No Author Given

Algorithmic Stochastic Convex Optimization

– Monograph –

August 31, 2022

Springer Nature

This book is dedicated to «three oracles»:
Boris Teodorovich Polyak,
Arkadi Semenovich Nemirovski,
Yurii Evgenievich Nesterov
– who invited us to modern numerical optimization methods.

Foreword

Use the template *foreword.tex* together with the document class SVMono (monograph-type books) or SVMult (edited books) to style your foreword.

Ask to write Foreword by A.S. Nemirovksi or B.T. Polyak.

The foreword covers introductory remarks preceding the text of a book that are written by a *person other than the author or editor* of the book. If applicable, the foreword precedes the preface which is written by the author or editor of the book.

Place, month year *Firstname Surname*

Preface

Use the template *preface.tex* together with the document class SVMono (monograph-type books) or SVMult (edited books) to style your preface.

A preface is a book's preliminary statement, usually written by the *author or editor* of a work, which states its origin, scope, purpose, plan, and intended audience, and which sometimes includes afterthoughts and acknowledgments of assistance.

When written by a person other than the author, it is called a foreword. The preface or foreword is distinct from the introduction, which deals with the subject of the work.

Customarily *acknowledgments* are included as last part of the preface.

Place(s), *Firstname Surname*
month year *Firstname Surname*

Acknowledgements

We would like to thank Boris Mordukhovich and Elizabeth Loew. Without their help, this book would hardly have been published by Springer.

This book has been in preparation since 2012. During all this time, the authors had a great opportunity to consult with S. Chukanov, Yu. Golubev, G. Gidel, O. Granichin, R. Hildebrand, A. Juditsky, A. Nedich, A. Nemirovski, Yu. Nesterov, B. Polyak, V. Protasov, P. Richtarik, G. Scutari, A. Shananin, A. Shapiro, V. Spokoiny, S. Stich, M. Takac, S. Tarasov, A. Taylor, A. Tsybakov. We would like to express our sincere gratitude to all of them.

Contents

Acronyms

Use the template *acronym.tex* together with the document class SVMono (monograph-type books) or SVMult (edited books) to style your list(s) of abbreviations or symbols.

Lists of abbreviations, symbols and the like are easily formatted with the help of the Springer-enhanced `description` environment.

ABC Spelled-out abbreviation and definition
BABI Spelled-out abbreviation and definition
CABR Spelled-out abbreviation and definition

Part I
Part Title

Use the template *part.tex* together with the document class SVMono (monograph-type books) or SVMult (edited books) to style your part title page and, if desired, a short introductory text (maximum one page) on its verso page.

Chapter 1
Stochastic optimization and Data Science

Abstract This chapter aims to motivate stochastic optimization problems from a statistical perspective and a statistical learning perspective, where the goal is to maximize the log-likelihood or minimize the population risk. We briefly describe the two main approaches: *offline* (Monte Carlo / Sample Average Approximation) and *online* (Stochastic Approximation) approaches – to solve the expectation minimization problems of the type

$$\min_{x \in Q} \left\{ f(x) := \mathbb{E}_{\xi \sim \mathcal{D}}[f(x, \xi)] \right\}. \tag{1.1}$$

1.1 Motivation to stochastic optimization

According to [114], «Optimization problems involving stochastic models occur in almost all areas of science and engineering, so diverse as telecommunication, medicine, or finance, to name just a few. This stimulates interest in rigorous ways of formulating, analyzing, and solving such problems. Due to the presence of random parameters in the model, the theory combines concepts of the optimization theory, the theory of probability and statistics, and functional analysis. Moreover, in recent years the theory and methods of stochastic programming have undergone major advances.» This «major advances» are strongly stimulated by the explosion of interest in *Data science* problems. In the last decade, several good books have appeared on the relationship between *Stochastic Optimization* and *Data Science* [114, 110, 8]. In this section, we briefly describe the two main origins of stochastic optimization problems in Data Science: 1) *Statistical* origin (*maximum likelihood estimation*) and 2) *Machine Learning* origin (*stochastic gradient descent* and *regularized expected risk minimization*).

1.1.1 Statistical motivation

We start with the simplest example. Let $x^* \in \mathbb{R}$ be an unknown scalar parameter, $\eta \sim \mathcal{N}\left(0, \sigma^2\right)$ be Gaussian noise. Assume that we can measure

$$\xi^k = x^* + \eta^k, \ \ k = 1, ..., N,$$

where η^k are i.i.d. (independent identically distributed as η). **The goal is to estimate x^* from $\left\{\xi^k\right\}_{k=1}^N$.**

The main observation is the following: x^* is a solution of Stochastic optimization problem

$$\min_{x \in \mathbb{R}} \mathbb{E}_\xi \left[f(x, \xi) := (\xi - x)^2 \right], \tag{1.2}$$

where $\xi \sim \mathcal{N}\left(x^*, \sigma^2\right)$. Indeed,

$$\mathbb{E}_\xi (\xi - x)^2 = \mathbb{E}_\xi \xi^2 - 2x\mathbb{E}_\xi \xi + x^2 = (x^*)^2 + \sigma^2 - 2xx^* + x^2 = (x^* - x)^2 + \sigma^2$$

attains minimum in $x = x^*$. However, x^* is unknown (and probably σ^2). How problem (1.2) can be solved? Since $\left\{\xi^k\right\}_{k=1}^N$ are available, the Monte Carlo approach can be employed. This approach consists in replacing problem (1.2) by its empirical version

$$\min_{x \in \mathbb{R}} \left[\frac{1}{N} \sum_{k=1}^N (\xi^k - x)^2 \right]. \tag{1.3}$$

The solution to the problem (1.3) can be easily provided

$$\bar{x}^N = \frac{1}{N} \sum_{k=1}^N \xi^k. \tag{1.4}$$

In Statistics, this average is known as the *Sample Mean*, which is the best known (unbiased and with the smallest variance, see Theorem 1.1 below) estimate for the unknown parameter in the described parametric model, see Theorem 1.1 hereinafter.

The solution (1.4) can be also obtained by the following online procedure

$$x^{k+1} = x^k - \frac{1}{2(k+1)} \nabla_x f(x^k, \xi^k) = x^k - \frac{1}{k+1}(x^k - \xi^k), \ \ k = 0, ..., N-1, \tag{1.5}$$

where $x^0 = \xi^1$. This procedure corresponds to the *Stochastic Gradient Descent* (SGD) for 2-strongly convex in the 2-norm stochastic optimization problem (1.2).

A natural question arising here: by what scheme was $f(x, \xi)$ selected in (1.2)? Probably there are many ways to choose $f(x, \xi)$. If so, what is the «best» way to do it? Further, we briefly describe the basics of the maximum likelihood theory, which allows us to answer these questions.

Assume that some random variable ξ depends on an unknown vector of parameters $x^* \in \mathbb{R}^n$. Let $p(x, \xi)$ be the probability (probability density function) that we observe ξ if the true vector of parameters is $x \in \mathbb{R}^n$. In the aforementioned example, $n = 1$

and probability density function was

$$p(x,\xi) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\xi - x)^2}{2\sigma^2}\right).$$

If i.i.d. samples $\left\{\xi^k\right\}_{k=1}^N$ are available let us introduce the likelihood

$$p\left(x, \left\{\xi^k\right\}_{k=1}^N\right) = \prod_{k=1}^N p(x,\xi^k).$$

Perhaps, one of the most productive ideas in Statistics is to estimate the true vector of parameters x^* as a vector that maximizes likelihood $p\left(x, \left\{\xi^k\right\}_{k=1}^N\right)$. This problem can be equivalently reformulated as minimization of (normalized) minus log-likelihood

$$\min_{x\in\mathbb{R}^n}\left[-\frac{1}{N}\log p\left(x, \left\{\xi^k\right\}_{k=1}^N\right) = -\frac{1}{N}\sum_{k=1}^N \log p(x,\xi^k)\right].$$

This minimization problem can be considered as the empirical (sometimes called *Monte Carlo*) version of the Stochastic optimization problem

$$\min_{x\in\mathbb{R}^n} \mathbb{E}_\xi\left[-\log p(x,\xi)\right]. \tag{1.6}$$

In particular, for the aforementioned Gaussian model, this problem looks like

$$\min_{x\in\mathbb{R}} \mathbb{E}_\xi\left[\frac{1}{2\sigma^2}(\xi - x)^2 + \frac{1}{2}\log\left(2\pi\sigma^2\right)\right],$$

which is equivalent to (1.2).

Moreover, the observation that the true value of unknown vector of parameters x^* is a solution of (1.6) holds in the general case, i.e.,

$$x^* \in \operatorname{Arg}\min_{x\in\mathbb{R}^n} \mathbb{E}_\xi\left[-\log p(x,\xi)\right].$$

Indeed,[1]

$$\mathbb{E}_\xi\left[-\log p(x,\xi)\right] = -\int p(x^*,\xi)\log p(x,\xi)d\xi \geq -\int p(x^*,\xi)\log p(x^*,\xi)d\xi$$

since (*Jensen's inequality* for the entropy)

$$KL\left(p(x^*,\cdot), p(x,\cdot)\right) = \int p(x^*,\xi)\log\left(\frac{p(x^*,\xi)}{p(x,\xi)}\right)d\xi \geq 0$$

and $KL\left(p(x^*,\cdot), p(x,\cdot)\right) = 0$, when $x = x^*$.

[1] For certainty, here $p(x,\xi)$ is assumed to be a probability density function.

Thus, we explained that in the general case $f(x,\xi) := -\log p(x,\xi)$ in (1.2) and the maximum likelihood approach is nothing more than the Monte Carlo approach for Stochastic optimization problem (1.6).

Definitely the main gem of Statistics is the Fisher's theorem about asymptotic properties of *maximum likelihood estimation* (MLE)

$$\hat{x}^N_{MLE} = \arg\max_{x\in\mathbb{R}^n} p\left(x, \left\{\xi^k\right\}_{k=1}^N\right) = \arg\min_{x\in\mathbb{R}^n}\left[-\log p\left(x, \left\{\xi^k\right\}_{k=1}^N\right)\right]. \tag{1.7}$$

The next theorem presents an informal variant of this theorem.

Theorem 1.1 *Assume that $p(x,\xi)$ is sufficiently smooth and the set*

$$\{\xi : p(x,\xi) > 0\}$$

does not depend on x.[2] *Then*

1. *for all unbiased statistics $\tilde{x}^N\left(\left\{\xi^k\right\}_{k=1}^N\right)$ with finite second moment, the Rao–Cramer inequality holds*[3]

$$\mathbb{E}_{\left\{\xi^k\right\}_{k=1}^N}\left[\left(\tilde{x}^N\left(\left\{\xi^k\right\}_{k=1}^N\right) - x^*\right)\left(\tilde{x}^N\left(\left\{\xi^k\right\}_{k=1}^N\right) - x^*\right)^T\right] \succcurlyeq [NI_{x^*}]^{-1},$$

   where

$$I_{x^*} = \mathbb{E}_\xi\left[\nabla_x \log p(x^*,\xi)\left(\nabla_x \log p(x^*,\xi)\right)^T\right]$$

   is the Fisher information matrix.[4]
2. *MLE $\hat{x}^N_{MLE}\left(\left\{\xi^k\right\}_{k=1}^N\right)$ (see (1.7)) has asymptotically*[5] *normal (Gaussian) distribution $\mathcal{N}\left(x^*, \left[NI_{x^*,N}\right]^{-1}\right)$ and the Rao–Cramer inequality turns into the equality. This means that MLE has the asymptotically smallest variance along all the directions and no matter what x^* is.*

As a consequence of this theorem, the asymptotically smallest confidence set around MLE can be constructed. The online approach (based on the SGD, proper stepsize policy and the *Polyak–Juditsky–Ruppert averaging*) leads to a similar asymptotic result.

Unfortunately, the asymptotic theory does not fully characterize the real state of affairs when N is not sufficiently large. Indeed, let us consider the Bernoullie parametric model (coin flipping) with likelihood $p(x,\xi) = x^\xi(1-x)^{1-\xi}$ and $x^* > 0$ small enough. Then while $N \lesssim 1/x^*$ with positive probability, for MLE $\hat{x}^N = 0$ [110]. Hence $\mathbb{E}_\xi[-\log p(0,\xi)] = \infty$ is not well defined.

---

[2] This is satisfied for Gaussian noise model $\xi = x + \eta$, but is not satisfied if the noise η is uniformly distributed on $[0, x]$. We emphasis, that this assumption is informal.

[3] $A \succcurlyeq B$ means that for all $z \in \mathbb{R}^n$ $\langle z, (A-B)z\rangle \geq 0$.

[4] Note that $KL\left(p(x^*, \cdot), p(x^*+y, \cdot)\right) \simeq \frac{1}{2}\langle y, I_{x^*} y\rangle$.

[5] Under assumption that $N \to \infty$.

Modern offline asymptotic theory of statistics [54] (le Cam's theory) was further developed in a partially non-asymptotic and misspecification[6] directions, see e.g. [117]. In this book, we mainly (except the next section) concentrate on non-asymptotic online approaches for (1.6) and more general problem formulations.

At the end of this section, we aim to demonstrate the role of regularization in the offline approach as a Bayesian prior. Assume that in the general scheme, which described by the parametric model $p(x,\xi)$, we have an additional information about vector of parameters x: x is a random vector that was priory independently generated from the distribution with density function $\pi(x)$.

A Bayesian estimator is an estimator that minimizes the posterior expected value of the loss function (we consider quadratic loss), which coincides with a posterior mean

$$\hat{x}_B^N = \arg\min_{x\in\mathbb{R}^n} \int\limits_{\mathbb{R}^n} \|x-z\|^2 p\left(z, \left\{\xi^k\right\}_{k=1}^N\right)\pi(z)dz = \int\limits_{\mathbb{R}^n} x \frac{p\left(x, \left\{\xi^k\right\}_{k=1}^N\right)\pi(x)}{\int\limits_{\mathbb{R}^n} p\left(y, \left\{\xi^k\right\}_{k=1}^N\right)\pi(y)dy}dx. \tag{1.8}$$

The next theorem presents an informal analogue of Theorem 1.1 in this case.

Theorem 1.2 *Assume that $p(x,\xi)$ and $\pi(x)$ are sufficiently smooth and the set*

$$\{\xi : p(x,\xi) > 0\}$$

does not depend on x. Then

1. *for all statistics $\tilde{x}^N\left(\left\{\xi^k\right\}_{k=1}^N\right)$ with finite second moment, the van Trees inequality holds*

$$\mathbb{E}_{\left(x,\left\{\xi^k\right\}_{k=1}^N\right)}\left[\left(\tilde{x}^N\left(\left\{\xi^k\right\}_{k=1}^N\right)-x\right)\left(\tilde{x}^N\left(\left\{\xi^k\right\}_{k=1}^N\right)-x\right)^T\right] \geqslant \left[NI_p + I_\pi\right]^{-1},$$

where

$$I_p = \mathbb{E}_{(x,\xi)}\left[\nabla_x \log p(x,\xi)\left(\nabla_x \log p(x,\xi)\right)^T\right]$$

is the Fisher information matrix and

$$I_\pi = \mathbb{E}_x\left[\nabla \log \pi(x)\left(\nabla \log \pi(x)\right)^T\right].$$

2. *Bayesian estimator $\hat{x}_B^N\left(\left\{\xi^k\right\}_{k=1}^N\right)$ (see (1.8)) has conditional (with a priori drawing $x = x^*$) asymptotically normal distribution $\mathcal{N}\left(x^*, \left[NI_{x^*}\right]^{-1}\right)$, where I_{x^*} was introduced in Theorem 1.1.*

[6] If the parametric model is wrong, MLE can be interpreted as the asymptotically best way to estimate the KL-projection of the true vector of parameters on the parametric model.

A close result is contained in the *Bernstein–von Mises theorem*: a posterior distribution has asymptotically normal distribution centered at MLE with covariance matrix NI_{x^*}.

In Bayesian statistics, a *maximum a posterior estimation* (MAP)

$$\hat{x}^N_{MAP} = \arg\max_{x\in\mathbb{R}^n} p\left(x, \left\{\xi^k\right\}_{k=1}^N\right)\pi(x) = \arg\min_{x\in\mathbb{R}^n}\left[-\log p\left(x, \left\{\xi^k\right\}_{k=1}^N\right) - \log\pi(x)\right].$$

plays also an important role. MAP has typically the same asymptotic behavior as Bayesian estimator.

Let us consider several examples. The first example is *Regularized Least Squares*.

Ridge Regression and LASSO

Let $x^* \in \mathbb{R}^n$ be an unknown vector of parameters and $\eta \sim \mathcal{N}\left(0, \sigma^2\right)$ be Gaussian noise. Assume that we can measure

$$\xi^k = \langle a_k, x^*\rangle + \eta^k,\ k = 1, ..., N,$$

where η^k are i.i.d. (independent identically distributed as η) and matrix $A = [a_1, ..., a_N]^T$ is known.[7] **The goal is to estimate** x^* **from** $\xi := \left\{\xi^k\right\}_{k=1}^N$**.** Simple calculations lead to the following formulas[8]

$$\hat{x}^N_{MLE} = \arg\min_{x\in\mathbb{R}^n}\left[\frac{1}{2\sigma^2}\|Ax - \xi\|_2^2\right],$$

$$\hat{x}^N_B = \hat{x}^N_{MAP} = \arg\min_{x\in\mathbb{R}^n}\left[\frac{1}{2\sigma^2}\|Ax - \xi\|_2^2 + \frac{1}{2\sigma_\pi^2}\|x - \bar{x}\|_2^2\right],$$

where a priory x_i, $i = 1, ..., n$ are assumed to be independent and identically distributed according to $\mathcal{N}\left(\bar{x}, \sigma_\pi^2\right)$ (Ridge Regression) and

$$\hat{x}^N_{MAP} = \arg\min_{x\in\mathbb{R}^n}\left[\frac{1}{2\sigma^2}\|Ax - \xi\|_2^2 + \lambda\|x\|_1\right],$$

where the prior probability density is (LASSO):

$$\pi(x) = \prod_{i=1}^n \frac{\lambda}{2}\exp(-\lambda|x_i|) = \left(\frac{\lambda}{2}\right)^n \exp\left(-\lambda\|x\|_1\right).$$

It is obvious that Bayesian estimator and MAP asymptotically ($N \to \infty$) coincide with MLE. Another important observation that Bayesian estimator and MAP asymptotically coincide with MLE when $\sigma_\pi^2 \to \infty$. Both of these observations take place in the general case. So *Bayesian prior* can be interpreted as a regularizer in Bayesian version of maximum likelihood optimization problem.

---

[7] Note that a_k can also be generated randomly. In this case, to preserve the results it is sufficient to require that $\{a_k\}_{k=1}^n$ and $\left\{\eta^k\right\}_{k=1}^n$ are independent.

[8] We assume that parameters σ^2, σ_π^2, λ are known.

The second example goes back to Vadim V. Mottl.

Soft-SVM

In this example, *Soft-Support-Vector Machine* (Soft-SVM) is derived based on Bayesian inference with

$$p\left(x, \xi^k := \left(y^k, a_k\right)\right) \propto \begin{cases} 1, & \text{if } y^k \langle x, a_k \rangle \geq 1 \\ \exp\left(-\left(1 - y^k \langle x, a_k \rangle\right)\right), & \text{else,} \end{cases}$$

where $y^k \in \{-1, 1\}$ and a priory x_i, $i = 1, ..., n$ are assumed to be independent and identically distributed according to $\mathcal{N}\left(0, \sigma_\pi^2\right)$. Improper probability density function $p\left(x, \xi^k\right)$ has a natural interpretation: there exists «true» hyperplane (determined by the vector x^*) such that the data points with $y^k = 1$ lie mostly from the one side of this hyperplane and the data points with $y^k = -1$ lie mostly from the other side. The goal is to recognize this hyperplane from the data points having a prior information about x^*. Simple calculations lead to the following formula

$$\hat{x}_{MAP}^N = \arg\min_{x \in \mathbb{R}^n} \left[\sum_{k=1}^N \max\left\{0, 1 - y^k \langle x, a_k \rangle\right\} + \frac{1}{2\sigma_\pi^2} \|x\|_2^2 \right].$$

1.1.2 Machine Learning motivation

In the statistical approach, the loss function is $f(x, \xi) := -\log p(x, \xi)$. It means that we require parametric model $p(x, \xi)$. In many practical situations, $p(x, \xi)$ is not available. However in *Regression problems* we can introduce the *least square loss function* $f\left(x, \xi := (y, a)\right) = (y - \langle a, x \rangle)^2$. Without any knowledge of probability nature of ξ, we can consider the expected loss minimization problem (stochastic optimization problem)

$$\min_{x \in \mathbb{R}^n} \mathbb{E}_{(y,a)} \left[(y - \langle a, x \rangle)^2\right].$$

In the offline approach, this problem has a form:

$$\min_{x \in \mathbb{R}^n} \frac{1}{N} \|Y - Ax\|_2^2,$$

where $Y = \left(y^1, ..., y^N\right)^T$, $A = [a_1, ..., a_N]^T$. Similarly, in *Classification problems* we can introduce the *hinge-loss function* $f\left(x, \xi := (y, a)\right) = \max\{0, 1 - y\langle x, a \rangle\}$ and corresponding stochastic optimization problems has the following form

$$\min_{x\in\mathbb{R}^n} \mathbb{E}_{(y,a)} \left[\max\left\{0, 1 - y\langle x, a\rangle\right\}\right].$$

In many real world applications, we have some prior information about how much could be (should be) x^*. Typically, this information is formalized as a constraint of the type $x \in Q$, where Q is often chosen as a ball $B_p^n(R_p)$ in p-norm ($p \geq 1$) centered at 0 with radius R_p or another convex compact set with simple structure, e.g. unit simplex $S_n(1)$. So the final stochastic optimization problem in general has a form[9]

$$\min_{x\in Q\subseteq\mathbb{R}^n} \left[f(x) := \mathbb{E}_\xi f(x,\xi).\right] \tag{1.9}$$

For $Q = B_2^n(R_2)$ (or $Q = B_1^n(R_1)$) if the constraint is reached it can be replaced by $\|x\|_2^2$-regularization (or $\|x\|_1$-regularization) with Lagrange multiplayer as a regularization parameter.

All aforementioned problems (Regression and Classification) have two things in common. The target functions

1. are convex: for all ξ and $x, z \in Q$

$$f(z,\xi) \geq f(x,\xi) + \langle \nabla_x f(x,\xi), z - x\rangle$$

and *M-Lipschitz continuous* in x in 2-norm: for all ξ and $x, z \in Q$

$$|f(z,\xi) - f(x,\xi)| \leq M\|z - x\|_2.$$

2. have *generalized linear* structure:

$$f(x,\xi) := g\left(y(\xi), \langle x, a(\xi)\rangle\right).$$

The first common thing guarantees the effectiveness of the *online* approach. Both of them guarantee the effectiveness of the *offline* approach.

Let us start with the *offline* approach. We introduce the *empirical loss*

$$\bar{f}(x) := \bar{f}\left(x, \left\{\xi^k\right\}_{k=1}^N\right) = \frac{1}{N}\sum_{k=1}^N f(x,\xi^k)$$

and its minimizer

$$\hat{x}^N \in \operatorname{Arg}\min_{x\in Q} \bar{f}\left(x, \left\{\xi^k\right\}_{k=1}^N\right).$$

Theorem 1.3 (Learnability for generalized linear models) *Consider the stochastic optimization problem* (1.9) *with* $f(x,\xi)$ *satisfying the aforementioned conditions 1 and 2 and convex* $Q \subseteq B_2^n(R)$. *Then with probability at least* $1 - \beta$

[9] Here and everywhere below we will denote the solution of this problem as x^*. If the solution is not unique x^* means one of the solutions, e.g. such that is the closest to the starting point (initial guess).

$$\sup_{x \in Q} \left| \bar{f}(x) - f(x) \right| = O\left(MR \sqrt{\frac{\log(1/\beta)}{N}} \right).$$

Hence, with probability at least $1 - \beta$ *the following holds*

$$f(x) - f(x^*) \leq \bar{f}(x) - \bar{f}(\hat{x}^N) + O\left(MR \sqrt{\frac{\log(1/\beta)}{N}} \right). \tag{1.10}$$

If additionally for all ξ *and* $x, z \in Q$,

$$f(z, \xi) \geq f(x, \xi) + \langle \nabla_x f(x, \xi), z - x \rangle + \frac{\mu}{2} \|z - x\|_2^2,$$

i.e., $f(x, \xi)$ *is* μ*-strongly convex in* x *in* 2*-norm, then with probability at least* $1 - \beta$

$$f(x) - f(x^*) \leq 2 \left(\bar{f}(x) - \bar{f}(\hat{x}^N) \right) + O\left(\frac{M^2 \log(1/\beta)}{\mu N} \right). \tag{1.11}$$

If the condition 2 is no longer met, then (1.11) *should be rewritten as follows: with probability at least* $1 - \beta$

$$f(x) - f(x^*) \leq \sqrt{\frac{2M^2}{\mu} \left(\bar{f}(x) - \bar{f}(\hat{x}^N) \right)} + \tilde{O}\left(\frac{M^2 \log(1/\beta)}{\mu N} \right). \tag{1.12}$$

Moreover, all these inequalities are optimal up to a constant factor.

This theorem reduces stochastic optimization problem to the empirical loss (risk) minimization problem

$$\min_{x \in Q} \frac{1}{N} \sum_{k=1}^{N} f(x, \xi^k) \tag{1.13}$$

with proper choice of N, see the next section.

Now we move to the *online* approach and explain why it is so called. The standard SGD is in the core of the online approach:

$$x^{k+1} = \pi_Q \left(x^k - \gamma_k \nabla_x f(x^k, \xi^k) \right), \tag{1.14}$$

where π_Q is the Euclidean projection onto Q. Note that

$$\begin{aligned} \|x^{k+1} - x^*\|_2^2 &= \left\| \pi_Q \left(x^k - \gamma_k \nabla_x f(x^k, \xi^k) - x^* \right) \right\|_2^2 \\ &\leq \|x^k - \gamma_k \nabla_x f(x^k, \xi^k) - x^*\|_2^2 \\ &= \|x^{k+1} - x^*\|_2^2 - 2\gamma_k \langle \nabla_x f(x^k, \xi^k), x^k - x^* \rangle + \gamma_k^2 \|\nabla_x f(x^k, \xi^k)\|_2^2 \\ &\leq \|x^{k+1} - x^*\|_2^2 - 2\gamma_k \langle \nabla_x f(x^k, \xi^k), x^k - x^* \rangle + \gamma_k^2 M^2. \end{aligned}$$

The last inequality holds as $f(x,\xi)$ is M-Lipschitz continuous in x in 2-norm and therefore, $\|\nabla_x f(x^k,\xi^k)\|_2 \le M$. From the convexity of $f(x,\xi)$ in x:

$$f(x^k,\xi^k) - f(x^*,\xi^k) \le \langle \nabla_x f(x^k,\xi^k), x^k - x^* \rangle \le \frac{1}{2\gamma_k}\left(\|x^k - x^*\|_2^2 - \|x^{k+1} - x^*\|_2^2\right) + \frac{\gamma_k M^2}{2}.$$

From the μ-strong convexity of $f(x,\xi)$ in x:

$$f(x^k,\xi^k) - f(x^*,\xi^k) \le \langle \nabla_x f(x^k,\xi^k), x^k - x^* \rangle - \frac{\mu}{2}\|x^k - x^*\|_2^2 \le \frac{1}{2}\left(\frac{1}{\gamma_k} - \mu\right)\|x^k - x^*\|_2^2 - \frac{1}{2\gamma_k}\|x^{k+1} - x^*\|_2^2 + \frac{\gamma_k M^2}{2}.$$

Summing for $k = 1, \ldots, N$ «convex» inequality with $\gamma_k \equiv \frac{R}{M\sqrt{N}}$ and «strongly convex» inequality with[10] $\gamma_k = \frac{1}{\mu k}$ we obtain after normalization (multiplication on N^{-1}):

$$\frac{1}{N}\sum_{k=1}^{N} f(x^k,\xi^k) \le \frac{1}{N}\sum_{k=1}^{N} f(x^*,\xi^k) + \frac{M\|x^1 - x^*\|_2}{\sqrt{N}}, \tag{1.15}$$

$$\frac{1}{N}\sum_{k=1}^{N} f(x^k,\xi^k) \le \frac{1}{N}\sum_{k=1}^{N} f(x^*,\xi^k) + \frac{M^2(1+\log N)}{2\mu N}. \tag{1.16}$$

Note that in (1.15), (1.16) $x^* \in Q$ can be chosen in an arbitrary manner, say such that to minimize RHS, i.e.,

$$\frac{1}{N}\sum_{k=1}^{N} f(x^k,\xi^k) \le \min_{x\in Q}\frac{1}{N}\sum_{k=1}^{N} f(x,\xi^k) + \frac{M\|x^1 - x^*\|_2}{\sqrt{N}},$$

$$\frac{1}{N}\sum_{k=1}^{N} f(x^k,\xi^k) \le \min_{x\in Q}\frac{1}{N}\sum_{k=1}^{N} f(x,\xi^k) + \frac{M^2(1+\log N)}{2\mu N}.$$

Since we still do not use the probability nature of ξ^k, the last two inequalities characterize the SGD (1.14) as online learning procedure in the standard online sense [20].

