., Recht, B. & Parrilo, P. *Compressed sensing and robust recovery of low rank matrices* in *2008 42nd Asilomar Conference on Signals, Systems and Computers* (2008), 1043–1047 (Cited on page 111).

150. Tsybakov, A., Koltchinskii, V. & Lounici, K. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Annals of Statistics* **39,** 2302–2329 (2011) (Cited on page 111).

151. Kotz, S. & Nadarajah, S. *Multivariate t-distributions and their applications* (Cambridge University Press, 2004) (Cited on page 115).

152. Nemirovski, A. & Yudin, D. *Problem complexity and method efficiency in optimization* (John Wiley, XV, 1983) (Cited on page 134).

153. Azuma, K. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series* **19,** 357–367 (1967) (Cited on page 134).

154. Freedman, D. A. On tail probabilities for martingales. *The Annals of Probability,* 100–118 (1975) (Cited on page 134).

155. Lan, G., Nemirovski, A. & Shapiro, A. Validation analysis of mirror descent stochastic approximation method. *Mathematical programming* **134,** 425–458 (2012) (Cited on page 135).

156. Kerdreux, T., d'Aspremont, A. & Pokutta, S. Local and global uniform convexity conditions. *arXiv preprint arXiv:2102.05134* (2021) (Cited on page 180).

157. Grant, M. & Boyd, S. *CVX: Matlab Software for Disciplined Convex Programming, version 2.1* https://cvxr.com/cvx. Mar. 2014 (Cited on page 183).

158. Karmarkar, N. *A new polynomial-time algorithm for linear programming* in *Proceedings of the sixteenth annual ACM symposium on Theory of computing* (1984), 302–311 (Cited on page 183).