If we remember now about i.i.d. nature of $\{\xi^k\}_{k=1}^N$, remember that: $\mathbb{E}_\xi f(x,\xi) \equiv f(x)$, $f(x,\xi)$ is M-Lipschitz continuous in x in 2-norm and $f(x)$ is convex, than (1.15), (1.16) could be further simplify (*online to batch conversion*).

Theorem 1.4 *Consider stochastic optimization problems* (1.9) *with* $f(x,\xi)$ *satisfies the condition 1. Then for* x^k *generated by* (1.14) *with probability at least* $1-\beta$*:*

$$f(\bar{x}^N) - f(x^*) = O\left(\frac{M\|x^1 - x^*\|_2 \log(1/\beta)}{\sqrt{N}}\right), \tag{1.17}$$

[10] In this case we have the telescopic property: $\frac{1}{\gamma_{k+1}} - \mu = \frac{1}{\gamma_k}$.

where

$$\bar{x}^N = \frac{1}{N}\sum_{k=1}^{N} x^k. \tag{1.18}$$

If additionally $f(x,\xi)$ is μ-strongly convex in x in 2-norm, then with probability at least $1-\beta$:

$$f(\bar{x}^N) - f(x^*) = O\left(\frac{M^2 \log(N/\beta)}{\mu N}\right). \tag{1.19}$$

Since $\|x^1 - x^*\|_2 \leq 2R$, it follows that (1.17) and (1.19) correspond to (1.10) and (1.11), (1.12) in *sample complexity* – the required number of samples N. However, online approach does not require to solve an auxiliary empirical problem (1.13) and was justified under weaker assumptions. More detailed comparison online and offline approaches is given in the next section.

To conclude this section, remind the main observation: statistical approach for data science problems is a particular case of the general machine learning (ML) approach, where the loss function has a specific form determined by log-likelihood functions. So further we will consider mainly the ML approach, which characterizes stochastic optimization problem (1.9).

1.2 Sample Average Approximation vs Stochastic Approximation

In this section, we consider stochastic optimization problem (1.9)

$$\min_{x\in Q\subseteq\mathbb{R}^n} \left[f(x) := \mathbb{E}_\xi\left[f(x,\xi)\right].\right] \tag{1.20}$$

We are mainly interested in the sample complexity of offline (also called *Sample Average Approximation*) and online (also called *Stochastic Approximation*) procedures, which generate $\tilde{x}^N\left(\{\xi^k\}_{k=1}^N\right)$ from the solution of the empirical problem (1.13) or from the procedure of type (1.14). More precisely, we are interested in estimating such $N := N(\varepsilon,\beta)$ that

$$\mathbb{P}\left(f\left(\tilde{x}\left(\{\xi^k\}_{k=1}^N\right)\right) - f(x^*) \leq \varepsilon\right) \geq 1-\beta.$$

Assume that $Q \subseteq B_p^n(R_p)$ $(p \geq 1)$ and for all ξ and $x, y \in Q$

$$|f(y,\xi) - f(x,\xi)| \leq M_p\|y-x\|_p. \tag{1.21}$$

Let $\bar{x}^N_{\delta,\tilde{\beta}} := \bar{x}^N_{\delta,\tilde{\beta}}\left(\{\xi^k\}_{k=1}^N\right)$ be the $(\delta,\tilde{\beta})$-solution of the empirical problem (1.13)

$$\min_{x\in Q}\left[\bar{f}(x) := \frac{1}{N}\sum_{k=1}^{N} f(x,\xi^k)\right],$$

that is, with probability at least $1-\tilde{\beta}$:

$$\bar{f}\left(\bar{x}^N_{\delta,\tilde{\beta}}\right) - \min_{x\in Q}\bar{f}(x) = \bar{f}\left(\bar{x}^N_{\delta,\tilde{\beta}}\right) - \bar{f}\left(\hat{x}^N\right) \le \delta.$$

1.2.1 Non-convex case and convex case

One of the first and quite unexpected results about the offline approach is the following

Theorem 1.5 *Assume that* (1.21) *is satisfied. Then for* $\bar{x}^N_{\varepsilon/2,\beta/2}\left(\{\xi^k\}_{k=1}^N\right)$,

$$N = O\left(\frac{M_p^2 R_p^2}{\varepsilon^2}\left(n\log\left(\frac{M_p R_p}{\varepsilon}\right) + \log\left(\frac{1}{\beta}\right)\right)\right). \tag{1.22}$$

This bound is optimal up to a logarithmic factor. Moreover, if we additionally assume that $f(x,\xi)$ *is convex and smooth in* x, (1.22) *is still an optimal bound.*

Proof Nazary, please add the proof of the Theorem based on [114] (https://cpn-us-w2.wpmucdn.com/sites.gatech.edu/dist/4/1470/files/2021/03/SPbook.pdf) Sections 5.3.1 and 5.3.2.

The proof consists of two parts. Firstly, we prove the result for finite set Q. And then we generalize the result to the case of bounded Q.

For $\varepsilon \ge 0$ denote by

$$S^\varepsilon := \left\{x \in Q : f(x) \le \min_{x\in Q} f(x) + \varepsilon\right\}, \quad \bar{S}^\varepsilon := \left\{x \in Q : \bar{f}(x) \le \min_{x\in Q}\bar{f}(x) + \varepsilon\right\}$$

the sets of ε-optimal solutions of the problem (1.20) and the empirical problem (1.13), respectively.

In the case of finite Q, the sets S^ε and $\bar{S}^\varepsilon$ are nonempty and finite. For parameters $\varepsilon \ge 0$ and $\delta \in [0,\varepsilon]$, consider the event $\{\bar{S}^\delta \subset S^\varepsilon\}$. This event means that any δ-optimal solution of the empirical problem (1.13) is an ε-optimal solution of the problem (1.20). Next, we estimate the probability of that event.

$$\{\bar{S}^\delta \not\subset S^\varepsilon\} = \bigcup_{x\in Q\setminus S^\varepsilon}\bigcap_{y\in Q}\left\{\bar{f}(x) \le \bar{f}(y) + \delta\right\}$$

$$\mathbb{P}\left(\bar{S}^\delta \not\subset S^\varepsilon\right) \le \sum_{x\in Q\setminus S^\varepsilon}\mathbb{P}\left(\bigcap_{y\in Q}\left\{\bar{f}(x) \le \bar{f}(y) + \delta\right\}\right)$$

Consider a mapping $u : Q\setminus S^\varepsilon \longrightarrow Q$. If the set $Q\setminus S^\varepsilon$ is empty, then any feasible point $x \in Q$ is an ε-optimal solution of the true problem. Therefore we assume that this set is nonempty. Then from the last inequality follows that

$$\mathbb{P}\left(\bar{S}^{\delta} \not\subset S^{\varepsilon}\right) \leq \sum_{x \in Q \setminus S^{\varepsilon}} \mathbb{P}\left(\bar{f}(x) \leq \bar{f}(u(x)) + \delta\right) \tag{1.23}$$

We assume that the mapping $u(\cdot)$ is chosen in such a way that $f(u(x)) \leq f(x) - \varepsilon^*$, for every $x \in Q \setminus S^{\varepsilon}$ and for some $\varepsilon^* \geq \varepsilon$. Such a mapping always exists. For example, if we use a mapping $u : Q \setminus S^{\varepsilon} \longrightarrow S$ (S – the set of minimizers of $f(x)$ over Q), then (1.23) holds with

$$\varepsilon^* := \min_{x \in Q \setminus S^{\varepsilon}} f(x) - \min_{x \in Q} f(x)$$

and that $\varepsilon^* > \varepsilon$ since the set Q is finite. Different choices of $u()$ give a certain flexibility to the following derivations.

We relax the condition (1.21) as follows: for $Y(x, y, \xi) := f(x, \xi) - f(y, \xi)$ and for all $x, y \in Q$ *sub-Gaussian variance* of a random variable $Y(x, y, \xi) - \mathbb{E}_{\xi} Y(x, y, \xi)$ bounded from above by $\lambda^2 \|y - x\|_p^2$, i.e. for all $t \in \mathbb{R}$:

$$\mathbb{E}_{\xi}\left[\exp\left(t \cdot \left(Y(x, y, \xi) - \mathbb{E}_{\xi} Y(x, y, \xi)\right)\right)\right] \leq \exp\left(t^2 \lambda^2 \|x - y\|_p^2 / 2\right). \tag{1.24}$$

The assumption holds, for example, if the support of ξ is a bounded subset of $\mathbb{R}^d$, or if $Y(x, y, \xi)$ grows at most linearly and ξ has a distribution from an exponential family. Note that if (1.21) holds, then $\lambda^2 \leq 2M_p^2$. [11]

For each $x \in Q \setminus S^{\varepsilon}$ and $x' := u(x)$, define

$$Y(x', x, \xi) := f(x', \xi) - f(x, \xi)$$

Note that $\mathbb{E}_{\xi}[Y(x', x, \xi)] = f(x') - f(x)$, and hence $\mathbb{E}_{\xi}[Y(x', x, \xi)] \leq -\varepsilon^*$ for all $x \in Q \setminus S^{\varepsilon}$ due to the mapping choice.

The corresponding sample average is

$$\bar{Y}(x', x) := \frac{1}{N} \sum_{k=1}^{N} Y(x', x, \xi^k) = \bar{f}(x') - \bar{f}(x).$$

By (1.23) we have

$$\mathbb{P}\left(\bar{S}^{\delta} \not\subset S^{\varepsilon}\right) \leq \sum_{x \in Q \setminus S^{\varepsilon}} \mathbb{P}\left(\bar{Y}(x', x) \geq -\delta\right). \tag{1.26}$$

---

[11] If the assumption (1.21) holds then the expectation function $f(x)$ is also Lipschitz continuous on Q with Lipschitz constant M_p, and hence the random variable $Y(x, y, \xi) - \mathbb{E}_{\xi} Y(x, y, \xi)$ can be bounded as $|Y(x, y, \xi) - \mathbb{E}_{\xi} Y(x, y, \xi)| \leq 2M_p \|x - y\|$ w.p. 1. Moreover, we have that $\mathbb{E}_{\xi}\left(Y(x, y, \xi) - \mathbb{E}_{\xi} Y(x, y, \xi)\right) = 0$, and hence it follows by Hoeffding's inequality that

$$\mathbb{E}_{\xi}\left[\exp\left(t \cdot \left(Y(x, y, \xi) - \mathbb{E}_{\xi} Y(x, y, \xi)\right)\right)\right] \leq \exp\left(t^2 \cdot 2M_p^2 \|x - y\|_p^2\right), \forall t \in \mathbb{R}. \tag{1.25}$$

Let $I_{x,x'}(\cdot)$ denote the (large deviations) rate function of the random variable $Y(x', x, \xi)$. The inequality (1.26) together with the LD upper bound is it enough for proof? implies

$$1 - \mathbb{P}\left(\bar{S}^{\delta} \subset S^{\varepsilon}\right) \leq \sum_{x \in Q \setminus S^{\varepsilon}} \exp\left(-N I_{x,x'}(-\delta)\right)$$

Let ε and δ be non-negative numbers. Then the latter inequality implies

$$1 - \mathbb{P}\left(\bar{S}^{\delta} \subset S^{\varepsilon}\right) \leq |Q| \exp\left(-N\eta(\delta, \varepsilon)\right), \tag{1.27}$$

where $\eta(\delta, \varepsilon) = \min_{x \in Q \setminus S^{\varepsilon}} I_{x,x'}(-\delta)$.

It follows from (1.24) that

$$\ln \mathbb{E}_{\xi} \exp\left\{tY(x, x', \xi)\right\} - t\mathbb{E}_{\xi} Y(x, x', \xi) \leq \frac{\lambda^2 \|x - x'\|_p^2}{2} \leq \frac{\lambda^2 R_p^2}{2}. \tag{1.28}$$

Note, that it is suffices for the proof to verify assumption (1.24) for every: $y = u(x) \in Q \setminus S^{\varepsilon}$.

Hence the rate function $I_{x,x'}(\cdot)$, of $Y(x, x', \xi)$, satisfies

$$I_{x,x'}(z) \geq \sup_{t \in \mathbb{R}} \left(t\left(z - \mathbb{E}_{\xi} Y(x, x', \xi)\right) - \frac{\lambda^2 R_p^2}{2} \right) = \frac{\left(z - \mathbb{E}_{\xi} Y(x, x', \xi)\right)^2}{2\lambda^2 R_p^2}, \forall z \in \mathbb{R}.$$

In particular, it follows that

$$I_{x,x'}(z) \geq \frac{\left(-\delta - \mathbb{E}_{\xi} Y(x, x', \xi)\right)^2}{2\lambda^2 R_p^2} \geq \frac{(\varepsilon^* - \delta)^2}{2\lambda^2 R_p^2} \geq \frac{(\varepsilon - \delta)^2}{2\lambda^2 R_p^2}$$

Consequently the constant $\eta(\delta, \varepsilon)$ satisfies

$$\eta(\delta, \varepsilon) \geq \frac{(\varepsilon - \delta)^2}{2\lambda^2 R_p^2}$$

and hence the bound (1.27) takes the form

$$1 - \mathbb{P}\left(\bar{S}^{\delta} \subset S^{\varepsilon}\right) \leq |Q| \exp\left(\frac{-N(\varepsilon - \delta)^2}{2\lambda^2 R_p^2}\right),$$

This leads to the following result giving an estimate of the sample size which guarantees that any δ-optimal solution of the SAA problem is an ε-optimal solution of the true problem with probability at least $1 - \beta$.

Then for $\varepsilon > 0$, $0 \leq \delta < \varepsilon$, and $\beta \in (0, 1)$, and for the sample size N satisfying

$$N \geq \frac{2\lambda^2 R_p^2}{(\varepsilon - \delta)^2} \ln\left(\frac{|Q|}{\beta}\right) \tag{1.29}$$

it follows that

$$1 - \mathbb{P}\left(\bar{S}^{\delta} \subset S^{\varepsilon}\right) \leq 1 - \beta.$$

Next, we again relax the condition (1.21), and use its non-uniform counterpart: for all ξ and $x, y \in Q$

$$|f(y,\xi) - f(x,\xi)| \leq M_p(\xi)\|y - x\|_p, \tag{1.30}$$

and the moment-generating function $\mathbb{E}_{\xi}\left[\exp\left(t \cdot M_p(\xi)\right)\right]$ of $M_p(\xi)$ is finite valued for all t in a neighborhood of zero. This assumption holds if (1.21) holds and implies that the expectation $\mathbb{E}_{\xi}\left[M_p(\xi)\right]$ is finite and the function $f(x)$ is Lipschitz continuous on Q with Lipschitz constant $\mathbb{E}_{\xi}\left[M_p(\xi)\right]$. By Cramer's large deviation theorem we have that for any $M'_p > \mathbb{E}_{\xi}\left[M_p(\xi)\right]$ there exists a positive constant $\zeta = \zeta(M'_p)$ such that

$$\mathbb{P}\left(\bar{M}_p > M'_p\right) \leq \exp\left(-N\zeta\right), \tag{1.31}$$

where $\bar{M}_p = \frac{1}{N}\sum_{k=1}^{N} M_p(\xi_k)$. Note that it follows from (1.30) that w.p. 1

$$\left|\bar{f}(x) - \bar{f}(x')\right| \leq \bar{M}_p\|x - x'\|_p, \quad \text{for all } x, x' \in Q,$$

i.e., $\bar{f}$ is Lipschitz continuous on Q with Lipschitz constant $\bar{M}_p$.

Let us set $\nu = \frac{(\varepsilon-\delta)^2}{4M'_p}$, $\varepsilon' := \varepsilon - M'_p\nu$ and $\delta' := \delta + M'_p\nu$. Note that $\nu > 0$, $\varepsilon' = \frac{3\varepsilon}{4} + \frac{\delta}{4} > 0$, $\delta' = \frac{\varepsilon}{4} + \frac{3\delta}{4} > 0$ and $\varepsilon' - \delta' = \frac{\varepsilon-\delta}{2} > 0$. Let $x^1, \ldots, x^m \in Q$ be such that for every $x \in Q$ exists x^i, $i \in \{1, \ldots, m\}$, such that $\|x - x^i\| \leq \nu$, i.e., the set $\mathcal{N} = \{x^1, \ldots, x^m\}$ forms a ν-net in Q. We can choose this net in such a way that Is it possible to use any p ?

$$m \leq \left(\frac{\rho R_p}{\nu}\right)^n \tag{1.32}$$

for a constant $\rho > 0$.

Let $x^1, \ldots, x^m \in Q \subseteq B_p^n(R_p)$ $(p \geq 1)$ be such that for every $x \in Q$ exists x^i, $i \in \{1, \ldots, m\}$, such that $\|x - x^i\|_p \leq \nu$, i.e., the set $\mathcal{N} = \{x^1, \ldots, x^m\}$ forms a ν-net in Q. We can choose this net in such a way that

$$m \leq \left(\frac{\rho R_p}{\nu}\right)^n \tag{1.33}$$

for a constant $\rho > 0$.

If $\mathcal{N} \setminus S^{\varepsilon'}$ is empty, then any point of $\mathcal{N}$ is an ε-optimal solution of the problem (1.20). Otherwise, choose a mapping $u : \mathcal{N} \setminus S^{\varepsilon'} \longrightarrow S$ and consider the sets $\tilde{S} := \bigcup_{x \in \mathcal{N}} u(x)$ and $\tilde{Q} := \mathcal{N} \bigcup \tilde{S}$. Note that $\tilde{Q} \subset Q$ and $|\tilde{Q}| \leq \left(\frac{2\rho R_p}{\nu}\right)^n$.

Now let us replace the set Q by its subset $\tilde{Q}$. We refer to the problem (1.20) and its empirical counterpart as a reduced one for such replacement. We have that $\tilde{S} \subset S$, any point of the set $\tilde{S}$ is an optimal solutions of the true reduced problem and the optimal value of the true reduced problem is equal to the optimal value of the true

(unreduced) problem (1.20). By (1.29) we have that with probability at least $1-\beta/2$ any δ'-optimal solution of the reduced empirical problem is an ε'-optimal solutions of the reduced (and hence unreduced) true problem provided that

$$N \geq \frac{8\lambda^2 R_p^2}{(\varepsilon-\delta)^2} n \ln\left(\frac{8\rho M_p' R_p}{\varepsilon-\delta}\right) + \ln\left(\frac{2}{\beta}\right). \tag{1.34}$$

Note that the right-hand side of (1.47) is greater than or equal to the estimate

$$N \geq \frac{2\sigma^2}{(\varepsilon'-\delta')^2} \ln\left(\frac{|2\tilde{Q}|}{\beta}\right) \tag{1.35}$$

required by (1.29).

We also have by (1.31) that for

$$N \geq \beta^{-1} \ln\left(\frac{2}{\beta}\right)$$

the Lipschitz constant $\bar{M}_p$ is less than or equal to M_p' with probability at least $1-\beta/2$.

Now let $\hat{x}$ be a δ-optimal solution of the unreduced empirical problem. Then there is a point $x' \in \tilde{Q}$ such that $\|\hat{x}-x'\| \leq \nu$, and hence $\bar{f}(x') \leq \bar{f}(\hat{x}) + M_p'\nu$. We also have that the optimal value of the unreduced empirical problem is smaller than or equal to the optimal value of its reduced counterpart. It follows that x' is a δ'-optimal solution of the reduced empirical problem, provided that $\bar{M}_p \leq M_p'$ Consequently, we have that x' is an ε'-optimal solution of the problem (1.20) with probability at least $1-\beta/2$ provided that N satisfies both inequalities (1.47) and (1.35). It follows that

$$f(\hat{x}) \leq f(x')+\nu\mathbb{E}_\xi\left[M_p(\xi)\right] \leq f(x')+M_p'\nu \leq \min_{x\in Q} f(x)+M_p'\nu+\varepsilon' \leq \min_{x\in Q} f(x)+\varepsilon$$

We obtain that if N satisfies both inequalities (1.47) and (1.35), then with probability at least $1-\beta$, any δ-optimal solution of the empirical problem is an ε-optimal solution of the problem (1.47).

By setting $\delta = \varepsilon/2$ the required estimate follows.

it remains to say that $\rho = O(1)$ □

In the close setting online approach gives a better result, see also (1.17) for $p = 2$.

Theorem 1.6 *Assume that* (1.21) *is satisfied and* $f(x,\xi)$ *is convex in* x *in* Q. *Then for* $\bar{x}^N\left(\{\xi^k\}_{k=1}^N\right)$ *(see* (1.18)*) generated by the proper modification of* (1.14)*:*[12]

$$N = \tilde{O}\left(\frac{M_p^2 R_p^2}{\varepsilon^2} \ln\left(\frac{1}{\beta}\right)\right), \text{ if } p \in [1,2],$$

[12] We will talk about «proper» (*Mirror Descent*) modification in the next chapter in more details. Note that for $p \geq 2$ it is proper to use (1.14). The factor $n^{1-2/p}$ appears since the diameter of $B_p^n(R_p)$ in the 2-norm is $O\left(n^{1/2-1/p}R_p\right)$.

$$N = O\left(n^{1-2/p}\frac{M_2^2 R_p^2}{\varepsilon^2}\log\left(\frac{1}{\beta}\right)\right), \text{ if } p > 2.$$

These bounds are optimal up to logarithmic factors in the wide class of all reasonable ways to generate $\bar{x}^N\left(\{\xi^k\}_{k=1}^N\right)$.[13]

It seems that online setting (e.g. for $p = 2$) is better than offline in the sample complexity for convex $f(x,\xi)$ in x. In the next section we show that the gap factor n in the sample complexity bounds between online and offline approaches can be eliminated by the proper regularization.

1.2.2 Strongly convex case and regularization

If $f(x,\xi)$ is μ_p-*strongly convex* in x in the p-norm ($p \geq 1$), that is for all ξ and $x, y \in Q$:

$$f(y,\xi) \geq f(x,\xi) + \langle \nabla_x f(x,\xi), y - x\rangle + \frac{\mu_p}{2}\|y - x\|_p^2, \tag{1.36}$$

then Theorem 1.5 can be improved.

Theorem 1.7 *Assume that* (1.21) *and* (1.36) *are satisfied. Then for*

$$\bar{x}^N_{\delta,\beta/2}\left(\{\xi^k\}_{k=1}^N\right),\ \delta = \frac{\varepsilon^2\mu_p}{8M_p^2} \text{ and } \bar{x}^N\left(\{\xi^k\}_{k=1}^N\right)$$

generated by the proper (restarted[14] *Mirror Descent) modification of* (1.14)*:*

$$N = \tilde{O}\left(\frac{M_p^2}{\mu_p\varepsilon}\log\left(\frac{\log\left(M_p^2/(\mu_p\varepsilon)\right)}{\beta}\right)\right), \quad p \in [1,2]. \tag{1.37}$$

This bound is optimal to up a logarithmic factor in the wide class of all reasonable ways to generate $\bar{x}^N\left(\{\xi^k\}_{k=1}^N\right)$. *Moreover, this bound corresponds* (1.12) *and* (1.19) *when* $p = 2$ *and the bound on* δ *derived from the condition that the first term in RHS of* (1.12) *equals* $\varepsilon/2$*. The bound on* δ *also cannot be improved up to a numerical constant.*

Proof For simplicity, we prove (1.37) only in terms of expectation, rather than high probability bounds.

Nazary, please add the proof of the Theorem based on [111] (https://home.ttic.edu/~nati/Publications/nonlinearTR.pdf) Section 4 and https:

[13] We discuss it also in more details in the next chapter. Also in the next chapter we mention that in the non-convex case online approach gives much worse results $N \propto \varepsilon^{-(n+1)}$, which is also optimal bound for non-convex class of $f(x,\xi)$. Note that the bound on $N \propto n^{1-2/p}M_2^2R_p^2\varepsilon^{-2}$ in the regime $p > 2$ can be refined in the dimension-free case $N \lesssim n : N \propto M_p^p R_p^p \varepsilon^{-p}$ [85].

[14] See the proof of Theorem 1.10 for $p = 2$ and the next chapter in the general case.

//www.jmlr.org/papers/volume2/bousquet02a/bousquet02a.pdf p. 508. Note that the proof of the theorem it's sufficient to describe in terms of expectation, rather that high probability bounds. The proof is based on the concept of uniform stability [16]. Denote

$$\bar{f}^{(i)}(x) = \frac{1}{N}\sum_{k\neq i}^{N} f(x,\xi^k)$$

the empirical average without the i-th sample and let $\bar{x}^{(i)} = \underset{x\in Q\subseteq\mathbb{R}^n}{\operatorname{argmin}} \bar{f}^{(i)}(x)$ be its minimizer. We first establish that the empirical minimizer is $\frac{2M_p^2}{\mu_p N}$ uniformly stable, i.e. that $|f(\bar{x},\xi) - f(\bar{x}^{(i)},\xi)| \leq \frac{2M_p^2}{\mu_p N}$ for all samples and all ξ.

To do so, we first calculate:

$$\begin{aligned}\bar{f}^{(i)}(\bar{x}^{(i)}) - \bar{f}^{(i)}(\bar{x}) &= \frac{f(\bar{x}^{(i)},\xi^i) - f(\bar{x},\xi^i)}{N} + \frac{\sum_{i\neq j}\left(f(\bar{x}^{(i)},\xi^i) - f(\bar{x},\xi^i)\right)}{N}\\ &= \frac{f(\bar{x}^{(i)},\xi^i) - f(\bar{x},\xi^i)}{N} + \frac{N-1}{N}\left(\bar{f}^{(i)}(\bar{x}^{(i)}) - \bar{f}^{(i)}(\bar{x})\right)\\ &\leq \frac{\left|f(\bar{x}^{(i)},\xi^i) - f(\bar{x},\xi^i)\right|}{N} \leq \frac{M_p}{N}\|\bar{x}^{(i)} - \bar{x}\|, \quad (1.38)\end{aligned}$$

where the first inequality follows from the fact that $\bar{x}^{(i)}$ is the minimizer of $\bar{f}^{(i)}(x)$ and in the second inequality we use the Lipschitz property (1.21). But from strong convexity of $\bar{f}(x)$ and the fact that $\bar{x}$ is the minimizer of $\bar{f}(x)$ we also have that $\frac{\mu_p}{2}\|\bar{x}^{(i)} - \bar{x}\|^2 \leq \bar{f}(\bar{x}^{(i)}) - \bar{f}(\bar{x})$. Combining this with (1.38) we obtain $\|\bar{x}^{(i)} - \bar{x}\| \leq \frac{2M_p}{\mu_p N}$ and from Lipschitz continuity (1.21) we get

$$\left|f(\bar{x}^{(i)},\xi) - f(\bar{x},\xi)\right| \leq \frac{2M_p^2}{\mu_p N}.$$

Now, from [16, p.508] we have

$$\mathbb{E}_\xi\left[f(\bar{x}) - \bar{f}(\bar{x})\right] \leq \frac{4M_p^2}{\mu_p N}$$

with $\bar{x}^{(+)} = \underset{x\in Q\subseteq\mathbb{R}^n}{\operatorname{argmin}} \frac{1}{N+1}\sum_{k=1}^{N} f(x,\xi^k) + \frac{1}{N+1} f(x,\xi')$

$$\begin{aligned}\mathbb{E}_{\xi,\xi'}\left[f(\bar{x},\xi') - \bar{f}(\bar{x})\right] &= \frac{1}{N}\sum_{k=1}^{N}\mathbb{E}_{\xi,\xi'}\left[f(\bar{x},\xi') - f(\bar{x},\xi^k)\right] \\ &= \frac{1}{N}\sum_{k\neq i}^{N}\mathbb{E}_{\xi,\xi'}\left[f(\bar{x},\xi') - f(\bar{x}^{(+)}) + f(\bar{x}^{(+)}) - f(\bar{x},\xi^k)\right] \\ &\qquad \text{since } \mathbb{E}_{\xi,\xi'} f(\bar{x}^{(+)},\xi') = \mathbb{E}_{\xi,\xi'} f(\bar{x}^{(+)},\xi^k) \\ &= \frac{1}{N}\sum_{k=1}^{N}\mathbb{E}_{\xi,\xi'}\left[f(\bar{x},\xi') - f(\bar{x}^{(+)},\xi') + f(\bar{x}^{(+)},\xi^k) - f(\bar{x},\xi^k)\right] \\ &\leq \frac{4M_p^2}{\mu_p(N+1)} \leq \frac{4M_p^2}{\mu_p N} \end{aligned} \tag{1.39}$$

Adding $\mathbb{E}_{\xi}\left[\bar{f}(x^*) - f(x^*)\right] = 0$ with $x^* = \underset{x\in Q\subseteq\mathbb{R}^n}{\operatorname{argmin}} f(x)$ to the left-hand side and using the fact that $\bar{x}$ is the minimizer of $\bar{f}(x)$:

$$\begin{aligned}\frac{4M_p^2}{\mu_p N} \geq \mathbb{E}_{\xi}\left[f(\bar{x}) - \bar{f}(\bar{x})\right] &= \mathbb{E}_{\xi}\left[f(\bar{x}) - f(x^*)\right] + \mathbb{E}_{\xi}\left[\bar{f}(x^*) - \bar{f}(\bar{x})\right] \\ &\geq \mathbb{E}_{\xi}\left[f(\bar{x}) - f(x^*)\right]. \end{aligned} \tag{1.40}$$

Due to the fact that the empirical objective $\bar{f}$ is strongly convex, any approximate empirical minimizer must be close to $\bar{x}$, and due to the fact that the expected objective F is Lipschitz-continuous any vector close to $\bar{x}$ cannot have a much worse value than $\bar{x}$. We therefore have, that with probability at least $1-\beta$, for all $x \in Q$:

$$\begin{aligned} f(x) - f(x^*) &= f(x) - f(\bar{x}) + f(\bar{x}) - f(x^*) \leq M_p\|x-\bar{x}\| + f(\bar{x}) - f(x^*) \\ &\leq \sqrt{\frac{2M_p^2}{\mu_p}}\sqrt{\bar{f}(x) - \bar{f}(\bar{x})} + f(\bar{x}) - f(x^*) \leq \sqrt{\frac{2M_p^2}{\mu_p}}\delta + f(\bar{x}) - f(x^*). \end{aligned} \tag{1.41}$$

Taking expectation of (1.41) and using (1.40) and using that x is a δ solution of empirical problem we obtain:

$$f(x) - f(x^*) \leq \sqrt{\frac{2M_p^2}{\mu_p}}\delta + \frac{4M_p^2}{\mu_p N}.$$

Setting $\delta = \frac{\varepsilon^2\mu_p}{8M_p^2}$ and $N = \frac{8M_p^2}{\varepsilon\mu_p}$, $p \in [1,2]$ we have that we have $f(x) - f(x^*) \leq \varepsilon$.

cite the LD result

We emphasise that in Theorem 1.5 $\delta \simeq \varepsilon$, but in Theorem 1.7 $\delta \simeq \frac{\varepsilon^2\mu_p}{M_p^2}$ and the last bound cannot be weakened!

Based on Theorem 1.7 one can derive the result that improve Theorem 1.5 in the convex case ($\mu_p \simeq 0$). Assume for the clarity that $p = 2$.

Lemma 1.1 (Tiknonov's regularization) *Consider regularized stochastic optimization problem*

$$\min_{x \in Q} \left[f_\mu(x) := \mathbb{E}_\xi f(x, \xi) + \frac{\mu}{2} \|x\|_2^2 \right] \tag{1.42}$$

with $\mu = \varepsilon / R_2^2$. *Assume that*

$$f_\mu(\tilde{x}) - \min_{x \in Q} f_\mu(x) \leq \frac{\varepsilon}{2}.$$

Then

$$f(\tilde{x}) - \min_{x \in Q} f(x) = f(\tilde{x}) - f(x^*) \leq \varepsilon.$$

Proof Indeed,

$$\begin{aligned} f(\tilde{x}) - f(x^*) &\leq f_\mu(\tilde{x}) - \left(f_\mu(x^*) - \frac{\mu}{2} \|x^*\|_2^2 \right) \\ &\leq f_\mu(\tilde{x}) - \min_{x \in Q} f_\mu(x) + \frac{\mu}{2} R_2^2 \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \end{aligned}$$

□

The combination of Theorem 1.7 and Lemma 1.1 allow to improve the result of Theorem 1.5 in the convex case.

Theorem 1.8 (the role of the regularization) *Assume that* (1.21) *is satisfied and* $f(x, \xi)$ *is convex in* x *in* Q. *Then for* $\bar{x}^N_{\delta, \beta/2} \left(\{\xi^k\}_{k=1}^N \right)$ *to be a* $\left(\delta = \frac{\varepsilon^3}{8 M_2^2 R_2^2}, \frac{\beta}{2} \right)$*-solution of the empirical version of* (1.42)*:*

$$\min_{x \in Q} \left[\frac{1}{N} \sum_{k=1}^N f(x, \xi^k) + \frac{\varepsilon}{2R_2^2} \|x\|_2^2 \right], \tag{1.43}$$

we have:

$$N = \tilde{O} \left(\frac{M_2^2 R_2^2}{\varepsilon^2} \log \left(\frac{\log (M_2 R_2 / \varepsilon))}{\beta} \right) \right).$$

Moreover, in the general case $p \in [1, 2]$ *the described above technique (with proper regularization) allows to obtain the bounds on* N *that correspond to the bounds in Theorem 1.6 up to logarithmic factors.*

To conclude, from the Theorem 1.8 we derive that in the sample complexity bounds online approach and offline approach (with proper regularization in the convex case) are equivalent up to a logarithmic factors.

1.2.3 s-growth condition

We say that $f(x) := \mathbb{E}_\xi f(x,\xi)$ satisfies *s-growth condition* ($s \geq 1$) on $Q_{2\varepsilon}$ if for all

$$x \in Q_{2\varepsilon} := \{x \in Q : \quad f(x) \leq f(x^*) + 2\varepsilon\} :$$

$$f(x) - f(x^*) \geq \mu_{p,s}\|x - x^*\|_p^s, \tag{1.44}$$

where x^* is a projection of x (in the p-norm) on the set of the solutions of (1.9).

We relax the condition (1.21) as follows: for all $x, y \in Q$ *sub-Gaussian variance* of $f(y,\xi) - f(x,\xi) - (f(y) - f(x))$ bounded from above by $\lambda^2\|y - x\|_p^2$, i.e. for all $t \in \mathbb{R}$:

$$\mathbb{E}_\xi \left[\exp\left(t \cdot (f(y,\xi) - f(x,\xi) - (f(y) - f(x)))\right)\right] \leq \exp\left(t^2\lambda^2\|y - x\|_p^2/2\right). \tag{1.45}$$

Note that if (1.21) holds, then $\lambda^2 \leq 2M^2$.

Theorem 1.9 *Assume that $f(x,\xi)$ is convex in x in Q and* (1.44), (1.45) *are satisfied. Then for $\bar{x}^N_{\varepsilon/2,\beta/2}\left(\{\xi^k\}_{k=1}^N\right)$:*

$$N = O\left(\frac{\lambda_p^2}{\mu_{p,s}^{2/s}\varepsilon^{2(s-1)/s}}\left(n\log\left(\frac{M_p R_{p,\varepsilon}}{\varepsilon}\right) + \log\left(\frac{1}{\beta}\right)\right)\right), \tag{1.46}$$

where $R_{p,\varepsilon}$ is the diameter of $Q_{2\varepsilon}$ in the p-norm. In particular, for $s = 1$ $R_{p,\varepsilon} \leq 4\varepsilon/\mu_{p,1}$. Thus in the case of «sharp minimum» ($s = 1$) N does not depend on ε at all.

The bound (1.50) *is optimal up to a logarithmic factor.*

Proof Nazary, please add the proof of the Theorem based on [114] (https://cpn-us-w2.wpmucdn.com/sites.gatech.edu/dist/4/1470/files/2021/03/SPbook.pdf) Section 5.3.2.

It was assumed in the proof of the Theorem 1.5 that the set Q has a finite diameter, i.e., that Q is bounded. For convex problems, this assumption can be relaxed. Assume that the problem is convex, the optimal value $\min\limits_{x\in Q} f(x)$ of the true problem is finite, and for some $a > \varepsilon$ the set S^a has a finite diameter R_p^a. (Recall that $S^a := x \in Q : f(x) \leq \min\limits_{x\in Q} f(x) + a$.) We refer here to the respective true and empirical problems, obtained by replacing the feasible set Q by its subset S^a, as reduced problems. Note that the set S^ε , of ε-optimal solutions, of the reduced and original true problems are the same. Let N be an integer satisfying the inequality from Theorem 1.5, e.g.

$$N \geq \max\left\{\frac{8\lambda^2 R_p^2}{(\varepsilon-\delta)^2} n\ln\left(\frac{8\rho M_p' R_p}{\varepsilon-\delta}\right) + \ln\left(\frac{2}{\beta}\right), \beta^{-1}\ln\left(\frac{2}{\beta}\right)\right\}. \tag{1.47}$$

with R_p replaced by R_p^a. Then, under the assumptions of Theorem 1.5, we have that with probability at least $1-\beta$ all δ-optimal solutions of the reduced SAA problem are ε-optimal solutions of the true problem. Let us observe now that in this case the set of δ-optimal solutions of the reduced SAA problem coincides with the set of δ-optimal solutions of the original SAA problem. Indeed, suppose that the original SAA problem has a δ-optimal solution $x^* \in Q \setminus S^a$. Let $\bar{x} \in \operatorname{Arg\,min}_{x\in S^a} \bar{f}(x)$, such a minimizer does exist since $\bar{x} \in S^\varepsilon$ and S^a is compact and $\bar{f}(x)$ is real valued convex and hence continuous. Then $\bar{x} \in S^\varepsilon$ and $\bar{f}(x^*) \le \bar{f}(\bar{x}) + \delta$. By convexity of $\bar{f}(x)$ it follows that $\bar{f}(x) \le \max\left\{\bar{f}(\bar{x}), \bar{f}(x^*)\right\}$ for all x on the segment joining $\bar{x}$ and x^*. This segment has a common point $\hat{x}$ with the set $S^a \setminus S^\varepsilon$. We obtain that $\hat{x} \in S^a \setminus S^\varepsilon$ is a δ-optimal solutions of the reduced SAA problem, a contradiction.

That is, with such sample size N^* we are guaranteed with probability at least $1-\beta$ that any δ-optimal solution of the SAA problem is an ε-optimal solution of the true problem. Also, assumptions (1.30) and (1.24) should be verified for x, x' in the set S^a only.

Consider further $a = 2\varepsilon$ and suppose that the set S of optimal solutions of the true problem is nonempty. Then it follows from the proof of Theorem 1.5 that it suffices to verify assumption (1.24) only for every $x \in Q \setminus S^{\varepsilon'}$ and $x' := u(x)$, where $u : Q \setminus S^\varepsilon \longrightarrow S$ and $\varepsilon' := \frac{3}{4}\varepsilon + \frac{\delta}{4}$. If the set S is closed, we can use, for instance, a mapping $u(x)$ assigning to each $x \in Q \setminus S^{\varepsilon'}$ a point of S closest to x. If, moreover, the set S is convex and the employed norm is strictly convex (e.g., the Euclidean norm), then such mapping (called metric projection onto S) is defined uniquely. Then for such x and x' we we can bound $\|x - x'\|_p$ in (1.28) by $\|x - x'\|_p \le \sup_{x\in Q\setminus S^{\varepsilon'}, x'\in S} \|x - x'\|_p$. The growth condition (1.44) implies that $S = \{x^*\}$ and that for any $x \in Q_{2\varepsilon}$ the inequality $\|x - x^*\|_p \le \left(\frac{2\varepsilon}{\mu_{p,s}}\right)^{1/s}$ holds.

Since the problem is convex we can use $Q_{2\varepsilon}$ instead of Q to reproduce step (1.33).

Reproducing the proof of Theorem 1.5 with the above refinements we obtain from the result of the theorem.

Optimality of (1.50)

Consider a simple example for $p = 2$, $Q = B_2^n(1)$:

$$f(x, \xi) = \|x\|_2^s - s\langle \xi, x\rangle,$$

$\xi \sim \mathcal{N}(0, \sigma^2 I_n)$, where I_n is the identity $n \times n$ matrix. Hence $f(x) = \|x\|_2^s$, $x^* = 0$, $\mu_{2,s} = 1$ in (1.44) and

$$f(y, \xi) - f(x, \xi) - (f(y) - f(x)) = s\langle \xi, y - x\rangle$$

has $\mathcal{N}\left(0, s^2\sigma^2\|y - x\|_2^2\right)$-distribution. Therefore $\lambda^2 = s^2\sigma^2$ in (1.45).

Note also that

$$\bar{f}(x) = \|x\|_2^s - s\langle \bar{\xi}_N, x\rangle,$$

where $\bar{\xi}_N \sim \mathcal{N}(0, \sigma^2 N^{-1} I_n)$. For this problem we can explicitly find the minimizer of the empirical loss

$$\hat{x}^N \in \arg\min_{x \in Q} \bar{f}(x) = \frac{\bar{\xi}_N}{\|\bar{\xi}_N\|_2^b},$$

where

$$b = \begin{cases} 1, & \text{if } \|\bar{\xi}_N\|_2 > 1 \\ \frac{s-2}{s-1}, & \text{else.} \end{cases}$$

Since $f(x) = \|x\|_2^s$, it follows that

$$f(\hat{x}^N) - f(x^*) \leq \varepsilon$$

is equivalent to

$$\|\bar{\xi}_N\|_2^{\frac{s}{s-1}} \leq \varepsilon$$

for sufficiently small ε. Combining this with $\bar{\xi}_N \sim \mathcal{N}(0, \sigma^2 N^{-1} I_n)$ we can get that for

$$\mathbb{P}\left(f\left(\hat{x}^N\right) - f(x^*) \leq \varepsilon\right) = \mathbb{P}_{\bar{\xi}_N \sim \mathcal{N}(0, \sigma^2 N^{-1} I_n)}\left(\left\|\bar{\xi}_N\right\|_2^{\frac{s}{s-1}} \leq \varepsilon\right) \geq 0.7$$

it is required that

$$N > \frac{n\sigma^2}{\varepsilon^{2(s-1)/s}}. \tag{1.48}$$

The lower bound (1.48) corresponds to (1.50) when $\mu_{2,s} = 1$ and s is finite. Note that when $s = 2$ and $\mu_{2,2} = \mu_2 \neq 1$ (1.48) can be clarified as follows

$$N \geq \frac{n\sigma^2}{\mu_2 \varepsilon}.$$

The last lower bound seems to be strange enough (n-factor in the lower bound looks wrong) due to the upper bound from (1.37). But there is no contradiction here even with the strengthened upper bound from (1.37)

$$N = \tilde{O}\left(\frac{\tilde{M}_2^2}{\mu_2 \varepsilon}\right),$$

since $\tilde{M}_2^2 := \mathbb{E}_\xi\left[M_2(\xi)^2\right] = ns^2\sigma^2$,[15] where $M_2(\xi)$ is defined with $p = 2$ according to the following: for all ξ and $x, y \in Q$:

$$|f(y, \xi) - f(x, \xi)| \leq M_p(\xi)\|y - x\|_p. \tag{1.49}$$

[15] In (1.37) it is assumed that there exists such M_2 that $M_2(\xi) \leq M_2$. Here we relax the notion of M_2 to $\tilde{M}_2$.

Theorem 1.10 *Assume that $f(x,\xi)$ is convex in x in Q, $f(x) := \mathbb{E}_\xi f(x,\xi)$ satisfies r-growth condition in Q and* (1.21) *is satisfied. Then for $\bar{x}^N\left(\{\xi^k\}_{k=1}^N\right)$ generated by the proper (restarted Mirror Descent) modification of* (1.14)*:*

$$N = \tilde{O}\left(\frac{M_p^2}{\mu_{p,s}^{2/s}\varepsilon^{2(s-1)/s}}\right), \quad p \in [1,2]. \tag{1.50}$$

This bound is optimal up to logarithmic factor in the wide class of all reasonable ways to generate $\bar{x}^N\left(\{\xi^k\}_{k=1}^N\right)$.

Proof For clarity we consider only the euclidean case $p = 2$. Since $f(x) := \mathbb{E}_\xi f(x,\xi)$ satisfies s-growth condition in Q, it follows from (1.17) that with probability at least $1 - \beta/\kappa$

$$\mu_{2,s}\|\bar{x}^N - x^*\|_2^s \le f(\bar{x}^N) - f(x^*) = O\left(\frac{M_2\|x^1 - x^*\|_2\sqrt{\log(\kappa/\beta)}}{\sqrt{N}}\right),$$

where $\bar{x}^N$ is calculated according to (1.18) based on (1.14). If we choose

$$N = O\left(\frac{M_2^2\log(\kappa/\beta)}{\mu_{2,s}^2\|x^1 - x^*\|_2^{2(s-1)}}\right),$$

then

$$\|\bar{x}^N - x^*\|_2^s = \frac{1}{2}\|x^1 - x^*\|_2^s.$$

The idea of the *restart technique* is to put

$$x^1 := \bar{x}^N$$

and to restart algorithm (1.14). By denoting $R_{2,l}$ the distance between the starting point and the solution x^* at l-th restart, we could guarantee that $R_{2,l+1}^s = R_{2,1}^s 2^{-l}$. Similarly, N_l is a number of iteration at l-th restart. Since we would like to solve the problem with probability at least $1 - \beta$ and with accuracy ε, the number of the restarts κ is determined from

$$\frac{M_2 R_{2,\kappa+1}\sqrt{\log(\kappa/\beta)}}{\sqrt{N_{\kappa+1}}} \simeq \mu_{2,r}R_{2,\kappa+1}^s = \mu_{2,r}R_{2,1}^s 2^{-(\kappa+1)} \simeq \varepsilon.$$

Therefore the total number of samples (iterations) is

$$\sum_{l=1}^{\kappa} O\left(\frac{M_2^2 \log(\kappa/\beta)}{\mu_{2,s}^2 R_{2,l}^{2(s-1)}}\right) = \frac{M_2^2 \log(\kappa/\beta)}{\mu_{2,s}^2 R_{2,1}^{2(s-1)}} \sum_{l=1}^{\kappa} O\left(2^{\frac{2(s-1)}{s} l}\right)$$
$$= O\left(\frac{M_2^2 \log(\kappa/\beta)}{\mu_{2,s}^2 R_{2,1}^{2(s-1)}} 2^{\frac{2(s-1)}{r}(\kappa+1)}\right)$$
$$= O\left(\frac{M_2^2 \log(\kappa/\beta)}{\mu_{2,s}^2 R_{2,1}^{2(s-1)}} \left(\frac{\mu_{2,s} R_{2,1}^s}{\varepsilon}\right)^{\frac{2(s-1)}{r}}\right)$$
$$= O\left(\frac{M_2^2 \log(\kappa/\beta)}{\mu_{2,s}^{2/s} \varepsilon^{2(s-1)/s}}\right).$$

□

1.3 Concluding remarks

For a better structure of this chapter we have collected various comments that clarify the results given above (but have not a primal interest) in this (separate) section at the end of the chapter. In more details most of this comments will be further developed in the next chapters.

1.3.1 Weakening of uniform Lipschitz condition in online approach

An important remark concerns online approach is that we can significantly relax uniform Lipschitz continuity property (1.21), assuming that $M_p(\xi)$ in (1.49) has a finite second moment $\mathbb{E}_\xi\left[M_p(\xi)^2\right] < \infty$. In this case, all the bound remain the same up to a logarithmic factor, see [81, 45] for $p = 2$, and [83, 57] for $p \geq 1$, but for the convergence in expectation. If we have only $\mathbb{E}_\xi\left[M_p(\xi)^{1+\alpha}\right] < \infty$, where $\alpha > 0$ then in the dimension-free case ($N \lesssim n$) the expected $N \sim \varepsilon^{-\max\{2,p\}}$ will get worse $N \sim \varepsilon^{-(1+\alpha_p)/\alpha_p}$, where $\alpha_p = \min\left\{1, \alpha, (p-1)^{-1}\right\}$ [85, 128]. Similarly, in the strongly convex case (s-growth condition) and in the case $N \gtrsim n$. Note that high-probability bound analysis has been developed in this generality mainly for the Euclidean proximal setup with $\mathbb{E}_\xi\left[M_2(\xi)^2\right] < \infty$.

For offline approach some particular results in this direction are also known, see the references in [114].

1.3.2 Weakening of the convexity condition

The principal difference between online and offline approach is that for optimal results in offline approach the convexity of $f(x,\xi)$ in x for all ξ is typically required. It was shown in [109] that for any regularizer there is a stochastic optimization problem with convex $F(x)$ such that regularized empirical loss minimization approach fails to learn. But for online approach the convexity of $F(x)$ is enough for the same rates of convergences in terms of convergence in expectation [83, 57].

In Section 1.2.1 we have observed that offline approach in non-convex case required $N \propto n\varepsilon^{-2}$ samples despite the fact that online approach in non-convex case required $N \propto \varepsilon^{-(n+1)}$ samples. Moreover, under different additional assumptions [110, 101, 8] (finite VC-dimension e.t.c.) the dependence of n in offline approach $N \propto n\varepsilon^{-2}$ can be relaxed.[16] So it seems that offline approach is much better than online. In terms of the sample complexity (number of different samples of ξ) it really is. But at the end in offline approach we should solve empirical loss (risk) minimization problem that would be non-convex. To solve this problem we required $N \propto n\varepsilon^{-(n+1)}$ stochastic gradient oracle calls[17] that corresponds (up to a factor n) to online approach.

Some results that were mentioned in the previous sections can be generalized if we replace (strong) convexity assumption by quasi-convexity or some growth condition [82] or Polyak–Lojasiewicz(–Lezansky) condition [60, 11]. For example, online and offline approaches under Polyak–Lojasiewicz condition are considered in [3] and [73].

1.3.3 How to make online approach adaptive?

To answer this question, we appeal to SGD (1.14)

$$x^{k+1} = \pi_Q\left(x^k - \gamma_k \nabla_x f(x^k, \xi^k)\right),$$

with

$$\gamma_k \equiv \frac{R}{M\sqrt{N}}.$$

The problem is that the stepsize policy requires the knowledge of parameters R and M. Moreover, this stepsize policy in not adaptive in N, i.e., we should know the desired N in advance. The last problem was solved in [83] by changing

$$\gamma_k \equiv \frac{R}{M\sqrt{k}}.$$

[16] Factor n is replaced by the «efficient» dimension, which could be much smaller.

[17] This bound can be improved a little bit by using the fact that all the terms in the sum (the empirical loss) have the same distribution. But this improvement will have a minor effect on the total oracle complexity.

This stepsize policy leads to the same convergence rate up to a logarithmic factor. This factor can be eliminated by the Nesterov's dual extrapolations scheme [88]. The problem of unknown M-parameter was further solved in [24, 25], where it was proved that for

$$\gamma_k \equiv \frac{R}{\sqrt{\sum_{j=1}^{k} \|\nabla_x f(x^j, \xi^j)\|_2^2}}$$

the convergence rate does not change up to a numerical constant factor. SGD with this stepsize policy is known as AdaGrad. The works [83, 24] largely determined the development of modern stochastic optimization. For example, one of the most cited stochastic optimization algorithm after SGD is Adam [61, 102], which is based on AdaGrad. This algorithm and its variations are one of the main tools to train Deep Neural Networks [70, 123].

Although in practice different adaptive algorithms show themselves well in the theory typically they converge in the worst case not better than non-adaptive analogues [9, 34].

1.3.4 Overparametrization

In practice for the strongly convex problems ($f(x)$ is μ-strongly convex in the 2-norm):

$$\min_{x \in \mathbb{R}^n} \left[f(x) := \mathbb{E}_\xi f(x, \xi) \right]$$

with uniformly Lipschitz continuous gradient: for all ξ and $x, y \in \mathbb{R}^n$:[18]

$$\|\nabla_x f(y, \xi) - \nabla_x f(x, \xi)\|_2 \leq L \|y - x\|_2 \tag{1.51}$$

it was observed that simple stochastic gradient method (SGD):

$$x^{k+1} = x^k - \gamma \nabla_x f(x^k, \xi^k)$$

converges with linear rate in the vicinity of the solution x^* [79]. That was also confirmed in the theory

$$\mathbb{E}_{\{\xi^k\}_{k=1}^N} \left[|x^{N+1} - x^*\|_2^2 \right] \leq \|x^1 - x^*\|_2^2 (1 - \gamma\mu)^N + \frac{2\gamma\sigma_*^2}{\mu}, \tag{1.52}$$

where the stepsize $\gamma \leq 1/(2L)$ and

$$\sigma_*^2 = \mathbb{E}_\xi \left[\|\nabla_x f(x^*, \xi) - \nabla f(x^*)\|_2^2 \right] = \mathbb{E}_\xi \left[\|\nabla_x f(x^*, \xi)\|_2^2 \right],$$

[18] As we will see in the next chapters it sufficiently to consider Lipschitz-type conditions only in some balls centered at starting point and radius determined (up to a logarithmic factor) by the distance between starting point and the closest to this point solution.

since $\nabla f(x^*) = 0$.

Indeed, from Section 1.1.2:

$$\|x^{k+1} - x^*\|_2^2 \leq \|x^k - x^*\|_2^2 - 2\gamma\langle\nabla_x f(x^k, \xi^k), x^k - x^*\rangle + \gamma^2\|\nabla_x f(x^k, \xi^k)\|_2^2.$$

Taking the conditional expectation in ξ^k under fixed x^k and using that

$$\begin{aligned}\mathbb{E}_\xi\left[\|\nabla_x f(x,\xi)\|_2^2\right] &\leq 2\mathbb{E}_\xi\left[\|\nabla_x f(x,\xi) - \nabla_x f(x^*,\xi)\|_2^2\right] + 2\mathbb{E}_\xi\left[\|\nabla_x f(x^*,\xi)\|_2^2\right] \\ &\leq 4L\mathbb{E}_\xi\left[f(x,\xi) - f(x^*,\xi) - \langle\nabla_x f(x^*,\xi), x - x^*\rangle\right] + 2\sigma_*^2 \\ &= 4L\left(f(x) - f(x^*)\right) + 2\sigma_*^2,\end{aligned}$$

we obtain[19]

$$\begin{aligned}\mathbb{E}_{\xi^k}\left[\|x^{k+1} - x^*\|_2^2 | x^k\right] &\leq \|x^k - x^*\|_2^2 - 2\gamma\langle\nabla f(x^k), x^k - x^*\rangle \\ &\quad + \gamma^2\left(4L\left(f(x^k) - f(x^*)\right) + 2\sigma_*^2\right) \\ &\leq \|x^k - x^*\|_2^2 - 2\gamma\left(f(x^k) - f(x^*) + \frac{\mu}{2}\|x^k - x^*\|_2^2\right) \\ &\quad + 4Lh^2\left(f(x^k) - f(x^*)\right) + 2\gamma^2\sigma_*^2.\end{aligned}$$

Rearranging the terms in the RHS and taking the expectation in x^k we come to the following:

$$\begin{aligned}\mathbb{E}_{\{\xi^j\}_{j=1}^k}\left[\|x^{k+1} - x^*\|_2^2\right] &\leq (1 - \gamma\mu)\,\mathbb{E}_{\{\xi^j\}_{j=1}^{k-1}}\left[\|x^k - x^*\|_2^2\right] \\ &\quad + 2\gamma\left(1 - 2L\gamma\right)\left(\mathbb{E}_{\{\xi^j\}_{j=1}^{k-1}}\left[f(x^k)\right] - f(x^*)\right) + 2\gamma^2\sigma_*^2 \\ &\leq (1 - \gamma\mu)\,\mathbb{E}_{\{\xi^j\}_{j=1}^{k-1}}\left[\|x^k - x^*\|_2^2\right] + 2\gamma^2\sigma_*^2,\end{aligned}$$

if $\gamma \leq 1/(2L)$. So we come to (1.52) by induction.

The overparametrization effect appears if σ_*^2 is small, that is $\nabla_x f(x^*, \xi) \simeq 0$ for almost all ξ.

For example if we consider offline approach

$$\min_{x\in\mathbb{R}^n}\left[\bar{f}(x) := \frac{1}{N}\sum_{k=1}^N f(x, \xi^k)\right]$$

and reformulate this problems as

$$\min_{x\in\mathbb{R}^n}\left[\bar{f}(x) := \mathbb{E}_k f(x, \xi^k),\right] \tag{1.53}$$

[19] The last inequality uses the weaker variant of μ-strong convexity assumption of $f(x)$: for all $x \in \mathbb{R}^n$

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

where $\mathbb{P}(k = l) = 1/N$ for $l = 1, ..., N$. In this case $L = \max_{l=1,...,N} L_l$, where L_l is Lipschitz gradient constant of $f(x, \xi^l)$ in x. The variance is

$$\sigma_*^2 = \frac{1}{N} \sum_{k=1}^{N} \|\nabla_x f(x^*, \xi^k)\|_2^2.$$

If $\nabla_x f(x^*, \xi^k) \simeq 0$, which could be possible due to the same stochastic nature of all the terms $f(x, \xi^k)$, then for all $k = 1, ..., N$ we have overparametrization and effect of linear convergence of SGD to a small vicinity of the solution.

Although overparameterized problems have attracted considerable attention in recent years, the results available here are still far from theory we have described in the previous sections. For example, in offline approach with $\sigma_*^2 \simeq 0$ we have only [73]:

$$\mathbb{E}_{\{\xi^k\}_{k=1}^N} \left[\|\hat{x}^N - x^*\|_2^2\right] \propto \frac{1}{\mu^2 N^2},$$

rather than we have in online approach with proper stepsize policy $\gamma = 1/(2L)$:

$$\mathbb{E}_{\{\xi^k\}_{k=1}^N} \left[\|x^{N+1} - x^*\|_2^2\right] \propto \left(1 - \frac{\mu}{2L}\right)^N.$$

Little is known about overparameterization in a non-Euclidean proximal setup.

1.3.5 Acceleration and batching for smooth convex optimization problems in online approach

Consider smooth convex optimization problem

$$\min_{x \in Q} f(x), \tag{1.54}$$

where for all $x, y \in Q$:

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|y - x\|_2. \tag{1.55}$$

Accelerated method [89, 69, 76] allows to solve smooth convex optimization problems with the rate

$$f(x^N) - f(x^*) \lesssim \frac{LR^2}{N^2},$$

where $R^2 = \|x^1 - x^*\|_2^2$ and x^* is the closest solution (in 2-norm) to x^1 if the set of the solutions contains more than one point. Below we describe how to build accelerated batch-parallelized algorithm that significantly outperform SGD in the number of subsequent iterations.

First of all, following [22, 31, 29] we introduce the notion of (δ_1, δ_2, L)-oracle. We say that for the problem (1.54) we have an access to (δ_1, δ_2, L)-oracle at a point

x if we can evaluate a vector $\nabla_\delta f(x)$ such that, for all $x, y \in Q$,

$$-\delta_1 \le f(y) - f(x) - \langle \nabla_\delta f(x), y - x\rangle \le \frac{L}{2}\|y - x\|_2^2 + \delta_2,$$

where $\mathbb{E}\delta_1 = 0$ (δ_1 is independently taken at each oracle call), $\mathbb{E}\delta_2 \le \delta$. Note that the left inequality corresponds to the definition of δ_1-(sub)gradient [95] and reduces to the convexity property in the case $\delta_1 = 0$. In this case the LHS holds with $\nabla_\delta f(x) = \nabla f(x)$. The right inequality in the case when $\delta_2 = 0$ is a consequence[20] of (1.55). Let us consider an algorithm $\mathbf{A}(L, \delta_1, \delta_2)$ that converges with the rate[21]

$$\mathbb{E}f(x^N) - f(x^*) = O\left(\frac{LR^2}{N^\alpha} + N^\zeta \delta\right). \tag{1.56}$$

The *batching technique*, applied to the problem (1.54) with L-Lipschitz gradient (in 2-norm), is based on the use of the mini-batch stochastic approximation of the gradient

$$\nabla_\delta f(x) = \frac{1}{r}\sum_{j=1}^{r} \nabla_x f(x, \xi^j)$$

in $\mathbf{A}(L, \delta_1, \delta_2)$, where $\{\xi^j\}_{j=1}^r$ are sampled independently and r is an appropriate batch size. The choice of r is based on the following relations

$$\langle \nabla_\delta f(x) - \nabla f(x), y - x\rangle \le \frac{1}{2L}\|\nabla_\delta f(x) - \nabla f(x)\|_2^2 + \frac{L}{2}\|y - x\|_2^2,$$

$$\mathbb{E}_{\{\xi^j\}_{j=1}^r}\left[\|\nabla_\delta f(x) - \nabla f(x)\|_2^2\right] \le \frac{\sigma^2}{r},$$

where σ^2 is the variance of unbiased stochastic gradient $\nabla_x f(x, \xi)$, which is available. Hence, if

$$\delta \le \frac{1}{2L}\max_{x\in Q} \mathbb{E}_{\{\xi^j\}_{j=1}^r}\left[\|\nabla_\delta f(x) - \nabla f(x)\|_2^2\right],$$

i.e. $\delta = \sigma^2/(2Lr)$, we have that $\mathbf{A}(2L, \delta_1, \delta_2)$ converges with the rate given in (1.56). From (1.56) we see that to obtain

$$\mathbb{E}f(x^N) - f(x^*) \le \varepsilon$$

it suffices to take

---

[20] Note, that the right inequality in the case when $\delta_2 = 0$ is not equivalent to (1.55), but is typically sufficient to obtain optimal (up to constant factors) bounds on the rate of convergence of different methods [126].

[21] N is a number of iterations which up to a constant factor is equal to the number of (δ_1, δ_2, L)-oracle calls. We can consider more specific rates of convergence for problems with additional structure and develop *batching technique* in a similar way.

$$N = O\left(\left(\frac{LR^2}{\varepsilon}\right)^{1/\alpha}\right) \quad \text{and} \quad r = O\left(\frac{\sigma^2 N^\beta}{L\varepsilon}\right).$$

In particular, for all known Accelerated gradient methods we have that $\alpha = 2$, $\zeta = 1$ [22, 31]. In this case, we obtain the complexity bounds for batched Accelerated gradient methods (assume that σ^2 is such that $T \geq N$, otherwise we put $T := N$):

$$N = O\left(\sqrt{LR^2/\varepsilon}\right), r = O\left(\sigma^2 R/\left(\sqrt{L}\varepsilon^{3/2}\right)\right), T = N \cdot r = O\left(\sigma^2 R^2/\varepsilon^2\right).$$

It is obvious that we can calculate batch in a parallel manner. This reduces the number of subsequent iterations from $N \propto \varepsilon^{-2}$ for standard SGD with small stepsize (see Section 1.1.2) and $N \propto \varepsilon^{-1}$ for SGD with special stepsize policy $\gamma \simeq \min\{1/L, 1/(\mu N)\}$ [124] (see Section 1.3.4) to the optimal rate $N \propto \varepsilon^{-1/2}$ [85, 133]. Recently [131] this result was generalized to overparametrized problems, see Section 1.3.4.

The described above batching technique is very important and universal technique, which allows to build (optimal) stochastic algorithms based on the (optimal) deterministic algorithms and their analysis of convergence with inexact oracle. We mention here only the two most recent examples. In [40] batching technique was successfully applied in gradient-free optimization. In [77] batching technique was successfully applied for distributed strongly convex-concave saddle-point problems with different constants of strong convexity and strong concavity.

Note that the described technique can be further generalized to strongly convex problems (problems with s-growth condition) and non-Euclidean proximal setup [32, 44].

1.3.6 Sum-type problems and offline approach

At the very end of the offline approach we should solve the empirical loss (risk) minimization problem

$$\min_{x \in Q}\left[\bar{f}(x) := \frac{1}{N}\sum_{k=1}^{N} f(x, \xi^k)\right]. \tag{1.57}$$

For clarity, we assume that $f(x, \xi)$ is μ-strongly convex and M-Lipscitz continuous in x in 2-norm, see (1.36), (1.21). According to Theorem 1.7 $N = \tilde{O}\left(M^2/(\mu\varepsilon)\right)$ and we should solve (1.57) with the accuracy $\delta \simeq \varepsilon^2\mu/M^2$. Unfortunately, without additional assumptions on the smoothness of $f(x, \xi)$ the complexity of this problem (the number of $\nabla_x f(x, \xi^k)$ calculations) is $\tilde{O}\left(M^2/(\mu\delta)\right)$ [85]. That is much worse than N. If we additionally assume that $f(x, \xi)$ has L-Lipschitz continuous gradient in x in the 2-norm, see (1.51), then we can apply batch-parallelization and acceleration in the number of subsequent iterations, see Section 1.3.5. But this trick does not solve the problem of oracle complexity. We still require $\tilde{O}\left(M^2/(\mu\delta)\right)$ calculations of $\nabla_x f(x, \xi^k)$. It seems that we come to some contradiction. Offline approach seems

to be worse everytime than online one in terms of the oracle complexity. Fortunately, this is not the case. There exist randomized Variance Reduced (VR) algorithms (see, e.g. [132, 69, 76]) that allow to solve (1.57) (with accuracy δ) with the complexity:[22]

$$\tilde{O}\left(\left(N + \sqrt{N\frac{L}{\mu}}\right)\log\left(\frac{\Delta f}{\delta}\right)\right). \tag{1.58}$$

Under the natural assumption $L/\mu \lesssim N \simeq M^2/(\mu\varepsilon)$, i.e.[23] $L \lesssim M^2/\varepsilon$ this complexity coincides with N up to a logarithmic factor.

Moreover, for many concrete problems (e.g. Soft-SVM, see Section 1.1.1) we can efficiently reduce originally non-smooth problems to smooth one [5] and apply the VR algorithms.

The modern theory of VR methods is well developed, see e.g. [69]. For example, it includes non-Euclidean proximal setup.

In the core of VR methods lies a very simple idea, which goes back to Monte Carlo theory. Instead of stochastic gradient $\nabla_x f(x,\xi)$ it is proposed to consider the reduced one

$$\tilde{\nabla}_x f(x,\xi) = \nabla_x f(x,\xi) - \nabla_x f(\hat{x}^N,\xi),$$

where $\hat{x}^N$ is the solution of (1.57). Note that with stochastic gradient we have overparametization effect $\tilde{\nabla}_x f(\hat{x}^N,\xi) = 0$ (for all ξ) and therefore we can expect a linear convergence. Unfortunately in this form VR trick is not practical, since it is required to know $\hat{x}^N$. The proper correction of the trick consist in replacing $\nabla_x f(x^k,\xi^{t(k)})$ (where $t(k)$ is an index that equally likely and independently selected among $1,...,N$ at k-th iteration) with

$$\tilde{\nabla}_x f(x^k,\xi^{t(k)}) := \nabla_x f(x^k,\xi^{t(k)}) - \nabla_x f(\bar{x}^k,\xi^{t(k)}) + \nabla\bar{f}(\bar{x}^k),$$

where $\bar{x}^k$ periodically updated as $\bar{x}^k := x^k$ according to the different policies [69, 67]. With this stochastic gradient we may also expect overparametrization along with the convergence $x^k \to \hat{x}^N$. Indeed,

$$\mathbb{E}_{\xi^{t(k)}}\left[\|\tilde{\nabla}_x f(x^k,\xi^{t(k)})\|_2^2\right] \lesssim L\left(\bar{f}(\bar{x}^k) - \bar{f}(\hat{x}^N)\right) \to 0 \tag{1.59}$$

along with $\bar{x}^k \to \hat{x}^N$.

[22] This bound is optimal [132, 69], i.e. there are no algorithms that work only with $\nabla_x f(x,\xi^k)$ and has a better complexity.

[23] One can always achieve this condition by smoothing a non-smooth problem. In this case $L \simeq M^2/\varepsilon$ [87, 132].

1.3.7 Composite optimization

From the previous sections we have known that regularizers in the empirical loss (risk) minimization approach play an important role. Sometimes this regularizers spoil the properties of the problem, e.g. $\|x\|_1$-regularizer in LASSO makes the problem non-smooth, see Section 1.1.1. We can solve this issue by using composite optimization approach. Let us remind that standard SGD (1.14) has a following structure:

$$\begin{aligned} x^{k+1} &= \pi_Q\left(x^k - \gamma_k \nabla_x f(x^k, \xi^k)\right) \\ &= \arg\min_{x \in Q}\left\{\langle \nabla_x f(x^k, \xi^k), x - x^k\rangle + \frac{1}{2\gamma_k}\|x - x^k\|_2^2\right\}. \end{aligned}$$

If the stochastic optimization problem is regularized (i.e. has composite term):

$$\min_{x \in Q}\left[\mathbb{E}_\xi\left[f(x, \xi)\right] + h(x)\right].$$

we could correct the described procedure as follows [136]:

$$x^{k+1} = \arg\min_{x \in Q}\left\{\langle \nabla_x f(x^k, \xi^k), x - x^k\rangle + \frac{1}{2\gamma_k}\|x - x^k\|_2^2 + h(x)\right\}.$$

The iteration complexity does not change. But the auxiliary (projection) problem becomes more difficult. Fortunately, for some concrete examples (e.g. LASSO) the auxiliary problem almost retains its complexity. In this case composite term is called «proximal-friendly». The same holds true for Accelerated batched algorithms [23] and VR algorithms [69].

In the case of non proximal-friendly composite terms it happens that we can split the oracle complexities for two terms [69, 64], see also Section 1.3.9. This turned out to be an extremely useful option in distributed optimization [69, 46, 107, 64].

Composite optimization was firstly developed in deterministic setup [10, 26, 86]. Moreover, in [84, 125] it was considered more general «model setup» with $f(x) := \min\{f_1(x), ..., f_m(x)\}$ and composite optimization as particular cases. Under some assumptions this model setup can be further developed on stochastic optimization problems [31].

1.3.8 Overfitting and early stopping for offline approach

Let us return to the empirical problem (1.57):

$$\min_{x \in Q}\left[\bar{f}(x) := \frac{1}{N}\sum_{k=1}^{N} f(x, \xi^k)\right].$$

In Section 1.2.2 we describe regularization trick, that allows to align sample complexities for offline and online approaches for convex, but non-strongly convex problems. Another (a little artificial) way to align sample complexities in both of the approaches is to change the way of obtaining $\bar{x}^N_{\delta,\beta/2}\left(\{\xi^k\}_{k=1}^N\right)$, which is based on sufficiently accurate solution of (1.57) (or its regularized version). The idea is trivial: to «solve» (1.57) by using SGD with samples $\{\xi^k\}_{k=1}^N$ without repeating. So the first N iterations of this algorithm is completely coincide with standard SGD iterations (1.14). An interesting phenomenon is that further iterations of SGD based on the same sample set $\{\xi^k\}_{k=1}^N$ not only not improve the quality of the solution (this quality is measured in terms of initial stochastic optimization problem!), but can also provably lead to a decrease in quality (overfitting).

This idea was further developed in seminal work [49], where it was shown that for the standard SGD (with output $\bar{x}^T$ after T iterations, see (1.18)) applied to smooth convex (but not strongly convex!) empirical problem (without any regularization!) in the expectation form (1.53):

$$f(\bar{x}^T) - f(x^*) \propto N^{-1/2} \text{ if } T \propto N.$$

This phenomenon sometimes called «early stopping» [41]. The work [49] initiated a lot of activity around overfitting properties of SGD applied to the empirical problems, see the survey in [73]. In particular, for smooth convex (but not strongly convex!) problems in [75] it was shown that

$$f(\bar{x}^T) - f(x^*) \propto N^{-\eta/(1+\eta)} \text{ if } T \propto N^{2/(1+\eta)}, \eta \in (0, 1].$$

It means that too many iteration lead to overfitting. For smooth strongly convex problems it was shown [73] that

$$f(\bar{x}^T) - f(x^*) \propto (\mu N)^{-1} \text{ if } T \propto (N/\mu)^2,$$

which corresponds to (1.12). So in the strongly convex case we do not expect the early stopping effect (this effect was described above as an alternative to regularization) and overfitting effect.

More stronger overfitting effect can be observed if one replace SGD with Gradient Descent (GD) [6, 109]:

$$x^{k+1} = x^k - \tilde{\gamma}\nabla\bar{f}(x^k).$$

In particular, for smooth convex empirical loss minimization problems the better rate of convergence than

$$f(\bar{x}^T) - f(x^*) \propto N^{-5/12}$$

is impossible (without additional assumptions) independently of what is T and h [109]. Remind that at the same assumptions for SGD we have $f(\bar{x}^T) - f(x^*) \propto N^{-1/2}$ if $T \propto N$. This rate is better, since $1/2 = 6/12 > 5/12$.

Despite all of this, in practice one can often meet that (1.57) with proper regularization is solved by fast deterministic algorithms, say, LBFGS or even by using high-order schemes, see Section 1.3.10. It works due to proper regularization!

1.3.9 Distributed optimization

In Section 1.3.5 we met with batch-parallelization consist in possibility to parallelize the batch calculation:

$$\frac{1}{r}\sum_{j=1}^{r}\nabla_x f(x,\xi^j).$$

If we assume that we have the number of nodes that is a division of r, then we can fully parallelize on these nodes batch calculation. But at each subsequent iteration of considered accelerated algorithm (after the batch calculation) all the nodes are required to share theirs sub-batches. In distributed optimization this is called – communication. So one iteration assume one communication. The natural question appears: does such number of communications is also optimal like the number of subsequent iterations? In general the answer is affirmative [130]. It means that without additional assumptions batched accelerated methods are the best ones in Federated Learning (FL) setup from the theoretical point of view [58]. This conclusion looks somewhat discouraging since from the practice it is well known that local steps (the main ingredient of FL) works good. To explain this contradiction let us consider unconstrained convex quadratic stochastic optimization problem. An important property of accelerated dynamics is its linearity (on average) in terms of x. This linearity generates superposition principle: instead of communication at each iteration we can to run independently at each node accelerated algorithm with reduced (to the number of nodes) batch-size and we communicate only one time at the very end (at the last iteration) by calculating an average of the outputs at all the nodes (this procedure is called «one shoot»). The total output of this approach will have the same quality as the approach we started with [129] (with many communications).

It means that for quadratic stochastic optimization problems (and close to quadratic ones) local steps give tangible benefits. Since quadratic problems are naturally appears as a local approximation of real problem in the vicinity of the solution or at each iteration as an inner problem (for example, iteration of Newton method [17]) we can still exploit local steps. One such example we consider at the very end of this section.

It is interesting to note, that rather than for deterministic distributed convex optimization problems for stochastic convex optimization problems there is a significant difference between the class of quadratic problems and convex ones [85, 130].

More naturally distributed setup appears when dealing with offline approach. For example, if we have m nodes (such that $N = m \cdot s$ for some natural s) we can rewrite the empirical loss minimization problem (1.57) as follows:

$$\min_{x\in Q}\left[\bar{f}(x) := \frac{1}{m}\sum_{k=1}^{m}\bar{f}_k(x) := \frac{1}{m}\sum_{k=1}^{m}\frac{1}{s}\sum_{j=1}^{s} f(x,\xi^{k,j})\right].$$

If we apply standard accelerated method [89] assuming that $\bar{f}(x)$ is μ-strongly convex in 2-norm and has L-Lipscitz gradient, then the number of iterations (communica-

tions) will be $\tilde{O}\left(\sqrt{L/\mu}\right)$ (here and below in this section we skip all the logarithmic for for a better visibility) and the number of incremental gradient oracle calls at each node will be $\tilde{O}\left(s\sqrt{L/\mu}\right)$ (the number of $\nabla_x f(x, \xi^{k,j})$ calculation).

Section 1.3.6 gives a hope that this bound can be further improved due to the sum-type structure of the functions stored at each node. Indeed, there exist a distributed accelerated VR scheme [72] with $\tilde{O}\left(\sqrt{L/\mu}\right)$ communication complexity and $\tilde{O}\left(s + \sqrt{sL/\mu}\right)$ oracle complexity in each node, where L in the last formula is a maximal Lipschitz gradient constant in x in 2-norm of functions $f(x, \xi^{k,j})$. This bound is optimal [52] if we do not use that $\left\{\xi^{k,j}\right\}$ are i.i.d. or do not use that among $\bar{f}_k(x)$ there exists some kind of similarity. In more details, if Lipschitz gradient constants of $\bar{f}(x) - \bar{f}_k(x)$ are bounded in 2-norm by l ($l \ll L$) than we may expect better communication complexity $\tilde{O}\left(\sqrt{l/\mu}\right)$, which corresponds to the lower bound under similarity [7].

To use similarity we describe **Accelerated gradient sliding** for unconstrained composite optimization problem:

$$\min_{x \in \mathbb{R}^n} \left[\bar{f}(x) := g(x) + h(x)\right],$$

where $g(x)$ has L_g-Lipschitz continuous gradient, $h(x)$ is convex and has L_h-Lipschitz continuous gradient ($L_g \leq L_h$); $\bar{f}(x)$ is μ-strongly convex function in 2-norm. Note that we do not assume $g(x)$ to be convex! The algorithm looks as follows [64]:

$$\tilde{x}^t = \tau x^t + (1-\tau)x_f^t,$$

$$x_f^{t+1} \approx \operatorname*{argmin}_{x \in \mathbb{R}^n} \left[A^t(x) := g(\tilde{x}^t) + \langle \nabla g(\tilde{x}^t), x - \tilde{x}^t \rangle + L_g \|x - \tilde{x}^t\|_2^2 + h(x)\right], \quad (1.60)$$

which means

$$\|\nabla A^t(x_f^{t+1})\|_2^2 \leq \frac{L_g^2}{3} \left\|\tilde{x}^t - \arg\min_{x \in \mathbb{R}^n} A^t(x)\right\|_2^2, \quad (1.61)$$

$$x^{t+1} = x^t + \eta\mu(x_f^{t+1} - x^t) - \eta\nabla\bar{f}(x_f^{t+1}),$$

where

$$\tau = \min\left\{1, \frac{\sqrt{\mu}}{2\sqrt{L_g}}\right\}, \quad \eta = \min\left\{\frac{1}{2\mu}, \frac{1}{2\sqrt{\mu L_g}}\right\}.$$

This algorithm (with output point x^N) has an iteration complexity

$$\tilde{O}\left(\sqrt{\frac{L_g}{\mu}}\right)$$

and solves several tasks at once:

- **(simple acceleration)** If $h(x) \equiv 0$ this algorithm becomes an ordinary accelerated method with
$$x_f^{t+1} = \tilde{x}^t - \frac{1}{2L_p}\nabla g(\tilde{x}^t);$$
- **(Catalyst)** If $g(x) \equiv 0$ this algorithm becomes a Catalyst-type proximal envelop [74], but less sensitive to the accuracy of the solution of (1.60);[24]
- **(Sliding)** If we apply to (1.60) Accelerated gradient sliding with $g(x) := h(x)$ and $h(x) := g(\tilde{x}^t) + \langle \nabla g(\tilde{x}^t), x - \tilde{x}^t \rangle + L_g \|x - \tilde{x}^t\|_2^2$ then obtain the total complexity of $\nabla h(x)$ oracle as
$$\tilde{O}\left(\sqrt{\frac{L_g}{\mu}}\right) \cdot O\left(\sqrt{\frac{L_g + L_h}{L_g}}\right) = \tilde{O}\left(\sqrt{\frac{L_h}{\mu}}\right).$$
That is, we have split the complexity of considered composite problem to the complexities correspond to the separate problems:
$$\tilde{O}\left(\sqrt{\frac{L_g}{\mu}}\right) \quad \text{for } \#\nabla g(x) \quad \text{and} \quad \tilde{O}\left(\sqrt{\frac{L_h}{\mu}}\right) \quad \text{for } \#\nabla h(x).$$

Let us rewrite the empirical problem as follows

$$\min_{x \in \mathbb{R}^n} \left[\bar{f}(x) := \left(\bar{f}(x) - \bar{f}_1(x)\right) + \bar{f}_1(x)\right]. \tag{1.62}$$

Denoting the first sum as $g(x)$ and the second one as $h(x)$ we can use Sliding trick to split the complexities. Note that we significantly use the fact that in this scheme $g(x)$ is not necessarily convex! So it remains only to notice that described Accelerated gradient sliding under this choice of $g(x)$ and $h(x)$ has a natural distributed interpretation.[25] It gives at the end a distributed algorithm that works according to the lower bounds for communications and oracle calls per node complexities under similarity [7]. Due to the statistical (i.i.d.) nature of $\left\{\xi^{k,j}\right\}$ (statistical similarity) one may expect that [53]: $L_g \propto s^{-1/2}$.

Thus, the number of communications for the developed algorithm is proportional to $\propto \sqrt{1/\sqrt{s}\mu}$ and the number of incremental gradient oracle calls at each node remains the same as for ordinary accelerated method $\tilde{O}\left(s\sqrt{L/\mu}\right)$. It means that we

[24] From Catalyst technique one can obtain (1.58) based on (1.59), restarts (see the proof of Theorem 1.10) and accelerated batched algorithm, see Section 1.3.5. Note also the paper [19], where the authors independently proposed stable version (to the accuracy of the solution of (1.60)) of Catalyst. Both of these versions are «logarithmic-free» (do not introduce additional logarithmic multipliers compared to direct acceleration), rather than initial one [74].

[25] Indeed, we can assign the node number 1 to be a master node that minimize at each iteration (1.60) with $g(x) := \bar{f}(x) - \bar{f}_1(x)$ and $h(x) := \bar{f}_1(x)$. It is obvious, that $h(x)$ is available to the master node and $\nabla g(\tilde{x}^t)$ can be available due to communications of the master node with the other ones. At each round of communications k-th node sends $\nabla \bar{f}_k(\tilde{x}^t)$ to the master node and receive in return x_f^{t+1}, which is calculated at the master node.

indeed improve the communication complexity by using statistical similarity. But at the same time we have worsened the oracle complexity per node in comparison with VR accelerated method, which uses sum-type structure of the terms stored in each nodes. An open question is to build an «intermediate» algorithm – some kind of convex combination of VR and Sliding with statistical similarity. The parameter of this convex combination is determined in practice by the ratio of arithmetic complexities of one oracle call to one communication.

Note that in (1.62) instead of $\bar{f}_1(x)$ we can take an arbitrary smooth convex functions. In particular we can take Taylor series expansion

$$\tilde{f}_1(x) := \bar{f}_1(\tilde{x}^t) + \langle \nabla \bar{f}_1(\tilde{x}^t), x - x^t \rangle + \frac{1}{2!} \langle \nabla^2 \bar{f}_1(\tilde{x}^t)(x - \tilde{x}^t), x - x^t \rangle.$$

Note that $\tilde{f}_1(x)$ – convex function, rather than $\bar{f}(x) - \tilde{f}_1(x)$. Under the third-order smoothness assumption one may expect that $\tilde{f}_1(x)$ has a close hessian to the hessian of $\bar{f}_1(x)$ in the vicinity of $\tilde{x}^t$. Thus we may expect this method to be required only few communication steps when the number of iteration t is large. Note that in this approach we not only have similarity on higher iterations, but also have a quadratic structure for auxiliary problem (1.60). In case of stochastic (randomized) oracle this structure allows to use accelerated one-shoot local methods for (1.60), which strength the effect of communications saving.

In this section we consider distributed centralized algorithms. Some of the results mentioned above have analogues also in decentralized setup, see [46] and references there in.

1.3.10 Accelerated tensor methods

Starting with the work [91] the interest in tensor methods (i.e. the methods that used high-order derivatives) in convex optimization began to grow steadily. In particular, an optimal[26] (up to a logarithmic complexity factor for line-search procedure) second-order method was proposed in [78] and an optimal (also, up to a logarithmic factor) high-order method was proposed in [39]. In [90] it was shown that second and third-order tensor methods are implementable – complexity of each iteration is roughly the same as for Newton method. Optimal methods without line-search (that work according to the lower bounds up to a constant factor) were recently proposed in [66, 18]. Thus the deterministic theory of tensor methods for convex (unconstrained) problems seems to be close to the final point. In Section 1.3.5 we have demonstrated the profit of acceleration in online approach for smooth problems. Fortunately, we can additionally improve the results of Section 1.3.5 by using accelerated tensor methods. For that we need to develop sensitivity analysis of these methods. Such an analysis was made in [1] for accelerated tensor methods according to Nesterov-type of acceleration under high-order smoothness assumption [90]. This acceleration is

[26] See [63, 38] and references there in for lower bounds.

a little bit worse than the best one Monteiro–Svaiter acceleration [78, 39, 66]. By using the results of [1] and batching technique one can improve the number of subsequent iterations in online approach from Section 1.3.5. If n is not too big then such improvement can be valuable also in terms of arithmetic complexity.

For offline approach the main motivation to use tensor methods is coming from the similarity approach, see Section 1.3.9. Where the reduced auxiliary problem (1.60)

$$\min_{x\in\mathbb{R}^n}\left[\langle\nabla\bar{f}(\tilde{x}^t)-\nabla\bar{f}_1(\tilde{x}^t),x-\tilde{x}^t\rangle+l\|x-\tilde{x}^t\|_2^2+\frac{1}{s}\sum_{j=1}^{s}f(x,\xi^{1,j})\right]$$

is a sum type problem with the reduced number of terms s ($s \ll N$). If $s \simeq n$ we have that for Newton-type methods the complexity of one iteration is upper bounded by the Hessian-matrix inversion, rather than the complexity of Hessian calculation by itself. In other words, in this case second and third-order tensor methods do not «feel» the sum-type structure of the problem and work with almost the same complexities as if $s = 1$. This idea reduces the number of subsequent iterations of second and third-order methods for inner (auxiliary) problems and simultaneously alleviates the main drawback of tensor methods related with expansive iterations [33].[27]

1.3.11 Saddle-point problems and variational inequalities

Offline approach to the stochastic Saddle-point problems (SPP) developed in [71, 137, 30, 93], see also [65] for distributed approach. Online approach to the stochastic Variational Inequalities (VI) – and as a consequence for saddle-point problems – developed in [56, 42, 43].

Roughly speaking, all the results for both of the approaches look very similar to the results mentioned in the previous sections for the stochastic optimization problems except absence of acceleration. But there still exist open problems for SPP and VI that were closed for optimization problems. For example, randomized VR algorithms for (strongly) convex problems match the lower complexity bound (see Section 1.3.6), rather than its SPP and VI analogues [4, 48].

1.3.12 Wasserstein barycenter example

Wasserstein barycenter (WB) problem and its dual entropy-smoothing version is an extremely interesting example in many ways at once. First of all, stochastic

[27] Since we have to calculate the sum the iteration must be expensive independently of the order of the method we use. This observation opens up the possibility to increase the order of the method by conserving the complexity of iteration.

optimization (population) WB problem formulation comes from Statistics, but is not due to the principle of maximum likelihood [14, 12]. So we may consider this example to be intermediate in terms of Section 1.1.1 and Section 1.1.2. Secondly, the empirical WB problem as a convex optimization problem has an efficient saddle-point and dual representations [27]. For example, when WB problem solved on the space of finite-support measures (on n points) the complexity of primal gradient oracle is $\tilde{O}\left(n^3\right)$ a.o. (arithmetic operations) and the complexity of dual gradient oracle is $O\left(n^2\right)$ a.o. Moreover, dual gradient oracle has a natural stochastic unbiased estimation with the complexity $O\left(n\right)$ a.o. For some real-world applications $n \simeq 10^6$. Hence mentioned above computational observations play an important role [27]. Thirdly, the possibility to use dual oracle appears only in offline approach. To make this approach correct we need proper regularization [28], see also Theorem 1.8 for the Euclidean case. This regularization should be non-Euclidean, since we have simplex constraint – barycenter is a measure, that is an element of probability simplex $S_n(1)$. Fourthly, WB problem is non-smooth, but strongly convex in 2-norm on $S_n(1)$ if we consider dual entropy-smoothing version [27]. Since the problem is non-smooth it is impossible to use batch-parallelization in online approach, see Section 1.3.5. But due to the strong convexity (comes from regularization or/and from dual entropy-smoothing) the dual problem (in offline approach) is smooth [104] and we can apply distributed (batched-parallelized) accelerated methods to solve it [27]. To conclude, WB problem is an interesting example of the problem for which offline approach motivated not only the ability to distribute calculations across nodes (what is typical of the offline approach in general), but also the possibility to solve dual problem with better properties: cheaper oracle and better iteration-complexities bounds, since smoothness without strong convexity (for dual problem) is better than strong convexity without smoothness (for primal problem).

At the end we mentioned that the empirical WB problem is not well suited for modern distributed Variance Reduced (VR) schemes and algorithms that use similarity. The reason is a simplex constraint. Although for Euclidean proximal setup distributed VR is well developed [65] as well as similarity [64], but for non-Euclidean proximal setup (generated by the simplex constraint) the results are absent.

With this remark, we wanted to demonstrate that despite the huge progress made in the last decade in convex stochastic programming, there are still a lot of open problems that looks like a minor generalization of already solved ones. Apparently, solving such problems will require the involvement of new ideas.

1.4 Historical Notes

Stochastic optimization has began to take shape in an independent field of knowledge for about 70 years ago starting with the seminal paper of H. Robbins and S. Monro [103]. This field was actively developed along with the usual optimization. In particular, in an outstanding book of A.S. Nemirovski and D.B. Yudin [85] (original version of the book was dated by 1979) the complexity theory of mod-

ern convex optimization was build. This theory included stochastic gradient oracle. So we may consider 1979 as a second (theoretical) birth of stochastic optimization. The third significant wave of the interest happened for about 20 years ago in accordance with Data Science applications. It is already impossible to imagine modern data analysis without stochastic optimization. For the moment many books were written around Stochastic optimization [35, 13, 108, 100, 116, 114]. In some books and surveys one can find Data Science applications of Stochastic Optimization [80, 120, 121, 110, 101, 134, 25, 15, 8, 135].

The results of Section 1.1.1 are rather standard and can be mainly find in [54, 119]. An example of Vadim V. Mottl was taken from [68]. Non-asymptotic results can be found in [117, 118]. Polyak–Juditsky–Ruppert averaging was separately proposed in [106] and [96, 97]. Online analogue of Theorem 1.1 was developed in [98, 99].

The results of Section 1.1.2 were motivated by the papers [59, 122, 111, 112]. Online learning is well presented in [20, 51, 92, 21]. Note that for the convex case (not strongly convex) the described results can be generalized to non-euclidean proximal setup. The most interesting applications related with unit simplex $Q = S_n(1)$ [20]. Note that in this section we started to use the notion of (unbiased) stochastic subgradient $\nabla_x f(x, \xi)$ without accurate definition of this subject in the non-smooth case. The problems appear when the subgradient is not unique. In this case we understand under $\nabla_x f(x, \xi)$ some kind of measurable selector (no matter what kind of selector). More accurate definitions and properties of stochastic gradient one can find in [114].

The results of Section 1.2.1 were taken from [115, 83, 114]. The tight lower bound for online case was obtained in [2]. The tight lower bound for offline case (for smooth convex problems) was obtained in [36].

Online results of Section 1.2.2 corresponds to [57]. Offline results of Section 1.2.2 corresponds to [111, 110]. High-probability bounds investigated in [37, 62]. Tikhonov's regularization was accurately developed in [127]. For non-euclidean case offline results were generalized in [28, 30].

Online results of Section 1.2.3 were taken from [115, 114]. Offline results of Section 1.2.3 were taken from [57] for the case $s = 2$ (s is growth parameter). For the case $s = 1$ (sharp minimum [94]) this result was obtained earlier in a different manner [55]. The idea of restarts for strongly convex problems goes back to [85, 84]. For the stochastic optimization problems it was developed in [57]. For a sharp minimum and deterministic optimization convex optimization problems restarts was developed in [105].

References

1. Artem Agafonov, Dmitry Kamzolov, Pavel Dvurechensky, Alexander Gasnikov, and Martin Takac. Inexact tensor methods and their application to stochastic convex optimization, 2020.
2. Alekh Agarwal, Peter L. Bartlett, Pradeep Ravikumar, and Martin J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012.

3. Ahmad Ajalloeian and Sebastian U. Stich. On the convergence of sgd with biased gradients. *arXiv preprint arXiv:2008.00051*, 2020.
4. Ahmet Alacaoglu and Yura Malitsky. Stochastic variance reduction for variational inequality methods. *arXiv preprint arXiv:2102.08352*, 2021.
5. Zeyuan Allen-Zhu and Elad Hazan. Optimal black-box reductions between optimization objectives. *Advances in Neural Information Processing Systems*, 29, 2016.
6. Idan Amir, Yair Carmon, Tomer Koren, and Roi Livni. Never go full batch (in stochastic convex optimization). In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 25033–25043. Curran Associates, Inc., 2021.
7. Yossi Arjevani and Ohad Shamir. Communication complexity of distributed convex learning and optimization. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.
8. Francis Bach. Learning theory from first principles. *e-print,* `https://www.di.ens.fr/~fbach/ltfp_book.pdf`, 2021.
9. Francis Bach and Kfir Y. Levy. A universal algorithm for variational inequalities adaptive to smoothness and noise. In *Conference on learning theory*, pages 164–194. PMLR, 2019.
10. Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.
11. Mikhail Belkin. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, 30:203–248, 2021.
12. Jérémie Bigot and Thierry Klein. Characterization of barycenters in the wasserstein space by averaging optimal transport maps. *ESAIM: Probability and Statistics*, 22:35–57, 2018.
13. John R. Birge and Francois Louveaux. *Introduction to stochastic programming*. Springer Science & Business Media, 2011.
14. Emmanuel Boissard, Thibaut Le Gouic, and Jean-Michel Loubes. Distribution's template estimate with wasserstein metrics. *Bernoulli*, 21(2):740–759, 2015.
15. Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.
16. Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2(Mar):499–526, 2002.
17. Brian Bullins, Kshitij Patel, Ohad Shamir, Nathan Srebro, and Blake E. Woodworth. A stochastic newton algorithm for distributed convex optimization. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 26818–26830. Curran Associates, Inc., 2021.
18. Yair Carmon, Danielle Hausler, Arun Jambulapati, Yujia Jin, and Aaron Sidford. Optimal and adaptive monteiro-svaiter acceleration. *arXiv preprint arXiv:2205.15371*, 2022.
19. Yair Carmon, Arun Jambulapati, Yujia Jin, and Aaron Sidford. Recapp: Crafting a more efficient catalyst for convex optimization. *arXiv preprint arXiv:2206.08627*, 2022.
20. Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.
21. Nicolo Cesa-Bianchi and Francesco Orabona. Online learning algorithms. *Annual review of statistics and its application*, 2021.
22. Olivier Devolder. *Exactness, inexactness and stochasticity in first-order methods for large-scale convex optimization*. PhD thesis, PhD thesis, 2013.
23. Benjamin Dubois-Taine, Francis Bach, Quentin Berthet, and Adrien Taylor. Fast stochastic composite minimization and an accelerated frank-wolfe algorithm under parallelization. *arXiv preprint arXiv:2205.12751*, 2022.
24. John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.
25. John C. Duchi. Introductory lectures on stochastic optimization. *The mathematics of data*, 25:99–186, 2018.
26. John C. Duchi, Shai Shalev-Shwartz, Yoram Singer, and Ambuj Tewari. Composite objective mirror descent. In *COLT*, volume 10, pages 14–26. Citeseer, 2010.

27. Darina Dvinskikh. Decentralized algorithms for wasserstein barycenters. *arXiv preprint arXiv:2105.01587*, 2021.
28. Darina Dvinskikh. Stochastic approximation versus sample average approximation for wasserstein barycenters. *Optimization Methods and Software*, pages 1–33, 2021.
29. Darina Dvinskikh and Alexander Gasnikov. Decentralized and parallel primal and dual accelerated methods for stochastic convex programming problems. *Journal of Inverse and Ill-posed Problems*, 29(3):385–405, 2021.
30. Darina Dvinskikh, Vitali Pirau, and Alexander Gasnikov. On the relations of stochastic convex optimization problems with empirical risk minimization problems on p-norm balls. *arXiv preprint arXiv:2202.01805*, 2022.
31. Darina Dvinskikh, Alexander Tyurin, Alexander Gasnikov, and Sergey Omelchenko. Accelerated and nonaccelerated stochastic gradient descent with model conception. *Math. Notes*, 108(4):511–522, 2020.
32. Pavel Dvurechensky and Alexander Gasnikov. Stochastic intermediate gradient method for convex problems with stochastic inexact oracle. *Journal of Optimization Theory and Applications*, 171(1):121–145, 2016.
33. Pavel Dvurechensky, Dmitry Kamzolov, Aleksandr Lukashevich, Soomin Lee, Erik Ordentlich, Cesar A Uribe, and Alexander Gasnikov. Hyperfast second-order local solvers for efficient statistically preconditioned distributed optimization. *arXiv preprint arXiv:2102.08246*, 2021.
34. Alina Ene, Huy L. Nguyen, and Adrian Vladu. Adaptive gradient methods for constrained convex optimization and variational inequalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7314–7321, 2021.
35. Yu.M. Ermoliev and R.J.-B. Wets. *Numerical techniques for stochastic optimization*. Springer-Verlag, 1988.
36. Vitaly Feldman. Generalization of erm in stochastic convex optimization: The dimension strikes back. *Advances in Neural Information Processing Systems*, 29:3576–3584, 2016.
37. Vitaly Feldman and Jan Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on Learning Theory*, pages 1270–1279. PMLR, 2019.
38. Ankit Garg, Robin Kothari, Praneeth Netrapalli, and Suhail Sherif. Near-optimal lower bounds for convex optimization for all orders of smoothness. *Advances in Neural Information Processing Systems*, 34:29874–29884, 2021.
39. Alexander Gasnikov, Pavel Dvurechensky, Eduard Gorbunov, Evgeniya Vorontsova, Daniil Selikhanovych, César A Uribe, Bo Jiang, Haoyue Wang, Shuzhong Zhang, Sébastien Bubeck, et al. Near optimal methods for minimizing convex functions with lipschitz p-th derivatives. In *Conference on Learning Theory*, pages 1392–1393. PMLR, 2019.
40. Alexander Gasnikov, Anton Novitskii, Vasilii Novitskii, Farshed Abdukhakimov, Dmitry Kamzolov, Aleksandr Beznosikov, Martin Takáč, Pavel Dvurechensky, and Bin Gu. The power of first-order smooth optimization for black-box non-smooth problems. *arXiv preprint arXiv:2201.12289*, 2022.
41. Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.
42. Eduard Gorbunov, Hugo Berard, Gauthier Gidel, and Nicolas Loizou. Stochastic extragradient: General analysis and improved rates. In *International Conference on Artificial Intelligence and Statistics*, pages 7865–7901. PMLR, 2022.
43. Eduard Gorbunov, Marina Danilova, David Dobre, Pavel Dvurechensky, Alexander Gasnikov, and Gauthier Gidel. Clipped stochastic methods for variational inequalities with heavy-tailed noise. *arXiv preprint arXiv:2206.01095*, 2022.
44. Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. *Advances in Neural Information Processing Systems*, 33:15042–15053, 2020.
45. Eduard Gorbunov, Marina Danilova, Innokentiy Shibaev, Pavel Dvurechensky, and Alexander Gasnikov. Near-optimal high probability complexity bounds for non-smooth stochastic optimization with heavy-tailed noise. *arXiv preprint arXiv:2106.05958*, 2021.

46. Eduard Gorbunov, Alexander Rogozin, Aleksandr Beznosikov, Darina Dvinskikh, and Alexander Gasnikov. Recent theoretical advances in decentralized distributed convex optimization. *arXiv preprint arXiv:2011.13259*, 2020.
47. Eduard Gorbunov, Evgeniya Vorontsova, and Alexander Gasnikov. On the upper bound for the expectation of the norm of a vector uniformly distributed on the sphere and the phenomenon of concentration of uniform measure on the sphere. *Mathematical Notes*, 106, 2019.
48. Yuze Han, Guangzeng Xie, and Zhihua Zhang. Lower complexity bounds of finite-sum optimization problems: The results and construction. *arXiv preprint arXiv:2103.08280*, 2021.
49. Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1225–1234, New York, New York, USA, 20–22 Jun 2016. PMLR.
50. Elad Hazan. Lecture notes: Optimization for machine learning. *arXiv preprint arXiv:1909.03550*, 2019.
51. Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.
52. Hadrien Hendrikx, Francis Bach, and Laurent Massoulié. An optimal algorithm for decentralized finite-sum optimization. *SIAM Journal on Optimization*, 31(4):2753–2783, 2021.
53. Hadrien Hendrikx, Lin Xiao, Sebastien Bubeck, Francis Bach, and Laurent Massoulie. Statistically preconditioned accelerated gradient method for distributed optimization. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4203–4227. PMLR, 13–18 Jul 2020.
54. Ildar Abdulovich Ibragimov and Rafail Zalmanovich HasMinskii. *Statistical estimation: asymptotic theory*, volume 16. Springer Science & Business Media, 2013.
55. Anatoli Juditsky. A stochastic estimation algorithm with observation averaging. *IEEE transactions on automatic control*, 38(5):794–798, 1993.
56. Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.
57. Anatoli Juditsky and Yuri Nesterov. Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization. *Stochastic Systems*, 4(1):44–80, 2014.
58. Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konecný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Hang Qi, Daniel Ramage, Ramesh Raskar, Mariana Raykova, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.
59. Sham M. Kakade and Ambuj Tewari. On the generalization ability of online strongly convex programming algorithms. *Advances in Neural Information Processing Systems*, 21, 2008.
60. Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak–łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 795–811. Springer, 2016.
61. Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
62. Yegor Klochkov and Nikita Zhivotovskiy. Stability and deviation optimal risk bounds with convergence rate $\mathcal{O}(1/n)$. *arXiv preprint arXiv:2103.12024*, 2021.

63. Guy Kornowski and Ohad Shamir. High-order oracle complexity of smooth and strongly convex optimization. *arXiv preprint arXiv:2010.06642*, 2020.
64. Dmitry Kovalev, Aleksandr Beznosikov, Ekaterina Borodich, Alexander Gasnikov, and Gesualdo Scutari. Optimal gradient sliding and its application to distributed optimization under similarity. *arXiv preprint arXiv:2205.15136*, 2022.
65. Dmitry Kovalev, Aleksandr Beznosikov, Abdurakhmon Sadiev, Michael Persiianov, Peter Richtárik, and Alexander Gasnikov. Optimal algorithms for decentralized stochastic variational inequalities. *arXiv preprint arXiv:2202.02771*, 2022.
66. Dmitry Kovalev and Alexander Gasnikov. The first optimal acceleration of high-order methods in smooth convex optimization. *arXiv preprint arXiv:2205.09647*, 2022.
67. Dmitry Kovalev, Samuel Horváth, and Peter Richtárik. Don't jump through hoops and remove those loops: Svrg and katyusha are better without the outer loop. In Aryeh Kontorovich and Gergely Neu, editors, *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, volume 117 of *Proceedings of Machine Learning Research*, pages 451–467. PMLR, 08 Feb–11 Feb 2020.
68. Olga Krasotkina and Vadim Mottl. A bayesian approach to sparse learning-to-rank for search engine optimization. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, pages 382–394. Springer, 2015.
69. Guanghui Lan. *First-order and stochastic optimization methods for machine learning*. Springer, 2020.
70. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.
71. Yunwen Lei, Zhenhuan Yang, Tianbao Yang, and Yiming Ying. Stability and generalization of stochastic gradient methods for minimax problems. In *International Conference on Machine Learning*, pages 6175–6186. PMLR, 2021.
72. Huan Li, Zhouchen Lin, and Yongchun Fang. Variance reduced extra and diging and their optimal acceleration for strongly convex decentralized optimization, 2020.
73. Shaojie Li and Yong Liu. Improved learning rates for stochastic optimization: Two theoretical viewpoints. *arXiv preprint arXiv:2107.08686*, 2021.
74. Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A universal catalyst for first-order optimization. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.
75. Junhong Lin, Raffaello Camoriano, and Lorenzo Rosasco. Generalization properties and implicit regularization for multiple passes sgm. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2340–2348, New York, New York, USA, 20–22 Jun 2016. PMLR.
76. Zhouchen Lin, Huan Li, and Cong Fang. Accelerated optimization for machine learning. *Nature Singapore: Springer*, 2020.
77. Dmitriy Metelev, Alexander Rogozin, Alexander Gasnikov, and Dmitry Kovalev. Decentralized saddle-point problems with different constants of strong convexity and strong concavity. *arXiv preprint arXiv:2206.00090*, 2022.
78. Renato DC Monteiro and Benar Fux Svaiter. An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods. *SIAM Journal on Optimization*, 23(2):1092–1125, 2013.
79. Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24, 2011.
80. Fomin Vladimir N. *Recurrent estimation and adaptive filtration*. M.: Nauka (in Russian), 1984.
81. Alexander V. Nazin, Arkadi S. Nemirovsky, Alexandre B. Tsybakov, and Anatoli B. Juditsky. Algorithms of robust stochastic optimization based on mirror descent method. *Automation and Remote Control*, 80(9):1607–1627, 2019.
82. Ion. Necoara, Yu. Nesterov, and Francois Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 175(1):69–107, 2019.

83. Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.
84. Arkadi Semenovich Nemirovski and Yurii Evgenievich Nesterov. Optimal methods of smooth convex minimization. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 25(3):356–369, 1985.
85. A.S. Nemirovski and D.B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. A Wiley-Interscience publication. Wiley, 1983.
86. Yu. Nesterov. Gradient methods for minimizing composite functions. *Mathematical programming*, 140(1):125–161, 2013.
87. Yu. Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1):381–404, 2015.
88. Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1):221–259, 2009.
89. Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.
90. Yurii Nesterov. Implementable tensor methods in unconstrained convex optimization. *Mathematical Programming*, 186(1):157–183, 2021.
91. Yurii Nesterov and Boris T Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.
92. Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.
93. Asuman Ozdaglar, Sarath Pattathil, Jiawei Zhang, and Kaiqing Zhang. What is a good metric to study generalization of minimax learners? *arXiv preprint arXiv:2206.04502*, 2022.
94. Boris Teodorovich Poljak. Sharp minimum. In *Generalized Lagrangians and applications*. Oxford: Pergamon Press, 1982.
95. Boris Teodorovich Polyak. Introduction to optimization. optimization software. *Inc., Publications Division, New York*, 1, 1987.
96. Boris Teodorovich Polyak. New method of stochastic approximation type. *Automation and remote control*, 51(7):937–946, 1990.
97. Boris Teodorovich Polyak and Anatoli Borisovich Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.
98. Boris Teodorovich Polyak and Yakov Zalmanovich Tsypkin. Adaptive estimation algorithms: convergence, optimality, stability. *Avtomatika i telemekhanika*, (3):71–84, 1979.
99. Boris Teodorovich Polyak and Yakov Zalmanovich Tsypkin. Optimal pseudogradient adaptation algorithms. *Avtomatika i Telemekhanika*, (8):74–84, 1980.
100. András Prékopa. *Stochastic programming*, volume 324. Springer Science & Business Media, 2013.
101. Alexander Rakhlin and Karthik Sridharan. Statistical learning and sequential prediction. *e-print,* `http://www.mit.edu/~rakhlin/courses/stat928/stat928_notes.pdf`, 2014.
102. Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019.
103. Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.
104. R. Tyrrell Rockafellar. *Convex analysis*, volume 18. Princeton university press, 1970.
105. Vincent Roulet and Alexandre d'Aspremont. Sharpness, restart and acceleration. *Advances in Neural Information Processing Systems*, 30, 2017.
106. David Ruppert. Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.
107. Abdurakhmon Sadiev, Darina Dvinskikh, Aleksandr Beznosikov, and Alexander Gasnikov. Decentralized and personalized federated learning. *arXiv preprint arXiv:2107.07190*, 2021.
108. Johannes Schneider and Scott Kirkpatrick. *Stochastic optimization*. Springer Science & Business Media, 2007.

109. Ayush Sekhari, Karthik Sridharan, and Satyen Kale. Sgd: The role of implicit regularization, batch-size and multiple-epochs. *Advances in Neural Information Processing Systems*, 34, 2021.
110. Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.
111. Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *COLT*, volume 2, page 5, 2009.
112. Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.
113. Ohad Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *Journal of Machine Learning Research*, 18:52:1–52:11, 2017. First appeared in arXiv:1507.08752.
114. Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczynski. *Lectures on stochastic programming: modeling and theory*. SIAM, 2021.
115. Alexander Shapiro and Arkadi Nemirovski. On complexity of stochastic programming problems. In *Continuous optimization*, pages 111–146. Springer, 2005.
116. James C. Spall. Stochastic optimization. In *Handbook of computational statistics*, pages 173–201. Springer, 2012.
117. Vladimir Spokoiny. Parametric estimation. finite sample theory. *The Annals of Statistics*, 40(6):2877–2909, 2012.
118. Vladimir Spokoiny. Bernstein-von mises theorem for growing parameter dimension. *arXiv preprint arXiv:1302.3430*, 2013.
119. Vladimir Spokoiny and Thorsten Dickhaus. *Basics of modern mathematical statistics*. Springer, 2015.
120. Suvrit Sra, Sebastian Nowozin, and Stephen J. Wright. *Optimization for machine learning*. Mit Press, 2012.
121. Karthik Sridharan. Learning from an optimization viewpoint. *arXiv preprint arXiv:1204.4145*, 2012.
122. Karthik Sridharan, S Shalev-Shwartz, and N Srebro. Fast convergence rates for excess regularized risk with application to svm, 2008.
123. Eli Stevens, Luca Antiga, and Thomas Viehmann. *Deep learning with PyTorch*. Manning Publications, 2020.
124. Sebastian U Stich. Unified optimal analysis of the (stochastic) gradient method. *arXiv preprint arXiv:1907.04232*, 2019.
125. Fedor Stonyakin, Alexander Tyurin, Alexander Gasnikov, Pavel Dvurechensky, Artem Agafonov, Darina Dvinskikh, Mohammad Alkousa, Dmitry Pasechnyuk, Sergei Artamonov, and Victorya Piskunova. Inexact model: A framework for optimization and variational inequalities. *Optimization Methods and Software*, pages 1–47, 2021.
126. Adrien B Taylor, Julien M Hendrickx, and François Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161(1):307–345, 2017.
127. Andrey Nikolaevich Tikhonov and Vasilii Iakovlevich Arsenin. *Solutions of Ill-posed Problems: Andrey N. Tikhonov and Vasiliy Y. Arsenin. Translation Editor Fritz John*. Wiley, 1977.
128. Nuri Mert Vural, Lu Yu, Krishnakumar Balasubramanian, Stanislav Volgushev, and Murat A Erdogdu. Mirror descent strikes again: Optimal stochastic convex optimization under infinite noise variance. *arXiv preprint arXiv:2202.11632*, 2022.
129. Blake Woodworth, Kumar Kshitij Patel, Sebastian Stich, Zhen Dai, Brian Bullins, Brendan Mcmahan, Ohad Shamir, and Nathan Srebro. Is local SGD better than minibatch SGD? In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10334–10343. PMLR, 13–18 Jul 2020.

130. Blake E. Woodworth, Brian Bullins, Ohad Shamir, and Nathan Srebro. The min-max complexity of distributed stochastic convex optimization with intermittent communication. In *Conference on Learning Theory*, pages 4386–4437. PMLR, 2021.
131. Blake E Woodworth and Nathan Srebro. An even more optimal stochastic optimization algorithm: Minibatching and interpolation learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 7333–7345. Curran Associates, Inc., 2021.
132. Blake E. Woodworth and Nati Srebro. Tight complexity bounds for optimizing composite objectives. *Advances in neural information processing systems*, 29, 2016.
133. Blake E. Woodworth, Jialei Wang, Adam Smith, Brendan McMahan, and Nati Srebro. Graph oracle models, lower bounds, and gaps for parallel stochastic optimization. *Advances in neural information processing systems*, 31, 2018.
134. Stephen J. Wright. Optimization algorithms for data science. *IAS/Park City Math. Ser.–2016.–* `http://www.optimization-online.org/DB_FILE/2016/12/5748.pdf`, 7:816–824, 2016.
135. Stephen J. Wright and Benjamin Recht. *Optimization for data analysis*. Cambridge University Press, 2022.
136. Lin Xiao. Dual averaging method for regularized stochastic learning and online optimization. *Advances in Neural Information Processing Systems*, 22, 2009.
137. Junyu Zhang, Mingyi Hong, Mengdi Wang, and Shuzhong Zhang. Generalization bounds for stochastic saddle point problems. In *International Conference on Artificial Intelligence and Statistics*, pages 568–576. PMLR, 2021.

Chapter 2
Stochastic Gradient Descent: Nonsmooth Case

Abstract This chapter .
In this chapter, we briefly describe

2.1 Stochastic Subgradient Method

In this section we consider the problem

$$\min_{x \in Q} \left\{ f(x) = \mathbb{E}_{\xi \sim \mathcal{D}} [f(x, \xi)] \right\}, \tag{2.1}$$

where the set Q is closed and convex.

2.1.1 Stochastic Subgradient Method

Plan:
Introduce assumptions on the objective: in each iteration k, the algorithm uses a random vector g^k called *stochastic subgradient* at the current point x^k.

$$\mathbb{E}_k [g^k] = \nabla f(x^k) \in \partial f(x^k), \tag{2.2}$$

where $\mathbb{E}_k [\cdot]$ denotes an expectation with respect to the randomness from the k-th iteration.

Assumption 2.1 (Uniformly Bounded Variance) There exists a constant $\sigma \geq 0$ such that for any $k \geq 0$ stochastic subgradient g^k is unbiased, i.e., (2.2) holds, and satisfies

$$\mathbb{E}_k \left[\|g^k - \nabla f(x^k)\|_2^2 \right] \leq \sigma^2. \tag{2.3}$$

(this assumption is made globally)

The main assumption of this section is that the objective function f has bounded subgradients such that

$$\|\nabla f(x^k)\|_2 \leq M.$$

Give equivalence to the Lipschitz continuity of f.

Give the Projected Stochastic Subgradient Method. See https://cpn-us-w2.wpmucdn.com/sites.gatech.edu/dist/f/330/files/2019/08/LectureOPTML.pdf (3.1.4) (the whole equations should be in our case too).

Prove the convergence rate in expectation in the convex case following the steps of the proof of Theorem 3.1 in https://cpn-us-w2.wpmucdn.com/sites.gatech.edu/dist/f/330/files/2019/08/LectureOPTML.pdf, but adopting it to the stochastic case as in the proof of Theorem 4.1 in the same book. After obtaining the result similar to Theorem 3.1/Theorem 4.1. do not assume that the set is bounded. Instead, assume that there is a known upper bound on the distance between x^ℓ and x^*. Keep the possibility to start averaging from some number not equal to 1. upper limit is N, lower limit is ℓ. Give the results for both types of the stepsizes as in Corollary 3.1 and Corollary 3.2 in the above book.

Prove the convergence rate in expectation in the strongly convex case following the steps of the proof of Theorem 3.2 in https://cpn-us-w2.wpmucdn.com/sites.gatech.edu/dist/f/330/files/2019/08/LectureOPTML.pdf, but adopting it to the stochastic case.

Remark on Lipschitz property on a ball since in expectation the distance will not be twice larger than at the starting point.

2.1.2 Lower bounds

Plan:

One-dimensional convex case. Take the example https://www.mccme.ru/dubna/2017/notes/gasnikov-slides4.pdf, slide 16, change the set to $[-R, R]$, the function to $M\varepsilon x$, and the distribution to $\mathcal{N}(0, \sigma^2)$ (confidence level in our book is β), write a more detailed explanations, in particular, estimate the second moment of the stochastic gradient, derive the Likelihood-ratio test, translate the bound to the language of the objective residual (to optimize below accuracy $|\varepsilon|$, we need to know exactly the sign of ε). May be helpful: https://en.wikipedia.org/wiki/Neyman%E2%80%93Pearson_lemma

One-dimensional strongly convex case. Take the example https://www.mccme.ru/dubna/2017/notes/gasnikov-slides4.pdf, slide 17, multiply the function by strong convexity parameter μ, change the distribution to $\mathcal{N}(0, \sigma^2)$ (confidence level in our book is β), write a more detailed explanations, in particular, estimate the second moment of the stochastic gradient assuming that $|x^*| \leq R$ and the unconstrained minimization problem is equivalent to minimization over a ball with center at 0 and radius $2R$, add more details to the derivations, underline that we have bounds both for the distance to the solution and for the objective residual. Underline

also that smoothness does not help, i.e. the bound is achieved on a nice smooth (even quadratic) function.

Multidimensional case will be in the section on MD with the bounds in general geometries. Cite the book by Duchi

2.2 Stochastic Composite Mirror Descent

In this section we consider composite/regularized optimization problems:

$$\min_{x \in Q \subseteq \mathbb{R}^n} \{f_h(x) = f(x) + h(x)\}, \tag{2.4}$$

where, as before, the function $f(x) : \mathbb{R}^n \to \mathbb{R}$ is given as $f(x) = \mathbb{E}_{\xi \sim \mathcal{D}}[f(x, \xi)]$, the set Q is closed convex, and function $h(x) : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a proper lower semicontinuous convex function called regularization/composite term. Moreover, function h is assumed to have simple structure. The exact meaning will be explained below.

2.2.1 Stochastic Composite Mirror Descent

Plan:

Introduce the proximal setup. $\mathbb{R}^n$, norm, dual norm, distance generating function d (be accurate and take it from `https://cpn-us-w2.wpmucdn.com/sites.gatech.edu/dist/f/330/files/2019/08/LectureOPTML.pdf` Sect 3.2 Mirror descent), Bregman divergence $V(x, y) = d(x) - d(y) - < d(y), x - y >$.

Give examples: Euclidean proximal setup, Entropy setup, give references to A. Nemirovski's lectures for more proximal setups.

Introduce assumptions on the objective: in each iteration k, the algorithm uses a random vector g^k called *stochastic gradient* at the current point x^k.

$$\mathbb{E}_k[g^k] = \nabla f(x^k) \in \partial f(x^k), \tag{2.5}$$

where $\mathbb{E}_k[\cdot]$ denotes an expectation with respect to the randomness from the k-th iteration. (can just shortly say an refer to the section 2.1)

Assumption 2.2 (Uniformly Bounded Variance) There exists a constant $\sigma \geq 0$ such that for any $k \geq 0$ stochastic gradient g^k is unbiased, i.e., (2.2) holds, and satisfies

$$\mathbb{E}_k \left[\|g^k - \nabla f(x^k)\|_*^2 \right] \leq \sigma^2. \tag{2.6}$$

We underline that in this section, this condition is considered w.r.t. the general norm.

The main assumption of this section is that the objective function f has bounded subgradients such that

$$\|\nabla f(x^k)\|_* \leq M.$$

We underline that in this section, this condition is considered w.r.t. the general norm.

Give the assumption on the simplicity of the function h meaning that a subproblem in composite mirror descent can be solved exactly or efficiently up to machine precision. The Bregman proximal operator associated with αh for any $\alpha \geq 0$ and V

$$\mathcal{P}^V_{\alpha h}(x; g) = \operatorname*{argmin}_{u \in Q} \{\alpha h(u) + \langle g, u - x\rangle + V(u, x)\} \tag{2.7}$$

should be easy to compute for any $x \in Q$.

Give the Stochastic Composite Mirror Descent algorithm. Take `https://www.dropbox.com/s/os6637d2zdvqzsf/Lecture%208%202021-01-08.pdf?dl=0` p. 5 and change the subgradient to the stochastic subgradient.

Throughout the book we use the following important result. State and prove the main technical lemma, see Lemma 1 on p.1 `https://wias-berlin.de/people/dvureche/HU%202019-2010%20OPT/AGD.pdf` (ψ - arbitrary convex function, better use x_+ instead of u_+)

Prove the convergence rate in expectation in the convex case combining the proof in the same file (Keep the composite in the resulting bound) and the proof in `https://cpn-us-w2.wpmucdn.com/sites.gatech.edu/dist/f/330/files/2019/08/LectureOPTML.pdf` Sect. 4.1.1 General nonsmooth convex functions. Important - do not assume that the feasible set is bounded, rather assume that there is an upper bound for $V(x^k, x^*)$. Give some remarks for particular cases: $h = 0$ or $h \geq 0$ and $h(x^1) = 0$. Then the bound is the same as for the stochastic subgradient method.

Give an example comparing the estimates in two setups - 2-setup and 1-setup, see pp. 1-2 `https://www.dropbox.com/s/txknghmvwxvn98q/Lecture%209%202021-01-15.pdf?dl=0` and Example 3.1 in `https://cpn-us-w2.wpmucdn.com/sites.gatech.edu/dist/f/330/files/2019/08/LectureOPTML.pdf`

Give the restarted mirror descent. (some thinking should be made or make as discussed in the video of the corresponding lecture) Questions: take strongly convex or uniformly convex case? If the second, which definition to take? `https://pubsonline.informs.org/doi/pdf/10.1287/10-SSY010` (2.1)? Growth condition as in Sect. 1.2.3? If the growth condition is sufficient, it is better to take it. Should we change the squared norm in the growth condition to the Bregman divergence? Most probably, no. Assumption on the distance generating function d - an appropriate modification of `https://arxiv.org/pdf/1710.06612.pdf` around (41). underline that the growth condition is sufficient to assume for the whole f_h. Thus, e.g., only the composite may be strongly convex.

Prove the convergence rate in expectation for the restarted algorithm. As a hint may use `https://arxiv.org/pdf/1411.2876.pdf` Sect. 4.1 or `https://arxiv.org/pdf/1710.06612.pdf` Sect. 4.3 (the second option may be used only as a hint since it uses a predefined stepsize based on accuracy ε. we need to avoid this and first consider general stepsizes.)

Questions to think about: inexact prox, biased stoch gradient

2.2.2 Composite Stochastic Dual Averaging

https://papers.nips.cc/paper/2009/file/7cce53cf90577442771720a370c3c723-Paper.pdf (plan to be developed. But just algorithm, remarks and convergence rates without the proof) Recent paper Juditskii Moulines

2.2.3 Lower bounds

Mirror Descent and Dual Averaging methods described in previous sections are of the most interest in a context of non-smooth stochastic optimization, because their convergence rates are optimal for the classes of Lipschitz continuous (strongly) convex and uniformly convex functions correspondingly. To demonstrate this, one need to obtain sharp lower oracle complexity bounds for both of the classes and show the equivalence their equivalence with convergence rates of the methods. We present the results without proofs, but introduce all the main underlying ideas. Note that lower bounds we consider correspond to the accuracy measured on average.

Let's start with a Lipschitz continuous convex and strongly convex cases. Following the Nemirovski and Yudin, we consider some classes of oracles $\phi \in O$, functions $f \in \mathcal{F}, f : Q \to \mathbb{R}$ and optimization methods $m \in \mathcal{M}_T$ performing at most T iterations before returning point x_T, and then estimate minimax error

$$\epsilon(\mathcal{F}, \phi; T) = \min_{m \in \mathcal{M}_T} \sup_{f \in \mathcal{F}} \mathbb{E}_\phi \left[f(x_T) - \min_{x \in Q} f(x) \right]. \tag{2.8}$$

We equip $\mathbb{R}^r$ with $\|\cdot\|_p$ norm with $p \in [1, +\infty]$, the corresponding dual norm is $\|\cdot\|_q$ with q such that $\frac{1}{p} + \frac{1}{q} = 1$. Further, each of $\phi \in O$ is a first-order stochastic oracle returning unbiased function and subgradient estimations $\hat{f}(x)$ and $\hat{g}(x)$ with bounded variance of the latter:

$$\mathbb{E}[\hat{f}(x)] = f(x), \quad \mathbb{E}[\hat{g}(x)] \in \partial f(x), \quad \mathbb{E}[\|\hat{g}(x)\|_q^2] \le \sigma^2 \tag{2.9}$$

for some $\sigma \ge L$, $\mathcal{F}_L$ is a class of convex Lipschitz continuous functions $f : Q \to \mathbb{R}$:

$$|f(x) - f(y)| \le L\|x - y\|_p, \quad \forall x, y \in Q. \tag{2.10}$$

Herewith $Q \subseteq \mathbb{R}^d$ and contains some closed ball of radius $r > 0$ in $\|\cdot\|_\infty$ norm with a center in origin. Next class we consider is $\mathcal{F}_{\mu,L}$ which consists of Lipschitz continuous functions satisfying μ-strong convexity assumption in $\|\cdot\|_2$ norm:

$$f(\alpha x + (1-\alpha)y) \ge \alpha f(x) + (1-\alpha) f(y) + \alpha(1-\alpha)\frac{\mu^2}{2}\|x-y\|_2^2, \quad \forall x, y \in Q, \alpha \in [0,1]. \tag{2.11}$$

Now, when we have introduced all the necessary notions, without further ado we present the lower complexity bounds.

Theorem 2.1 *Let d, q, ϕ, S, r, L be such as they were introduced above. Then, there exists $c > 0$ such that*

1. *For $q \in [1, 2]$,*

$$\sup_{\phi \in O} \epsilon(\mathcal{F}_L, \phi; T) \geq \min \left\{ cLr\sqrt{\frac{d}{T}}, \frac{Lr}{144} \right\}. \tag{2.12}$$

2. *For $q \in (2, +\infty]$,*

$$\sup_{\phi \in O} \epsilon(\mathcal{F}_L, \phi; T) \geq \min \left\{ cLr\frac{d^{1-\frac{1}{q}}}{\sqrt{T}}, \frac{Ld^{1-\frac{1}{q}}r}{72} \right\}. \tag{2.13}$$

Theorem 2.2 *Let d, q, ϕ, S, r, μ, L be such as they were introduced above. Then, there exist $c_1, c_2 > 0$ such that*

1. *For $q \in [1, 2]$,*

$$\sup_{\phi \in O} \epsilon(\mathcal{F}_{\mu,L}, \phi; T) \geq \min \left\{ c_1\frac{L^2}{\mu^2 T}, c_2 Lr\sqrt{\frac{d}{T}}, \frac{L^2}{1152\mu^2 d}, \frac{Lr}{144} \right\}. \tag{2.14}$$

2. *For $q \in (2, +\infty]$,*

$$\sup_{\phi \in O} \epsilon(\mathcal{F}_{\mu,L}, \phi; T) \geq \min \left\{ c_1\frac{L^2 d^{1-\frac{2}{q}}}{\mu^2 T}, c_2\frac{Lrd^{1-\frac{1}{p}}}{\sqrt{T}}, \frac{L^2 d^{1-\frac{2}{q}}}{1152\mu^2}, \frac{Lrd^{1-\frac{1}{q}}}{144} \right\}. \tag{2.15}$$

The proofs of these results are out of the scope of this book and can be easily found by the reader in the original papers we list in references below. Nonetheless, we would like to explain the main underlying concepts because they are interesting by themselves. Presented lower complexity bounds can be obtained based on lower bounds on hypothesis tests error. This reduction is achieved by the proper choice of functions, such that optimization of one of them leads to unambiguous identification of this one function (uniqueness for enough accuracy is guaranteed). One also need to choose a special oracle (it is allowed to choose it manually, because of $\sup_{\phi \in O}$ in the results) evaluation of which needs in sampling from Bernoulli distribution (coin tossing) with a parameter, specified by the function. Then we notice that T iterations produce T Bernoulli samples and desired accuracy can be achieved only if it is possible to estimate parameters of Bernoulli distribution with some confidence based on T observations (in other words, optimization implies parameter identification, because parameters are chosen to be distinct enough). This possibility is determined by information theory lower bounds based on Fano's inequality and Le Cam's bound.

2.2.4 Constrained Mirror Descent

TODO: this should be in the middle of §4 of Chapter 1 according to the plan in DOCX-file.

Another attractive property of the mirror descent algorithm MD is a possibility to handle not only non-trivial geometry but functional constraints.

We consider the following convex constrained optimization problem over a simple closed convex set $Q \subseteq E$ for E is a finite dimensional normed space, i.e. $\mathbb{R}^n$ endowed with some norm $\|\cdot\|$

$$\begin{aligned} &\min_{x \in Q} f(x), \\ &\text{s.t. } g^{(\ell)}(x) \leq 0, \quad \ell = 1, \ldots, n_g. \end{aligned} \tag{2.16}$$

In the book by A. Nemirovsku and D. Yudin [85] it was observed that the following gradient-type algorithm can handle this type of the problem (with $Q = \mathbb{R}^n$)

$$x^{k+1} = \begin{cases} x^k - \eta \nabla f(x^k), & \max\limits_{\ell=1,\ldots,n_g} g^{(\ell)}(x^k) \leq \varepsilon; \\ x^k - \eta \nabla g^{(\ell(k))}(x^k), & \max\limits_{\ell=1,\ldots,n_g} g^{(\ell)}(x^k) > \varepsilon, \end{cases} \tag{2.17}$$

where $\ell(k) = \arg\max_{\ell=1,\ldots,n_g} g^{(\ell)}(x^k)$ and ε is a desired accuracy and constraint feasibility.

With a slight abuse of notation, next we define a stochastic version of the problem (2.16)

$$\begin{aligned} &\min_{x \in Q} \; f(x) = \mathbb{E}_\xi [f(x, \xi)], \\ &\text{s.t. } g^{(\ell)}(x) = \mathbb{E}_{\xi_\ell} [g^{(\ell)}(x, \xi_\ell)] \leq 0, \quad \ell = 1, \ldots, n_g. \end{aligned} \tag{2.18}$$

Thus particular type of the problem is quite important for the theory of Markov Decision Processes.

Markov Decision Processes and Reinforcement Learning

TODO Daniil: Fill it

$$? \tag{2.19}$$

$$? \tag{2.20}$$

Next in this section we describe a generalization of the gradient-based scheme (2.17) for the problems of type (2.18). Most intriguing, this new scheme is capable to produce solution to the dual problem (2.20) while algorithm works only with the

primal formulation. This fact is essential to extract the near-optimal policy and solve the MDP problem.

Let us start from the providing the list of assumptions Probably first two assumptions could be defined in the previous paragraph.

Assumption 2.3 (Lipschitz continuity) F and all $G^{(\ell)}$ are Lipschitz continuous with constant M for the objective function and for all constraints;

Assumption 2.4 (Uniformly bounded noise) Stochastic gradients are bounded $\|\nabla f(x,\xi)\|_* \le M, \|\nabla g^{(l)}(x,\xi_{(l)})\|_* \le M$ a.s.;

Assumption 2.5 (ϵ-approximation.) There are a set of functions $g_\epsilon^{(\ell)}(x)$ such that $\|g^{(\ell)} - g_\epsilon^{(\ell)}\|_\infty < \epsilon$.

The first assumption is standard for non-smooth optimization, whereas the second one could be potentially weaken. The last assumption could be guaranteed by additional approximation of the expectation within ϵ-accuracy.

Daniil: Next I define here prox-function, Bregman divergence and Mirror step, but it has to be placed in the previous paragraph.

Let us define a *prox-function* $d\colon Q \to \mathbb{R}$ as a continuous 1-strongly convex function d with respect to the norm $\|\cdot\|$ on $\mathbb{R}^n$ that admits continuous selection of subgradients $\nabla d(x)$ where they exist. *Bregman divergence* that corresponds to a prox-function d is a function $V(x,y) = d(y) - d(x) - \langle \nabla d(x), y - x\rangle$.

Given vectors $x \in E$ and $v \in E^*$, the mirror step is defined as

$$x^+ = \mathrm{Mirr}(x,v) = \operatorname*{argmin}_{y\in Q} \{\langle v, y\rangle + V(x,y)\}.$$

We assume that the mirror step can be easily computed.

Next we may describe the algorithm Stochastic Mirror Descent with functional constraints SMD-Constraints. First, let us fix a stepsize η, a desired constraint feasibility ε, an initial point $x^0 = \operatorname{argmin}_{x\in Q} d(x)$, and a number of iterates N.

Then for each step $k = 0, \dots, N-1$ we firstly compute

$$\ell(k) = \arg\max_{\ell=1,\dots,n_g} g_\epsilon^{(\ell)}(x^k),$$

and then

$$x^{k+1} = \begin{cases} \mathrm{Mirr}(x^k, \eta\nabla f(x^k,\xi^k)) & \text{if } g_\epsilon^{(\ell(k))} < \epsilon + \varepsilon, \quad \text{"productive step"} \\ \mathrm{Mirr}(x^k, \eta\nabla g^{(\ell(k))}(x^k,\xi^k)) & \text{if } g_\epsilon^{(\ell(k))} \ge \epsilon + \varepsilon, \quad \text{"non-productive step".} \end{cases} \tag{2.21}$$

In other words, we produce steps to the direction of the stochastic gradient of the objective function if our current point is feasible enough, and we are trying to improve the most violated constraint otherwise. Additionally, we define $I = \{k : g_\epsilon^{(\ell(k))} < \epsilon + \varepsilon\}$ a set of productive iterates, and J is a set of non-productive iterates.

Next we define the first performance criteria for our problem.

Definition 2.1 (Approximate primal solution) A vector $\hat{x} \in Q$ is called an $(\varepsilon_f, \varepsilon_g, \delta)$-solution to the primal problem (2.18), if the following holds with probability at least $1 - \delta$

$$\begin{aligned} f(\hat{x}) - f(x_*) &\leq \varepsilon_f, \\ g^{(\ell)}(\hat{x}) &\leq \varepsilon_g \quad \ell = 1, \ldots, n_g \end{aligned} \tag{2.22}$$

where x_* is a solution to the problem (2.18).

Next we are going to state the convergence result.

Theorem 2.3 *Let $\Theta_0^2 = d(x_*) - d(x^0)$. Then Algorithm* SMD-Constraints *with $\eta = \varepsilon/M^2$ and $\hat{x} = \frac{1}{|I|}\sum_{k\in I} x^k$ produces $(\varepsilon, \varepsilon + 2\epsilon, \delta)$-solution to the primal problem* (2.18) *in*

$$N = O\left(\frac{\Theta_0^2 M^2 \log(1/\delta)}{\varepsilon^2}\right)$$

iterates.

We start from the technical statement that is extremely useful for the analysis.

Lemma 2.1 *Let h be some convex function over a set Q, $\eta > 0$ is a stepsize, $x \in Q$. Let the point $x^+ = \mathrm{Mirr}(x, \eta(\nabla h(x) + \Delta))$, where Δ is some vector from the dual space. Then, for any $y \in Q$*

$$\eta(h(x) - h(y) + \langle \Delta, x - y\rangle) \leq \eta\langle \nabla h(x) + \Delta, x - y\rangle \leq \frac{\eta^2}{2}\|\nabla h(x) + \Delta\|_*^2 + V(x, y) - V(x^+, y).$$

Proof Daniil: Probably it appears in earlier proofs. □

Daniil: It is a kind of the chaos in the notation. Is it possible to simplify it? Denote $\hat{\nabla}_k f = \nabla f(x^k, \xi^k), \nabla_k f = \nabla f(x^k)$ and $\hat{\nabla}_k g^{(\ell)} = \nabla g^{(\ell)}(x^k, \xi_\ell^k), \nabla_k g^{(\ell)} = \nabla g^{(\ell)}(x^k)$. Next we define the stochastic gradient noise function

$$\gamma_k(y) = \begin{cases} \eta\langle \hat{\nabla}_k f - \nabla_k f, y - x^k\rangle, & k \in I; \\ \eta\langle \hat{\nabla}_k g^{(\ell(k))} - \nabla_k g^{(\ell(k))}, y - x^k\rangle, & k \in J. \end{cases} \tag{2.23}$$

Then we provide very useful lemma that produces objective residual decomposition.

Lemma 2.2 *For a point $\hat{x}$ defined in Theorem 2.3 and any $y \in Q$ the following holds*

$$\begin{aligned} \eta|I| \cdot (f(\hat{x}) - f(y)) \leq \frac{\eta^2 M^2}{2} N + [d(y) - d(x^0)] - |J|\eta \cdot \varepsilon \\ + \sum_{k=0}^{N-1} \gamma_k(y) + \eta \sum_{k\in J} g^{(\ell(k))}(y). \end{aligned} \tag{2.24}$$

Proof By the construction of productive and non-productive steps in (2.21) and Lemma 2.1, we have for all $y \in Q$

$$\begin{aligned}
\eta(F(x^k) - F(y)) &\leq \frac{\eta^2 M^2}{2} + V(x^k, y) - V(x^{k+1}, y) && k \in I;\\
&\quad + \eta\langle \hat{\nabla}_k f - \nabla_k f, y - x^k\rangle,\\
\eta(g^{(\ell(k))}(x^k) - g^{(\ell(k))}(y)) &\leq \frac{\eta^2 M^2}{2} + V(x^k, y) - V(x^{k+1}, y) && k \in J.\\
&\quad + \eta\langle \hat{\nabla}_k g^{(\ell(k))} - \nabla_k g^{(\ell(k))}, y - x^k\rangle
\end{aligned}$$

By definition of ϵ-approximation, we have the following inequalities for "productive" and "non-productive" steps respectively

$$\begin{aligned}
\eta(f(x^k) - f(y)) &\leq \frac{\eta^2 M^2}{2} + V(x^k, y) - V(x^{k+1}, y) + \gamma_k(y),\\
\eta(g_\epsilon^{(\ell(k))}(x^k) - g^{(\ell(k))}(y)) &\leq \frac{\eta^2 M^2}{2} + V(x^k, y) - V(x^{k+1}, y) + \gamma_k(y) + \eta\epsilon.
\end{aligned}$$

Sum all these inequalities over all $k \in I$ and $k \in J$ and use the fact that $I \cup J = \{0, \ldots, N-1\}$

$$\begin{aligned}
&\sum_{k\in I} \eta(f(x^k) - f(y)) + \sum_{k\in J} \eta(g_\epsilon^{(\ell(k))}(x^k) - g^{(\ell(k))}(y))\\
&\leq \frac{\eta^2 M^2}{2}|I| + \frac{\eta^2 M^2}{2}|J| + \sum_{k=0}^{N-1}[V(x^k, y) - V(x^{k+1}, y)] + \sum_{k=0}^{N-1}\gamma_k(y) + |J|\eta\epsilon.
\end{aligned} \tag{2.25}$$

By the choice of $x^0 = \operatorname{argmin}_{x\in Q} d(x)$, we have

$$\sum_{k=0}^{N-1}[V(x^k, y) - V(x^{k+1}, y)] \leq V(x^0, y) = d(y) - d(x^0) - \langle \nabla d(x^0), y - x^0\rangle = d(y) - d(x^0).$$

Using definition of "non-productive" steps $g_\epsilon^{(\ell(k))}(x^k) > \varepsilon + \epsilon$ and convexity of f

$$\begin{aligned}
&\sum_{k\in I} \eta(f(x^k) - f(y)) + \sum_{k\in J} \eta(g_\delta^{(\ell(k))}(x^k) - g^{(\ell(k))}(y))\\
&\geq \eta|I|(f(\hat{x}) - f(y)) + \eta|J|(\varepsilon + \epsilon) - \eta\sum_{k\in J} g^{(\ell(k))}(y).
\end{aligned}$$

By application of inequality (2.25) and regrouping of terms, we finish the proof. □

Proof (Proof of Theorem 2.3) Let us analyze the statement of Lemma 2.2 for $y = x_*$. In this case we notice that $g^{(\ell)}(x_*) \leq 0$ due to feasibility of the optimal point and we have

$$\eta|I| \cdot (f(\hat{x}) - f(x_*)) \leq \frac{\eta^2 M^2}{2} N + \Theta_0^2 - |J|\eta \cdot \varepsilon + \sum_{k=0}^{N-1} \gamma_k(y).$$

Therefore, it is sufficient to analyze $\sum_{k=0}^{N-1} \gamma_k(y)$. Define filtration of σ-algebras $\mathcal{F}_k = \sigma\left(\{\xi^i, \xi^i_{(\ell)}\}_{i \leq k}\right)$. It is easy to see that $\gamma_k(x^*)$ is a martingale-difference sequence adapted to $\mathcal{F}_k$. Additionally, notice that for any $y \in Q$: $\|y - x^0\|^2 \leq 2(d(y) - d(x^0))$.

Thus, by Azuma-Hoeffding inequality and recalling the definition $\eta = \varepsilon/M^2$ we may produce the following high-probability bound

Lemma 2.3 *Define the event $\mathcal{E}$ such that the following inequalities holds*

$$\sum_{k=0}^{N-1} \gamma_k(x_*) < \frac{2\Theta_0}{M} \cdot \sqrt{2N\varepsilon^2 \log(1/\delta)}.$$

Then $\mathbb{P}[\mathcal{E}] \geq 1 - \delta$. □

Therefore we have the following bound on objective residual

$$\eta|I| \cdot (f(\hat{x}) - f(x_*)) \leq \eta|I|\varepsilon \underbrace{- \frac{\varepsilon^2 N}{2M^2} + \Theta_0^2 + \frac{4\Theta_0\sqrt{2N\varepsilon^2 \log(1/\delta)}}{M}}_{\text{remainder}}.$$

We notice that the remainder term is negative for sufficiently large value of $N\varepsilon^2$. Formally speaking, if

$$N \geq \frac{280 \cdot \Theta_0^2 M^2 \log(1/\delta)}{\varepsilon^2}$$

then

$$f(\hat{x}) - f(x_*) \leq \varepsilon.$$

The bound on constraint feasibility simply follows from the definition on "productive" steps, ϵ-approximation, and convexity

$$g^{(\ell)}(\hat{x}) \leq \frac{1}{|I|} \sum_{k \in I} g^{(\ell)}(x^k) \leq \frac{1}{|I|} \sum_{k \in I} (g_\epsilon^{(\ell)}(x^k) + \delta) \leq \varepsilon + 2\epsilon.$$

Next we show that Algorithm SMD-Constraints also can generate solution to the Lagrange dual problem and optimize the duality gap, that is very useful for solving MDPs. Let us recall the definition of the Lagrange dual problem

$$\max_{\lambda \in \mathbb{R}_+^{n_g}} \left\{ \phi(\lambda) = \min_{x \in Q} \left\{ f(x) + \sum_{\ell=1}^{n_g} \lambda_\ell g^{(\ell)}(x) \right\} \right\}. \tag{2.26}$$

It is well-known that for any $x \in Q : g^{(\ell)}(x) \leq 0 \; \forall \ell \in \{1, \ldots, n_g\}$ and $\lambda \in \mathbb{R}_+^{n_g}$ *the weak duality* holds: $\Delta(x, \lambda) = f(x) - \phi(\lambda) \geq 0$, where Δ is so-called *the duality*

gap. We assume that for our primal problem (2.18) the Slater's condition holds, i.e. $\exists x \in Q : \forall \ell \in \{1, \dots, m\} : g^{(\ell)}(x) < 0$. It implies that the dual problem has a solution and there is *the strong duality*: $\Delta(x^*, \lambda^*) = 0$ for any x^* and λ^* are solutions to the primal and the dual problems respectively.

Then we may define another performance criteria for a pair of approximate solution to the primal and dual problems under Slater's condition.

Definition 2.2 (Approximate primal-dual solution) Let us call the pair $(\hat{x}, \hat{\lambda})$ a primal-dual $(\varepsilon_\Delta, \varepsilon_g, \sigma)$-solution to (2.18) if the following holds with probability at least $1 - \delta$

$$\begin{aligned} \Delta(\hat{x}, \hat{\lambda}) = f(\hat{x}) - \phi(\hat{\lambda}) \leq \varepsilon_\Delta, \\ g^{(\ell)}(\hat{x}) \leq \varepsilon_g \quad \ell = 1, \dots, n_g. \end{aligned} \tag{2.27}$$

Our estimate of dual variables is defined as follows

$$\hat{\lambda}_\ell = \frac{1}{|I|} \sum_{k \in J} \mathbb{I}\{\ell = \ell(k)\}. \tag{2.28}$$

Theorem 2.4 *Let* $\overline{\Theta}_0^2 = \sup_{y \in Q}(d(y) - d(x^0))$. *Choose* $\hat{\lambda} \in \mathbb{R}_+^m$ *as defined in* (2.28), $\hat{x} = \frac{1}{|I|} \sum_{k \in I} x^k$, *and a constant stepsize* $\eta = \varepsilon / M^2$. *Then the pair* $(\hat{x}, \hat{\lambda})$ *is an primal-dual* $(\varepsilon, \varepsilon + 2\epsilon, \delta)$*-solution for* $\varepsilon > 0, \epsilon \geq 0, \delta \in (0, 1/2)$ *after*

$$N \geq N_0' = O\left(\frac{\overline{\Theta}_0^2 M^2 (\log(1/\delta) + \kappa(E^*))}{\varepsilon^2}\right),$$

where $\kappa(E^*)$ *is a constant of Nemirovski's inequality TODO: Reference to appendix for the dual space.*

If E has a finite dimension n, then we always have $\kappa(E^*) \leq n$. Additionally, if E is endowed with ℓ_p norm, then E^* is endowed with ℓ_q norm, where $1/p + 1/q = 1$, and there is a more precise bound

$$\kappa(E^*) \leq K\left(\frac{p}{p-1}, d\right) = \begin{cases} d^{\frac{2}{p}-1}, & p \in [1, 2] \\ d^{1-\frac{2}{p}}, & p \in (2, +\infty] \end{cases}$$

In particular, if E has ℓ_2 norm, $\kappa(E^*) = 1$. For $p \in [2, +\infty]$ this bound is tight, however, in the case $p \in [1, 2)$ and $d \geq 3$ it could be improved to, for instance, a logarithmic bound $\kappa(E^*) \leq 2e \log(d) - e$, that could be useful in the case of ℓ_1-norm and an entropy prox-function.

The only difference with guarantees in the primal setting is a presence of the constant $\kappa(E^*)$ that handle the geometry of the space of stochastic gradients. Next we are going to prove this theorem.

Proof (Proof of Theorem 2.4) Now we start from the result of Lemma 2.2 for arbitrary $y \in Q$.

$$\eta|I|\cdot(f(\hat{x})-f(y))\le\frac{\eta^2M^2}{2}N+[d(y)-d(x^0)]-|J|\eta\cdot\varepsilon+\sum_{k=0}^{N-1}\gamma_k(y)+\eta\sum_{k\in J}g^{(\ell(k))}(y).$$

Notice that $\hat{\lambda}_l$ corresponds to the number of times $g^{(\ell(k))}(y)$ appears in the last sum up to a scaling factor $|I|$. Therefore, it could be rewritten as follows

$$\eta|I|\cdot\left(f(\hat{x})-\left\{f(y)+\sum_{\ell=1}^{n_g}\hat{\lambda}_\ell g^{(\ell)}(y)\right\}\right)\le\frac{\eta^2M^2}{2}N+\overline{\Theta}_0^2-|J|\eta\cdot\varepsilon+\sum_{k=0}^{N-1}\gamma_k(y).$$

If we take supremum over $y\in Q$, in the left-hand side we receive exactly the duality gap

$$\eta|I|\cdot\Delta(\hat{x},\hat{\lambda})\le\frac{\eta^2M^2}{2}N+\overline{\Theta}_0^2-|J|\eta\cdot\varepsilon+\sup_{y\in Q}\sum_{k=0}^{N-1}\gamma_k(y).\tag{2.29}$$

To control the right-hand side we will use the definition of $\gamma_k(y)$, its linearity in y, and Hölder inequality

$$\sum_{k=0}^{N-1}\gamma_k(y)=\sum_{k=0}^{N-1}\langle\Delta_k,y-x^0+x^0-x^k\rangle\le\|y-x^0\|\left\|\sum_{k=0}^{N-1}\Delta_k\right\|_*+\sum_{k=0}^{N-1}\gamma_k(x^0),$$

where Δ_k are defined as follows

$$\Delta_k=\begin{cases}\eta\left(\hat{\nabla}_k f-\nabla_k f\right), & k\in I\\ \eta\left(\hat{\nabla}_k g^{(l(k))}-\nabla_k g^{(l(k))}\right), & k\in J.\end{cases}$$

By a uniform bound on $\|y-x^0\|$ in terms of $\overline{\Theta}_0$ we have

$$\sup_{y\in Q}\sum_{k=0}^{N-1}\gamma_k(y)\le\sqrt{2\overline{\Theta}_0^2}\left\|\sum_{k=0}^{N-1}\Delta_k\right\|_*+\sum_{k=0}^{N-1}\gamma_k(x^0).$$

Thus, we have to control these two terms. For the first one we may apply McDiarmid's inequality TODO: refer to appendix. Indeed, the change of one Δ_k could change the norm of the sum only by $2\|\Delta_k\|_*\le 4\eta M=4\varepsilon M^{-1}$, and all noise ξ_ℓ^k is independent. Thus with probability at least $1-\delta/2$

$$\left\|\sum_{k=0}^{N-1}\Delta_k\right\|_*\le\frac{4\varepsilon}{M}\cdot\sqrt{2N\log(2/\delta)}+\mathbb{E}\left[\left\|\sum_{k=0}^{N-1}\Delta_k\right\|_*\right].$$

To bound the expectation term in this representation, we apply Jensen's and Nemirovski's inequalities

$$\left(\mathbb{E}\left\|\sum_{k=0}^{N-1}\Delta_k\right\|_*\right)^2 \leq \mathbb{E}\left[\left\|\sum_{k=0}^{N-1}\Delta_k\right\|_*^2\right] \leq \kappa(E^*)\sum_{k=0}^{N-1}\mathbb{E}\left[\|\Delta_k\|_*^2\right] \leq \kappa(E^*)\frac{4N\varepsilon^2}{M^2}.$$

Finally, bound on $\sum_{k=0}^{N-1}\gamma_k(x^0)$ follows from Azuma-Hoeffding inequality with probability at least $1-\delta/2$

$$\sum_{k=0}^{N-1}\gamma_k(x^0) \leq \frac{2\overline{\Theta}_0}{M}\cdot\sqrt{2N\varepsilon^2\log(2/\delta)}.$$

Thus, by union bound with probability at least $1-\delta$ we have

$$\sup_{y\in Q}\sum_{k=0}^{N-1}\gamma_k(y) \leq \frac{\overline{\Theta}_0\sqrt{2N\varepsilon^2}}{M}\left((4\sqrt{2}+2)\sqrt{\log(2/\delta)}+2\sqrt{\kappa(E^*)}\right).$$

The bound on the duality gap (2.30) became the following

$$\eta|I|\cdot\Delta(\hat{x},\hat{\lambda}) \leq \varepsilon\eta|I|\underbrace{-\frac{\varepsilon^2 N}{2M^2}+\overline{\Theta}_0^2+\frac{\overline{\Theta}_0\sqrt{2N\varepsilon^2}}{M}\left((4\sqrt{2}+2)\sqrt{\log(2/\delta)}+2\sqrt{\kappa(E^*)}\right)}_{\text{remainder}}. \tag{2.30}$$

Here we see that the remainder term has exactly the same structure as in the proof of Theorem 2.3. Thus, taking $N = O\left(\overline{\Theta}_0^2 M^2\varepsilon^{-2}(\log(1/\delta)+\kappa(E^*))\right)$ we conclude the statement.

2.3 Stochastic Mirror Descent: Online Setting

In Section 1.1.2, it was mentioned the online approach to the optimization problem in the Euclidean setting and was shown the SGD as an online learning procedure in the standard online sense (see (1.14)).

In this section, we mention some results of the online approach in the non-Euclidean setting, where, for example, the feasible set is a simplex, which seems essential for the problem of experts [**?**, **?**, 51, 20] and other problems (see also the listed examples below).

In general, an online convex optimization problem can be formulated as a repeated game between a learner and an adversary: at each iteration k, the learner first presents a solution $x_k \in Q$, where $Q \subset \mathbb{R}^n$ is a convex set, it then receives a convex function $f_k(x) : Q \longrightarrow \mathbb{R}$, and suffers the loss $f_k(x_k)$ for the submitted solution x_k. The objective of the learner is to generate a sequence of solutions $x_k \in Q, k = 1,\dots N$ that minimizes the regret Regret_N, which is defined as follows:

$$\text{Regret}_N := \frac{1}{N}\sum_{k=1}^{N} f_k(x_k) - \min_{x\in Q}\frac{1}{N}\sum_{k=1}^{N} f_k(x). \tag{2.31}$$

This regret measures the difference between the cumulative loss of the learner's strategy and the minimum possible loss had the sequence of loss functions been known in advance and the learner could choose the best-fixed action in hindsight.

Many successful algorithms have been developed over the past decade to minimize the regret in online convex optimization. In the seminal work of Zinkevich [**?**], it was presented an algorithm based on gradient descent with projection (Euclidean setting) that guarantees a regret of $O\left(\sqrt{N}\right)$ when the loss functions are Lipschitz continuous in the convex set Q. For strongly convex loss functions, in [**?**, **?**] it was proposed algorithms with logarithmic regret bound, for the Euclidean setup of the problem also.

It is worthy mention that the estimate $O\left(\sqrt{N}\right)$ of the bound of the regret is an optimal for the class of problems with Lipschitz continuous functions. Where any algorithm for online convex optimization incurs $\Omega\left(DM\sqrt{N}\right)$ regret in the worst case, where M is the Lipschitz constant of the objective function and D is an upper bound of the diameter of the convex set Q [50].

For the non-Euclidean case, which is very important in many applications where the Euclidean setting failed, it is convenient to use the Mirror Descent approach.

2.3.1 Examples of problems that can be modeled via online optimization

Multi-armed bandits

The problems where we do not receive the full-information, i.e., we do not observe the loss vector, are called **bandit problems**. The name comes from the problem of a gambler who plays a pool of slot machines, that can be called "one-armed bandits". On each round, the gambler places his bet on a slot machine and his goal is to win almost as much money as if he had known in advance which slot machine would return the maximal total reward [92].

Let we assume that there are n different handles. The game is repeated $N \gg 1$ times (this number may be unknown in advance). At each step k, we must choose a handle $i(k)$ to tug. Tugging a handle involves certain, generally random, losses $r^k_{i(k)}$ (for definiteness we assume that $r^k_{i(k)} \in [0,1]$) which depend on the number of the step, the number of the handle, and the tugging strategy which used up to step k. Our strategy at the step k is described by the vector of probability distributions $x^k \in S_n(1)$ according to which we chooses a handle to tug independently of anything else. The only information which we have at the step k is the vector

$$\left(\left(x^1, i(1), r^1_{i(1)}\right), \ldots, \left(x^{k-1}, i(k-1), r^{k-1}_{i(k-1)}\right)\right).$$

We assume that the losses r^k at the k-th step depend on x^k but do not depend on the result of playing the distribution x^k; they also depend on $(x^1, \ldots, x^{k-1})$ and on the results of the corresponding playing and on $(r^1, \ldots, r^{k-1})$. Thus, our purpose is to organize the handle tugging procedure so as to minimize the expected total losses. Let us introduce the function $f_k(x, \xi^k) = r^k_i$, with probability x_i, $\forall i = 1, \ldots, n$, where r^k is independent of the result of playing the probability distribution ξ^k determined by x. The generalized stochastic gradient of this function, with probability x_i, is

$$\nabla_x f_k\left(x, \xi^k\right) = (\underbrace{0, \ldots, r^k_i/x_i}_{i}, \ldots, 0)^{\mathrm{T}}, \quad i = 1, \ldots, n.$$

To this setting of the problem we have the following estimates [**?**]

$$O\left(\sqrt{\frac{n \ln n}{N}}\right) \text{ on average, and } O\left(\sqrt{\frac{n \ln (n/\sigma)}{N}}\right) \text{ with probability } \geq -\sigma,$$

which are optimal for the given class of problems [**?**, 20, **?**].

Weighting expert solutions and linear losses

Consider the problem of weighting expert solutions [20, **?**], There are n different experts. Each expert plays at market. The game is repeated $N \gg 1$ times (this number may be unknown in advance). Let l^k_i be the loss of an expert i at a step k, such that $|l^k_i| \leq M$. At each step k, we share one dollar among the experts according to a vector $x^k \in S_n(1)$. Our losses involved are calculated from the losses $\langle l^k, x^k \rangle$ of the experts. The purpose is to organize the procedure for distributing a dollar at each step so as to minimize our total losses. The expert losses l^k are allowed to depend also on the current move x_k. To this setting of the problem, with

$$f_k(x, \xi^k) = f_k(x) = \langle l^k, x \rangle,$$

we have the following estimate (for the losses, i.e. regret) $O\left(M\sqrt{\frac{\ln n}{N}}\right)$ [**?**], which is optimal for the given class of problems.

Weighting of expert solutions and convex losses

Let Q be a convex set. Under the conditions of the previous example, suppose that, at the k-th step, the i expert uses a strategy $\zeta^k_i \in Q$, which involves losses $\lambda(\omega^k, \xi^k_i)$, where ω^k is a move, possibly of the hostile "nature", which knows, in particular,

our current strategy. The function $\lambda(\cdot,\cdot)$ is convex in the second argument, and $|\lambda(\cdot,\cdot)| \le M$. At each step, we must choose a new strategy $x := \sum_{i=1}^{n} x_i \cdot \zeta_i^k \in Q$ involving losses $\lambda(\omega^k, x)$ so as to minimize our total losses. The function $\lambda(\omega^k, x)$ is convex in x for any ω^k, and therefore,

$$\sum_{k=1}^{N} \lambda\left(\omega^k, x^k\right) - \min_{i\le n} \sum_{k=1}^{N} \le \lambda\left(\omega^k, \zeta_i^k\right) \le \sum_{k=1}^{N} f_k\left(x^k\right) - \min_{x\in S_n(1)} \sum_{k=1}^{N} f_k(x)$$

To this setting of the problem, with

$$f_k\left(x, \xi^k\right) = f_k(x) = \sum_{i=1}^{n} x_i \lambda\left(\omega^k, \zeta_i^k\right) \ge \lambda\left(\omega^k, x\right),$$

we have the following estimate (for the losses, i.e. regret) $O\left(M\sqrt{\frac{\ln n}{N}}\right)$ [**?**], which is optimal for the given class of problems [20, **?**].

Weighting of expert solutions and non-convex losses

Under the conditions of the previous example connected with the weighting of expert solutions and convex losses, suppose that we cannot guarantee the convexity of $\lambda(\cdot,\cdot)$ in the second argument. Then we choose a strategy (a probability distribution on the set of expert strategies) and play the random variable according to this probability distribution. To this setting of the problem, with

$$f_k\left(x, \xi^k\right) = f_k(x) = \sum_{i=1}^{n} x_i \lambda\left(\omega^k, \zeta_i^k\right),$$

we have the following estimates (for the losses, i.e. regret) [**?**]

$$O\left(M\sqrt{\frac{\ln n}{N}}\right) \text{ on average, and } O\left(M\sqrt{\frac{\ln (n/\sigma)}{N}}\right) \text{ with probability } \ge -\sigma,$$

which are optimal for the given class of problems [20, **?**].

2.3.2 Online optimization for strongly convex functions

It is known that in offline convex optimization, the convergence bounds are varying over different classes of convex objective functions. The same fact holds in online convex optimization and for important classes of objective functions (such

strongly convex) significantly better regret bounds are possible. Such regret is called **logarithmic regret**, which is significantly better than $O\left(\sqrt{N}\right)$.

Let $f : Q \longrightarrow \mathbb{R}$ be a μ-strongly convex function and Lipschitz continuous with some constant $M > 0$ (see (B.7)), and let we consider the the following online variant of subgradient method

$$x_{k+1} = \text{Proj}_Q\left(x_k - \frac{1}{\mu k}\nabla f_k(x_k)\right), \quad k = 0, 1, \ldots, N \tag{2.32}$$

Then this algorithm achieves the following estimate of the regret (see [51] for more details and proof)

$$\text{regret}_N \leq \frac{M^2}{2\mu N}\left(1 + \log N\right). \tag{2.33}$$

We emphasize here, that the result about the bounds for the regret for the strongly convex functions is in the Euclidean setting, where until now to the best of our knowledge there is no generalization of this result with arbitrarily prox function and non-Euclidean settings.

Chapter 3
Convex Stochastic Optimization: Smooth Case

Abstract TODO for Eduard: write an abstract

TODO for Eduard: write a short introduction

Eduard: I wrote the problem formulation and L-smoothness assumption here for now. Once we have a structure of the book, we will put this to the better place.

In this chapter, we focus on the optimization methods for solving expectation minimization problems

$$\min_{x \in \mathbb{R}^n} \left\{ f(x) = \mathbb{E}_{\xi \sim \mathcal{D}}[f_\xi(x)] \right\} \tag{3.1}$$

with f being L-smooth meaning that f is differentiable and for all $x, y \in \mathbb{R}^n$ its gradient is L-Lipschitz:

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2. \tag{3.2}$$

3.1 Stochastic Gradient Descent

Stochastic Gradient Descent (SGD) is the first stochastic optimization method for solving problems of the form (3.1). From the given starting point $x^0 \in \mathbb{R}^n$ SGD generates the points $\{x^k\}_{k \geq 0}$ satisfying the following update rule:

$$x^{k+1} = x^k - \gamma_k g^k. \tag{3.3}$$

That is, at iteration k the method computes a random vector g^k called *stochastic gradient* at the point x^k and calculates x^{k+1} using formula (3.3) for some choice of stepsize γ_k. Such a description of the method leaves at least two significant questions: 1) how to choose stepsize γ_k and 2) how to choose stochastic gradient g^k?

Despite the seeming simplicity, each of these questions is complex on its own and was studied by many researchers throughout the years. Moreover, the significant part of this book is devoted to these two questions. We provide state-of-the-art results

regarding the first question in Chapter here we should cite the chapter on adaptive algorithms. In this chapter, we mainly focus on the second question and assume (unless the different is claimed) that $\gamma_k \equiv \gamma > 0$, where γ is some constant (in general, problem-dependent) stepsize.

3.1.1 Analysis of SGD: Uniformly Bounded Variance Case

To study the convergence of SGD we need to specify the assumptions on the stochastic gradient g^k. Typically, g^k is assumed to be an unbiased estimator of $\nabla f(x^k)$ at the given point x^k:

$$\mathbb{E}_k[g^k] = \nabla f(x^k), \tag{3.4}$$

where $\mathbb{E}_k[\cdot]$ denotes a conditional expectation with frozen $x^0, x^1, \ldots, x^k$. However, the unbiasedness property is not enough to ensure the convergence of the method since the noise in the stochastic gradient can be quite heavy-tailed and the difference between g^k and $\nabla f(x^k)$ can be large with noticeable probability. Therefore, some assumptions about second order moments of g^k are usually introduced. Perhaps, the most popular one is the assumption of uniformly bounded variance.

Assumption 3.1 (Uniformly Bounded Variance) There exists a constant $\sigma \geq 0$ such that for any $k \geq 0$ stochastic gradient g^k is unbiased, i.e., (3.4) holds, and satisfies

$$\mathbb{E}_k\left[\|g^k - \nabla f(x^k)\|^2\right] \leq \sigma^2. \tag{3.5}$$

Let us discuss the introduced assumption. First of all, the word "uniformly" in the name of the assumption means that the stochastic gradient should have bounded variance at the whole space $\mathbb{R}^n$. Indeed, without any additional assumptions on the method and/or stochastic gradient SGD can leave any bounded set containing the solution of the problem with a positive probability. Next, we emphasize that boundedness of the variance of the stochastic gradient does not imply boundedness of the norm of the (stochastic) gradient.

To better illustrate the assumption, we provide some examples when Assumption 3.1 holds.

Deterministic Gradient

Let $g^k = \nabla f(x^k)$ for all $k \geq 0$, i.e., there is no stochasticity in the method. Then Assumption 3.1 holds with $\sigma = 0$.

Additive Noise

Assume that $g^k = \nabla f(x^k) + \xi^k$, where ξ^k is generated from normal distribution $\mathcal{N}\left(0, \frac{\sigma^2}{n}\mathbf{I}_n\right)$ independently from the previous iterations. Then, we have

$$\mathbb{E}_k\left[\|g^k - \nabla f(x^k)\|^2\right] = \mathbb{E}\left[\|\xi^k\|^2\right] = \sum_{i=1}^{n} \mathbb{E}\left[(\xi_i^k)^2\right] = n \cdot \frac{\sigma^2}{n} = \sigma^2,$$

meaning that Assumption 3.1 is satisfied in this case as well.

Logistic Regression

Consider the following optimization problem called *logistic regression*, which is a classical example of the loss function for binary classification tasks:

$$\min_{x \in \mathbb{R}^n} \left\{ f(x) = \frac{1}{m} \sum_{i=1}^{m} f_i(x) \right\}, \quad \text{where } f_i(x) = \log\left(1 + \exp\left(-y_i \cdot [Ax]_i\right)\right), \quad (3.6)$$

and $A \in \mathbb{R}^{m \times n}$, $y \in \{-1, 1\}^m$. Here vector x denotes the vector of learnable parameters, rows of matrix A define training examples, and the components of vector $y = (y_1, \ldots, y_m)^\top$ are the labels/answers on the corresponding examples. Consider $g^k = \nabla f_{i_k}(x^k)$, where i_k is sampled uniformly at random from $\{1, \ldots, m\}$ independently from the previous iterations. Then, g^k is conditionally unbiased estimate of $\nabla f(x^k)$:

$$\mathbb{E}_k\left[g^k\right] = \frac{1}{m} \sum_{i=1}^{m} \nabla f_i(x^k) = \nabla f(x^k).$$

Next, for convenience, we also define the rows of matrix A as $a_1, a_2, \ldots, a_m \in \mathbb{R}^n$. Then, for each $i \in \{1, \ldots, m\}$ we have

$$\nabla f_i(x) = -\frac{y_i}{1 + \exp(y_i a_i^\top x)} a_i$$

and

$$\nabla f(x) = -\frac{1}{m} \sum_{i=1}^{m} \frac{y_i}{1 + \exp(y_i a_i^\top x)} a_i.$$

Therefore, using $1 + \exp(y_i a_i^\top x) \geq 1$ for all $x \in \mathbb{R}^n$, $i \in \{1, \ldots, n\}$ and the classical relation between the second moment and the variance, we derive the following inequality:

$$\mathbb{E}_k\left[\|g^k - \nabla f(x^k)\|^2\right] = \mathbb{E}_k\left[\|g^k\|^2\right] = \frac{1}{m} \sum_{i=1}^{m} \left\| \frac{y_i a_i}{1 + \exp(y_i a_i^\top x)} \right\|^2$$

$$\leq \frac{1}{m} \sum_{i=1}^{n} \|a_i\|^2 = \sigma^2.$$

That is, Assumption 3.1 holds with σ defined above.

Under Assumption 3.1 the analysis of SGD goes almost the same lines as the analysis of Gradient Descent.

TODO for Eduard: put the definition of quasi strong convexity somewhere in the text (I guess, we will have a special section for this).

Lemma 3.1 *Let the objective function f be μ-quasi strongly convex and L-smooth and Assumption 3.1 holds. Assume that the stepsize γ satisfies $0 < \gamma < 1/L$. Then, for each $k \geq 0$ the iterations of* SGD *satisfy*

$$\begin{aligned}\mathbb{E}\left[\|x^{k+1} - x^*\|^2\right] \leq (1-\gamma\mu)\mathbb{E}\left[\|x^k - x^*\|^2\right] + \gamma^2\sigma^2 \\ -2\gamma(1-\gamma L)\mathbb{E}\left[f(x^k) - f(x^*)\right].\end{aligned} \tag{3.7}$$

Proof We start with expanding the square:

$$\begin{aligned}\|x^{k+1} - x^*\|^2 &= \|x^k - x^* - \gamma g^k\|^2 \\ &= \|x^k - x^*\|^2 - 2\gamma\langle x^k - x^*, g^k\rangle + \gamma^2\|g^k\|^2.\end{aligned}$$

Next, taking the conditional expectation $\mathbb{E}_k[\cdot]$ from the above inequality and using the unbiasedness of g^k, we derive

$$\begin{aligned}\mathbb{E}_k\left[\|x^{k+1} - x^*\|^2\right] &= \|x^k - x^*\|^2 - 2\gamma\langle x^k - x^*, \mathbb{E}_k[g^k]\rangle + \gamma^2\mathbb{E}_k\left[\|g^k\|^2\right] \\ &\overset{(3.4)}{=} \|x^k - x^*\|^2 - 2\gamma\langle x^k - x^*, \nabla f(x^k)\rangle + \gamma^2\mathbb{E}_k\left[\|g^k\|^2\right].\end{aligned}$$

The second term from the right-hand side of the above inequality can be upper-bounded using quasi-strong convexity of f

$$-2\gamma\langle x^k - x^*, \nabla f(x^k)\rangle \leq -2\gamma\left(f(x^k) - f(x^*)\right) - \gamma\mu\|x^k - x^*\|^2,$$

and to bound the third term one can apply variance decomposition, Assumption 3.1, and upper bound for the squared norm of the smooth function (Eduard: might be better to create the list of basic inequalities to refer to them throughout the text)

$$\begin{aligned}\gamma^2\mathbb{E}_k\left[\|g^k\|^2\right] &= \gamma^2\|\nabla f(x^k)\|^2 + \gamma^2\mathbb{E}_k\left[\|g^k - \nabla f(x^k)\|^2\right] \\ &\overset{(3.5)}{\leq} \gamma^2\|\nabla f(x^k)\|^2 + \gamma^2\sigma^2 \\ &\leq 2\gamma^2 L\left(f(x^k) - f(x^*)\right) + \gamma^2\sigma^2.\end{aligned} \tag{3.8}$$

Putting all together, we get the following inequality

$$\begin{aligned}\mathbb{E}_k\left[\|x^{k+1} - x^*\|^2\right] \leq (1-\gamma\mu)\|x^k - x^*\|^2 + \gamma^2\sigma^2 \\ -2\gamma(1-\gamma L)\left(f(x^k) - f(x^*)\right).\end{aligned}$$

It remains to take the full expectation $\mathbb{E}[\cdot]$ from the above inequality to get (3.7). □

Using this lemma, one can derive several results about the convergence of SGD. In this subsection, we cover the simplest one, but provide more results further.

Theorem 3.1 *Let the objective function f be μ-quasi strongly convex with $\mu > 0$ and L-smooth and Assumption 3.1 holds. Assume that the stepsize γ satisfies $0 < \gamma \leq 1/L$. Then, for each $k \geq 0$ the iterations of* SGD *satisfy*

$$\mathbb{E}\left[\|x^k - x^*\|^2\right] \leq (1 - \gamma\mu)^k \|x^0 - x^*\|^2 + \frac{\gamma\sigma^2}{\mu}. \tag{3.9}$$

Proof Since $0 < \gamma < 1/L$ the last term in the right-hand side of (3.7) is non-positive, i.e., it can be omitted:

$$\mathbb{E}\left[\|x^{k+1} - x^*\|^2\right] \leq (1 - \gamma\mu)\mathbb{E}\left[\|x^k - x^*\|^2\right] + \gamma^2\sigma^2.$$

Unrolling the recurrence, we get the result:

$$\begin{aligned}
\mathbb{E}\left[\|x^{k+1} - x^*\|^2\right] &\leq (1 - \gamma\mu)^{k+1}\|x^0 - x^*\|^2 + \gamma^2\sigma^2 \sum_{t=0}^{k} (1 - \gamma\mu)^t \\
&\leq (1 - \gamma\mu)^{k+1}\|x^0 - x^*\|^2 + \gamma^2\sigma^2 \sum_{t=0}^{\infty} (1 - \gamma\mu)^t \\
&= (1 - \gamma\mu)^{k+1}\|x^0 - x^*\|^2 + \frac{\gamma\sigma^2}{\mu}.
\end{aligned}$$

The above theorem establishes linear convergence of SGD to the neighborhood of the solution. The rate of convergence depends on the stepsize and parameter μ: the larger stepsize is the faster the method converges. However, the size of the neighborhood is proportional to the stepsize. Therefore, there is a certain trade-off: for larger stepsize SGD converges faster but to larger neighborhood of the solution.

Eduard: I think it would be nice to put here some plots in order to illustrate the idea better. For example, we can have 2d-plots like here `https://fa.bianp.net/teaching/2018/COMP-652/stochastic_gradient.html` to illustrate that SGD quickly reaches the neighborhood of the solution, but then it starts to oscillate around the solution without the progress.

Next, we also notice that the neighborhood is proportional to the variance σ^2. Therefore, to reach more accurate solution there are two main options:

1. Use smaller stepsizes (or decrease them using some schedule).
2. Use tighter estimator g^k (with lower variance).

In the first case, the method makes smaller progress at each iteration and, as a result, converges slower but is able to reach tighter approximation of the solution. In Section TODO for Eduard: put the reference to the right section in the future, we discuss in detail some decreasing stepsize policies allowing to achieve the convergence to any predefined accuracy of the solution.

In the second case, one needs to clarify how the tighter estimator is obtained. Usually, it is done via a special trick called *mini-batching*. In the simplest case, mini-batching is a way of choosing g^k such that

$$g^k = \frac{1}{r}\sum_{i=1}^{r} g_i^k,$$

where $g_1^k, \ldots, g_r^k$ are i.i.d. samples from some distribution such that g_i^k satisfies Assumption 3.1 for all $i = 1, \ldots, r$. If g^k is defined this way, we have

$$\begin{aligned}
\mathbb{E}_k\left[g^k\right] &= \frac{1}{r}\sum_{i=1}^{r}\mathbb{E}_k\left[g_i^k\right] = \frac{1}{r}\sum_{i=1}^{r}\nabla f(x^k) = \nabla f(x^k),\\
\mathbb{E}_k\left[\|g^k - \nabla f(x^k)\|^2\right] &= \mathbb{E}_k\left[\left\|\frac{1}{r}\sum_{i=1}^{r}(g_i^k - \nabla f(x^k))\right\|^2\right]\\
&\overset{①}{=} \frac{1}{r^2}\sum_{i=1}^{r}\mathbb{E}_k\left[\|g_i^k - \nabla f(x^k)\|^2\right] \overset{(3.5)}{\leq} \frac{\sigma^2}{r},
\end{aligned}$$

where ① follows from the independence of $g_1^k, \ldots, g_r^k$. That is, mini-batched stochastic gradient with batchsize r has r times smaller variance. Therefore, SGD with such an estimator achieves r times better neighborhood of the solution, but each iteration of the method requires r times more oracle calls than before. However, we emphasize here that the increase of the oracle calls per iteration does not imply that the time needed to execute the step of SGD is increased accordingly. When r is not too large, then it is possible to organize mini-batch computation in a parallel way, e.g., using GPU computations or distributed optimization that we will discuss in Section put here the reference to the Section on parallel optimization.

Although the considered analysis of SGD provides some valuable insights about methods behavior, the results are derived under quite restrictive assumption – Assumption 3.1. As we already noticed before, this assumption implies that the variance is uniformly bounded on the whole space. This is not the case even for simple quadratic optimization problems.

In particular, consider the minimization problem

$$\min_{x\in\mathbb{R}^n}\left\{f(x) = \frac{1}{m}\sum_{i=1}^{m} f_i(x)\right\}, \quad \text{where } f_i(x) = \frac{1}{2}x^\top A_i x + b_i^\top x,$$

where $A_i \in \mathbb{R}^{n\times n}$ are positive semidefinite, and, as in the example with logistic regression, assume that $g^k = \nabla f_{i_k}(x^k)$, where i_k is sampled uniformly at random from $\{1, \ldots, m\}$ independently from the previous iterations. It is easy to check that the quantity

$$\mathbb{E}_k\left[\|g^k - \nabla f(x^k)\|^2\right] = \frac{1}{m}\sum_{i=1}^m \left\| A_i x^k + b_i - \frac{1}{m}\sum_{j=1}^m (A_j x^k + b_j)\right\|^2$$
$$= \frac{1}{m}\sum_{i=1}^m \left\| \frac{1}{m}\sum_{j=1}^m \Big((A_i - A_j)x^k + b_i - b_j\Big)\right\|^2$$

is unbounded unless $A_1 = A_2 = \ldots = A_m$, since x^k can be arbitrarily far from the origin. Therefore, even for such a good (convex, L-smooth) problem, Assumption 3.1 does not hold. To alleviate this issue, a different assumption is needed.

3.1.2 Analysis of SGD: Convex Smooth Stochastic Realizations

As an alternative to Assumption 3.1, consider the following

Assumption 3.2 (Convex Smooth Stochastic Trajectories) Stochastic gradient g^k is computed as $g^k = \nabla f_{\xi^k}(x^k)$, where ξ^k is sampled from some distribution independently from previous iterations. Moreover, there exists a positive constant $L_{\max} > 0$ such that stochastic trajectory $f_\xi(x)$ is convex and $L_{\max}$-smooth almost surely in ξ.

Unlike Assumption 3.1, the above one allows the variance $\mathbb{E}_k\left[\|g^k - \nabla f(x^k)\|^2\right]$ to be unbounded on the whole space and even grow when $\|x^k - x^*\|_2 \to \infty$. However, Assumption 3.1 implies neither convexity nor smoothness of the stochastic trajectories, allowing them to be arbitrary bad unless (3.5) is violated. Therefore, in this particular aspect, Assumption 3.2 is more restrictive than Assumption 3.1.

Below we provide two examples when Assumption 3.2 is satisfied.

Linear Regression

Consider the classical *linear regression* problem:

$$\min_{x\in\mathbb{R}^n} \frac{1}{2m}\|Ax - y\|^2, \tag{3.10}$$

and $A \in \mathbb{R}^{m\times n}$, $y \in \mathbb{R}^n$. For convenience, let us define the rows of matrix A as $a_1, \ldots, a_m \in \mathbb{R}^n$. Then, the above problem can be seen as a finite-sum minimization problem:

$$\min_{x\in\mathbb{R}^n}\left\{ f(x) = \frac{1}{m}\sum_{i=1}^m f_i(x)\right\}, \quad \text{where } f_i(x) = \frac{1}{2}(a_i^\top x - y_i)^2. \tag{3.11}$$

Here each function f_i is convex and L_i-smooth with $L_i = \|a_i\|^2$, since $\nabla^2 f_i(x) = a_i a_i^\top$, which is a positive semi-definite matrix with the largest eigenvalue $\lambda_{\max}(a_i a_i^\top) = \|a_i\|^2$. Therefore, Assumption 3.2 holds with $L_{\max} = \max_{i\in[m]} L_i$ and ξ^k being

sampled uniformly at random from $\{1, \ldots, m\}$ independently from the previous iterations.

Logistic Regression

In Section 3.1.1, we show that logistic regression problem (3.6) with standard stochastic gradient (gradient of the summand picked uniformly at random) fits Assumption 3.1. It turns out that functions $f_i(x)$ are L_i-smooth with $L_i = \|a_i\|^2$. Indeed, using matrix-vector differentiation, one can easily derive that

$$\nabla^2 f_i(x) = \frac{1}{\left(\exp\left(-\frac{1}{2} y_i a_i^\top x\right) + \exp\left(\frac{1}{2} y_i a_i^\top x\right)\right)^2} a_i a_i^\top,$$

which is positive semi-definite. Next, since $t + t^{-1} \geq 2$ for all $t > 0$, we have $\lambda_{\max}(\nabla^2 f_i(x)) \leq \frac{1}{4}\|a_i\|^2$. Therefore, Assumption 3.2 holds with $L_{\max} = \max_{i \in [m]} L_i$ and ξ^k being sampled uniformly at random from $\{1, \ldots, m\}$ independently from the previous iterations.

The analysis of SGD under Assumption 3.2 is quite similar to the one under Assumption 3.1.

Lemma 3.2 *Let the objective function f be μ-quasi strongly convex and L-smooth and Assumption 3.2 holds. Assume that the stepsize γ satisfies $0 < \gamma \leq 1/2L_{\max}$. Then, for each $k \geq 0$ the iterations of* SGD *satisfy*

$$\begin{aligned}\mathbb{E}\left[\|x^{k+1} - x^*\|^2\right] \leq (1 - \gamma\mu)\mathbb{E}\left[\|x^k - x^*\|^2\right] + 2\gamma^2\sigma_*^2 \\ -2\gamma(1 - 2\gamma L_{\max})\mathbb{E}\left[f(x^k) - f(x^*)\right],\end{aligned} \tag{3.12}$$

where $\sigma_*^2 = \mathbb{E}_\xi \|\nabla f_\xi(x^*)\|^2$.

Proof Following the same steps as in the proof of Lemma 3.1, we get

$$\mathbb{E}_k\left[\|x^{k+1} - x^*\|^2\right] = \|x^k - x^*\|^2 - 2\gamma\langle x^k - x^*, \nabla f(x^k)\rangle + \gamma^2 \mathbb{E}_k\left[\|g^k\|^2\right].$$

Next, we apply μ-quasi strong convexity

$$-2\gamma\langle x^k - x^*, \nabla f(x^k)\rangle \leq -2\gamma\left(f(x^k) - f(x^*)\right) - \gamma\mu\|x^k - x^*\|^2$$

and derive

$$\begin{aligned}\mathbb{E}_k\left[\|x^{k+1} - x^*\|^2\right] = (1 - \gamma\mu)\|x^k - x^*\|^2 - 2\gamma\left(f(x^k) - f(x^*)\right) \\ +\gamma^2 \mathbb{E}_k\left[\|g^k\|^2\right].\end{aligned} \tag{3.13}$$

It remains to handle the last term. Assumption 3.2 says that $f_\xi(x)$ is convex and $L_{\max}$-smooth almost surely in ξ. Therefore, using the standard properties of convex smooth functions, we obtain

$$\begin{aligned}
\mathbb{E}_k\left[\|g^k\|^2\right] &= \mathbb{E}_{\xi^k}\left[\|\nabla f_{\xi^k}(x^k) - \nabla f_{\xi^k}(x^*) + \nabla f_{\xi^k}(x^*)\|^2\right] \\
&\leq 2\mathbb{E}_{\xi^k}\left[\|\nabla f_{\xi^k}(x^k) - \nabla f_{\xi^k}(x^*)\|^2\right] + 2\mathbb{E}_{\xi^k}\left[\|\nabla f_{\xi^k}(x^*)\|^2\right] \\
&\overset{\text{(B.10)}}{\leq} 4L_{\max}\mathbb{E}_{\xi^k}\left[f_{\xi^k}(x^k) - f_{\xi^k}(x^*) - \langle \nabla f_{\xi^k}(x^*), x^k - x^*\rangle\right] \\
&\quad +2\mathbb{E}_{\xi}\left[\|\nabla f_{\xi}(x^*)\|^2\right] \\
&= 4L_{\max}\left(f(x^k) - f(x^*) - \langle \nabla f(x^*), x^k - x^*\rangle\right) + 2\sigma_*^2 \\
&= 4L_{\max}\left(f(x^k) - f(x^*)\right) + 2\sigma_*^2, \qquad (3.14)
\end{aligned}$$

where in the fourth step we use $\mathbb{E}_{\xi^k}[f_\xi(x)] = f(x)$ and $\mathbb{E}_\xi[\nabla f_\xi(x)] = \nabla f(x)$. Together with (3.13) the obtained upper bound for $\mathbb{E}_k\left[\|g^k\|^2\right]$ give (3.12). □

This lemma implies several useful facts about the convergence of SGD. In particular, it implies the following

Theorem 3.2 *Let the objective function f be μ-quasi strongly convex with $\mu > 0$ and L-smooth and Assumption 3.2 holds. Assume that the stepsize γ satisfies $0 < \gamma < 1/2L_{\max}$. Then, for each $k \geq 0$ the iterations of* SGD *satisfy*

$$\mathbb{E}\left[\|x^k - x^*\|^2\right] \leq (1-\gamma\mu)^k\|x^0 - x^*\|^2 + \frac{2\gamma\sigma_*^2}{\mu}, \qquad (3.15)$$

where $\sigma_*^2 = \mathbb{E}_\xi\|\nabla f_\xi(x^*)\|^2$.

Proof Given Lemma 3.2, the proof is identical to the proof of Theorem 3.1 up to the following changes: $L \to L_{\max}$ and $\sigma \to \sigma_*$. □

Similarly to Theorem 3.1, the above result establishes linear convergence of SGD to the neighborhood of the solution and similar observations are valid. The main differences can be summarized as follows: 1) the upper bound for γ is smaller in Theorem 3.2, than in Theorem 3.1, and 2) the size of the neighborhood in Theorem 3.2 is proportional to the variance of the stochastic gradient at the solution, while in Theorem 3.1 this size is proportional to the uniform upper bound on the variance $\sigma^2 \geq \sigma_*^2$.

3.1.3 SGD for Finite-Sum Problems and Variance Reduction

Finite-sum problems

$$\min_{x\in\mathbb{R}^n}\left\{f(x) = \frac{1}{m}\sum_{i=1}^m f_i(x)\right\} \qquad (3.16)$$

is an important special case of (3.1): it is sufficient to take ξ as a random integer uniformly distributed on $\{1, \dots, m\}$. Assume that $f_i(x)$ is convex and L_i-smooth for each $i \in [m]$. Then, two examples that we consider in the previous subsection – linear and logistic regression – fall in this setup. Moreover, Assumption 3.2 holds with $L_{\max} = \max_{i \in [m]} L_i$ and Theorem 3.2 implies the following result.

Theorem 3.3 *Let the objective function f be μ-quasi strongly convex with $\mu > 0$ and L-smooth, has a finite-sum form* (3.16)*, and summands f_i be convex and L_i-smooth functions. Assume that the stepsize γ satisfies $0 < \gamma \leq 1/2L_{\max}$, where $L_{\max} = \max_{i \in [m]} L_i$. Then, for each $k \geq 0$ the iterations of* SGD *satisfy*

$$\mathbb{E}\left[\|x^k - x^*\|^2\right] \leq (1 - \gamma\mu)^k \|x^0 - x^*\|^2 + \frac{2\gamma\sigma_*^2}{\mu}, \tag{3.17}$$

where $\sigma_*^2 = \frac{1}{m}\sum_{i=1}^m \|\nabla f_i(x^*)\|^2$.

Although the implications of this result are discussed in the previous subsection (after Theorem 3.2), it is important to write it explicitly for the future comparison with more sophisticated algorithms. So far, we consider a quite straightforward way of generating unbiased estimator g^k for SGD: $g^k = \nabla f_{\xi^k}(x^k)$, where ξ^k is picked uniformly at random from $\{1, \dots, m\}$ independently from previous steps. In particular, when $\gamma = 1/2L_{\max}$ SGD converges to some neighborhood of the solution after $\widetilde{O}(L_{\max}/\mu)$ iterations/oracle calls. In contrast, GD requires $\widetilde{O}(mL/\mu)$ oracle calls to converge to any predefined accuracy of the solution. One can show that $L \leq L_{\max} \leq mL$, meaning that in the worst case there is no benefit in using SGD. However, when $L_{\max} = O(L)$ SGD is m times faster (in terms of the oracle calls) than GD in finding not too accurate approximation of the solution. Therefore, when m is big, the difference between two methods behavior might be significant. However, SGD with constant stepsize does not achieve any predefined accuracy of the solution with linear rate. It leads us to the natural question: *are there any stochastic methods for solving* (3.16) *that have linear convergence to the exact solution asymptotically?*

It turns out that the answer to the above question is positive and the key to construct such a method is in changing the formula for g^k. The general idea is to add something to g^k that is zero in expectation but reduces the variance:

$$g^k = \nabla f_{\xi^k}(x^k) + s^k, \quad \text{where} \quad \mathbb{E}_k[s^k] = 0.$$

Moreover, it is important to have s^k such that it requires a comparable number of oracle calls with $\nabla f_{\xi^k}(x^k)$ needed for computation. Finally, it is crucial to ensure that the variance of estimator g^k reduces during the work of the method. In particular, it implies that $\mathbb{E}[\|g^k\|^2]$ should converge to zero since $\mathbb{E}[\|x^k - x^*\|^2]$ converges to zero. This partially explains, why SGD in all the setups considered above does not converge linearly to the solution: upper bounds (3.8) and (3.14) for $\mathbb{E}_k[\|g^k\|^2]$ contain constant terms proportional either to σ^2 or to σ_*^2.

Taking these observations into account, one can set $s^k = -\nabla f_{\xi^k}(w^k) + \nabla f(w^k)$, where the point w^k is updated from time to time and ξ^k does not depend on w^k.

Then, we have

$$g^k = \nabla f_{\xi^k}(x^k) - \nabla f_{\xi^k}(w^k) + \nabla f(w^k), \tag{3.18}$$
$$\mathbb{E}_k[g^k] = \frac{1}{m}\sum_{i=1}^{m}\left(\nabla f_i(x^k) - \nabla f_i(w^k) + \nabla f(w^k)\right) = \nabla f(x^k),$$

i.e., unbiasedness is satisfied. Next, g^k should require a comparable number of oracle calls with $\nabla f_{\xi^k}(x^k)$ needed for computation. We make the following observation: if $w^k = w^{k-1}$, then $\nabla f(w^k)$ can be taken from the previous iteration, i.e., full gradient computations are needed iff w^k is updated. Therefore, w^k should be updated rarely, e.g., one can set w^{k+1} as x^k once per every $b \sim m$ iterations:

$$w^{k+1} = \begin{cases} w^k, & \text{if } k+1 \mod b \neq 0, \\ x^{k+1}, & \text{if } k+1 \mod b = 0. \end{cases} \tag{3.19}$$

The resulting method, i.e., SGD (3.3) with g^k defined in (3.18) and w^k given in (3.19), is called Stochastic Variance Reduced Gradient (SVRG) and the parameter b is usually called inner loop size. The oracle cost of b subsequent steps of SVRG equals $2 \cdot (b-1) + (m+2) \cdot 1 = 2b + m$, while for SGD it equals b. Therefore, when $b \geq m$, the oracle costs of running SVRG and SGD for $K \geq m$ iterations are the same (up to numerical factor of 3).

Finally, the choice of w^k from (3.19) is promising for the following reason. Assume that we proved that the iterates of SVRG satisfy $\mathbb{E}[\|x^k - x^*\|^2] \to 0$. Then, we also have that $\mathbb{E}[\|w^k - x^*\|^2] \to 0$ and $\mathbb{E}[\|w^k - x^k\|^2] \to 0$ due to (3.19). Since $\nabla f_i(x)$ is Lipschitz continuous, we have $\mathbb{E}[\|\nabla f_i(x^k) - \nabla f_i(w^k)\|^2] \to 0$ for all $i \in [m]$ and $\mathbb{E}[\|\nabla f(w^k)\|^2] \to 0$. That is why $\mathbb{E}[\|g^k\|^2]$ converges to 0 as well, i.e., the variance of the estimator reduces during the work of the method.

It remains to formalize the observations enough to get a rigorous proof of convergence. However, for simplicity, we will analyze a slightly different method called Loopless Stochastic Variance Reduced Gradient (L-SVRG). The key difference between L-SVRG and SVRG is in the choice of w^k: while SVRG updates w^k once per every $b \sim m$ iterations, L-SVRG does it at each iteration with probability $p \sim 1/m$, which is typically small. In other words, instead of (3.19) L-SVRG relies on the following update rule for w^k:

$$w^{k+1} = \begin{cases} w^k, & \text{with probability } 1-p, \\ x^k, & \text{with probability } p. \end{cases} \tag{3.20}$$

L-SVRG can be seen as SVRG with random inner loop sizes having geometrical distribution with parameter p. If $p \sim 1/m$, all intuition regarding SVRG is valid for L-SVRG as well. In particular, when $p = 1/m$, the expected oracle cost of 1 iteration of L-SVRG equals $2 \cdot (1-p) + (m+2) \cdot p = 2 + mp = 3$, which is comparable with the oracle cost of the one step of SGD.

As we mention before, it is crucial to derive good upper bound for $\mathbb{E}_k[\|g^k\|^2]$ that does not contain constant terms like σ^2 or σ_*^2 from the analysis of SGD. The following lemma provides such an upper bound.

Lemma 3.3 *Let f_i be convex and L_i-smooth for all $i \in [m]$. Then, for all $k \geq 0$ the iterates produced by* L-SVRG *satisfy*

$$\mathbb{E}_k\left[\|g^k\|^2\right] \leq 4L_{\max}\left(f(x^k) - f(x^*)\right) + 2\sigma_k^2, \tag{3.21}$$

where $\sigma_k^2 = \frac{1}{m}\sum_{i=1}^m \|\nabla f_i(w^k) - \nabla f_i(x^*)\|^2$.

Proof Using the definition of g^k (3.18) and a well-known fact that $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ for all $a, b \in \mathbb{R}^n$, we derive

$$\begin{aligned}
\mathbb{E}_k\left[\|g^k\|^2\right] &= \mathbb{E}_{\xi^k}\left[\|\nabla f_{\xi^k}(x^k) - \nabla f_{\xi^k}(w^k) + \nabla f(w^k)\|^2\right] \\
&= \frac{1}{m}\sum_{i=1}^m \|\nabla f_i(x^k) - \nabla f_i(w^k) + \nabla f(w^k)\|^2 \\
&= \frac{1}{m}\sum_{i=1}^m \|\nabla f_i(x^k) - \nabla f_i(x^*) + \nabla f_i(x^*) - \nabla f_i(w^k) + \nabla f(w^k)\|^2 \\
&\leq \frac{2}{m}\sum_{i=1}^m \|\nabla f_i(x^k) - \nabla f_i(x^*)\|^2 \\
&\quad + \frac{2}{m}\sum_{i=1}^m \|\nabla f_i(w^k) - \nabla f_i(x^*) - \nabla f(w^k)\|^2.
\end{aligned}$$

To bound the first term in the right-hand side of the above inequality, we apply (B.10):

$$\begin{aligned}
\mathbb{E}_k\left[\|g^k\|^2\right] &\leq \frac{4}{m}\sum_{i=1}^m L_i\left(f_i(x^k) - f_i(x^*) - \langle \nabla f_i(x^k), x^k - x^*\rangle\right) \\
&\quad + \frac{2}{m}\sum_{i=1}^m \|\nabla f_i(w^k) - \nabla f_i(x^*) - \nabla f(w^k)\|^2 \\
&\leq 4L_{\max}\left(f(x^k) - f(x^*)\right) \qquad (3.22) \\
&\quad + \frac{2}{m}\sum_{i=1}^m \|\nabla f_i(w^k) - \nabla f_i(x^*) - \nabla f(w^k)\|^2. \qquad (3.23)
\end{aligned}$$

Next, the last term can be considered as the variance of $\nabla f_\xi(w^k) - \nabla f_\xi(x^*)$ for fixed w^k and ξ having an uniform distribution on $\{1, 2, \ldots, m\}$. Since the variance is not bigger than the second moment, we can continue our derivation as

$$\mathbb{E}\left[\|g^k\|^2 \mid x^k\right] \leq 4L_{\max}\left(f(x^k) - f(x^*)\right) + \frac{2}{m}\sum_{i=1}^m \|\nabla f_i(w^k) - \nabla f_i(x^*)\|^2.$$

Taking into account that $\sigma_k^2 = \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(w^k) - \nabla f_i(x^*)\|^2$, we get the result. □

Next, it is important to show that σ_k^2 converges (in expectation) to zero. An important step to achieve this goal is in deriving a good upper bound for $\mathbb{E}_k[\sigma_{k+1}^2]$.

Lemma 3.4 *Let f_i be convex and L_i-smooth for all $i \in [m]$. Then, for all $k \geq 0$ the iterates produced by* L-SVRG *satisfy*

$$\mathbb{E}_k\left[\sigma_{k+1}^2\right] \leq (1-p)\sigma_k^2 + 2pL_{\max}\left(f(x^k) - f(x^*)\right), \tag{3.24}$$

where $\sigma_k^2 = \frac{1}{m}\sum_{i=1}^m \|\nabla f_i(w^k) - \nabla f_i(x^*)\|^2$.

Proof By definition of w^k (3.20), we have

$$\begin{aligned}
\mathbb{E}_k\left[\sigma_{k+1}^2\right] &= \frac{1}{m}\sum_{i=1}^m \mathbb{E}_k\left[\|\nabla f_i(w^{k+1}) - \nabla f_i(x^*)\|^2\right] \\
&= \frac{1-p}{n}\sum_{i=1}^n \|\nabla f_i(w^k) - \nabla f_i(x^*)\|^2 \\
&\quad + \frac{p}{m}\sum_{i=1}^m \|\nabla f_i(x^k) - \nabla f_i(x^*)\|^2 \\
&= (1-p)\sigma_k^2 + \frac{p}{m}\sum_{i=1}^m \|\nabla f_i(x^k) - \nabla f_i(x^*)\|^2.
\end{aligned}$$

Applying (B.10) to the above inequality, we derive

$$\begin{aligned}
\mathbb{E}_k\left[\sigma_{k+1}^2\right] &\leq (1-p)\sigma_k^2 + \frac{p}{m}\sum_{i=1}^m 2L_i\left(f_i(x^k) - f_i(x^*) - \langle\nabla f_i(x^k), x^k - x^*\rangle\right) \\
&\leq (1-p)\sigma_k^2 + 2pL_{\max}\left(f(x^k) - f(x^*)\right),
\end{aligned} \tag{3.25}$$

which concludes the proof. □

Finally, it remains to combine the bounds from Lemmas 3.3 and 3.4.

Theorem 3.4 *Let the objective function f be μ-quasi strongly convex with $\mu > 0$ and L-smooth, has a finite-sum form* (3.16)*, and summands f_i be convex and L_i-smooth functions. Assume that the stepsize γ satisfies $0 < \gamma \leq 1/6L_{\max}$, where $L_{\max} = \max_{i\in[m]} L_i$. Then, for each $k \geq 0$ the iterations of* L-SVRG *satisfy*

$$\mathbb{E}[V_k] \leq \left(1 - \min\left\{\gamma\mu, \frac{p}{2}\right\}\right)^k V_0, \tag{3.26}$$

where $V_k = \|x^k - x^*\|^2 + \frac{4\gamma^2}{p}\sigma_k^2$ and $\sigma_k^2 = \frac{1}{m}\sum_{i=1}^m \|\nabla f_i(w^k) - \nabla f_i(x^*)\|^2$.

Proof We start with noticing that inequality (3.13) is derived for SGD based on three facts: (1) $x^{k+1} = x^k - \gamma g^k$, (2) g^k is an unbiased estimate of $\nabla f(x^k)$ for fixed x^k, and (3) function f is μ-quasi strongly convex. These three properties hold in the setup we consider right now. Therefore, we have

$$\begin{aligned}\mathbb{E}_k\left[\|x^{k+1} - x^*\|^2\right] &\overset{(3.13)}{\leq} (1-\gamma\mu)\|x^k - x^*\|^2 - 2\gamma\left(f(x^k) - f(x^*)\right)\\ &\quad +\gamma^2\mathbb{E}_k\left[\|g^k\|^2\right].\end{aligned}$$

Next, we apply Lemma 3.3 to upper bound $\mathbb{E}_k\left[\|g^k\|^2\right]$ and rearrange the terms:

$$\begin{aligned}\mathbb{E}_k\left[\|x^{k+1} - x^*\|^2\right] &\overset{(3.21)}{\leq} (1-\gamma\mu)\|x^k - x^*\|^2 - 2\gamma\left(f(x^k) - f(x^*)\right)\\ &\quad +\gamma^2\left(4L_{\max}\left(f(x^k) - f(x^*)\right) + 2\sigma_k^2\right)\\ &= (1-\gamma\mu)\|x^k - x^*\|^2 + 2\gamma^2\sigma_k^2\\ &\quad -2\gamma\left(1 - 2\gamma L_{\max}\right)\left(f(x^k) - f(x^*)\right).\end{aligned}$$

Summing up the above inequality with $\frac{4\gamma^2}{p}$-multiple of (3.24), we get

$$\begin{aligned}\mathbb{E}_k\Bigg[\underbrace{\|x^{k+1} - x^*\|^2 + \frac{4\gamma^2}{p}\sigma_{k+1}^2}_{V_{k+1}}\Bigg] &\leq (1-\gamma\mu)\|x^k - x^*\|^2 + 2\gamma^2\sigma_k^2\\ &\quad -2\gamma\left(1 - 2\gamma L_{\max}\right)\left(f(x^k) - f(x^*)\right)\\ &\quad +\frac{4\gamma^2}{p}\left((1-p)\sigma_k^2 + 2pL_{\max}\left(f(x^k) - f(x^*)\right)\right)\\ &= (1-\gamma\mu)\|x^k - x^*\|^2 + \left(1 - \frac{p}{2}\right)\frac{4\gamma^2}{p}\sigma_k^2\\ &\quad -2\gamma\left(1 - 6\gamma L_{\max}\right)\left(f(x^k) - f(x^*)\right)\\ &\leq \left(1 - \min\left\{\gamma\mu, \frac{p}{2}\right\}\right)\Bigg(\underbrace{\|x^k - x^*\|^2 + \frac{4\gamma^2}{p}\sigma_k^2}_{V_k}\Bigg),\end{aligned}$$

where in the last step we use $0 < \gamma \leq 1/6L_{\max} \Rightarrow 2\gamma\left(1 - 6\gamma L_{\max}\right)\left(f(x^k) - f(x^*)\right) \geq 0$. Finally, we take the full expectation from the obtained inequality, unroll the recurrence, and get (3.26). □

Let us discuss the obtained result and compare it with the result for SGD (Theorem 3.3). First of all, unlike the convergence guarantee for SGD, the guarantee for L-SVRG is given in terms of the Lyapunov function $V_k = \|x^k - x^*\|^2 + \frac{4\gamma^2}{p}\sigma_k^2$

(in expectation). Of course, since $\sigma_k^2 \geq 0$, inequality (3.26) gives an upper bound for $\mathbb{E}[\|x^k - x^*\|^2]$ as well. However, it also implies that $\mathbb{E}[\sigma_k^2]$ converges to zero supporting the intuition behind the method and this quantity, in particular. Moreover, one could guess the form of the Lyapunov function V_k from the discussion around Lemmas 3.3 and 3.4.

Next, perhaps, the most important difference between SGD and LSVRG is that the latter one *converges linearly to the exact solution* (asymptotically, in expectation) unlike SGD that converges linearly only to some neighborhood of the solution. To illustrate this phenomenon better, we consider the following example.

L-SVRG vs SGD: Solving Logistic Regression with ℓ_2-Regularization

TODO for Eduard: write about the details.

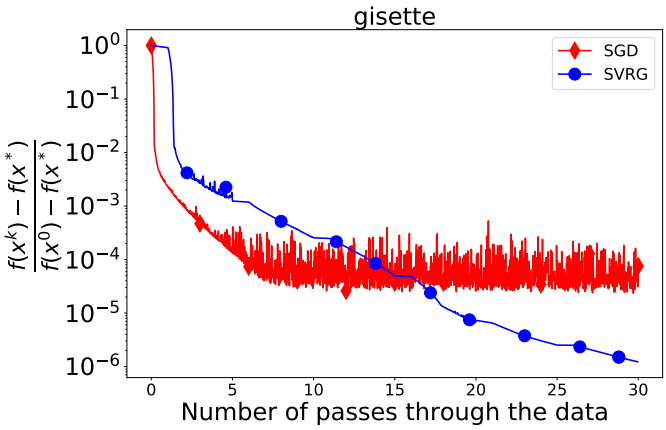

Fig. 3.1: TODO for Eduard: replace this figure by the polished one. Write a caption.

Finally, it remains to derive an explicit complexity upper bound for L-SVRG.

Corollary 3.1 *Let the assumptions of Theorem 3.4 hold. Assume that* $\gamma = 1/6L_{\max}$ *and* $p = 1/m$. *Then, to ensure* $\mathbb{E}[\|x^k - x^*\|^2] \leq \varepsilon$, L-SVRG *requires*

$$O\left(\left(m + \frac{L_{\max}}{\mu}\right) \log \frac{V_0}{\varepsilon}\right) \tag{3.27}$$

oracle calls in expectation.

Proof When $\gamma = 1/6L_{\max}$ and $p = 1/m$, Theorem 3.4 implies

$$\begin{aligned}\mathbb{E}[\|x^k - x^*\|^2] \leq \mathbb{E}\left[V_k\right] &\leq \left(1 - \min\left\{\gamma\mu, \frac{p}{2}\right\}\right)^k V_0 \\ &= \left(1 - \min\left\{\frac{\mu}{6L_{\max}}, \frac{1}{2m}\right\}\right)^k V_0 \\ &\leq \exp\left(-\min\left\{\frac{\mu}{6L_{\max}}, \frac{1}{2m}\right\} k\right) V_0.\end{aligned}$$

Therefore, after

$$k = O\left(\max\left\{\frac{L_{\max}}{\mu}, m\right\} \log\frac{V_0}{\varepsilon}\right) = O\left(\left(m + \frac{L_{\max}}{\mu}\right) \log\frac{V_0}{\varepsilon}\right)$$

iteration L-SVRG guarantees that $\mathbb{E}[\|x^k - x^*\|^2] \leq \varepsilon$. Taking into account that the expected total number of oracle calls equals $m + O(k)$, we get the result. □

In contrast, GD has $O\left(m\frac{L}{\mu}\log\frac{\|x^0 - x^*\|^2}{\varepsilon}\right)$ oracle complexity, where L is the smoothness constant of function f. Since $L \leq L_{\max} \leq mL$, in the worst case, L-SVRG has the same oracle complexity as GD up to the differences in the logarithmic factor. However, when $L_{\max} = O(L)$, the complexity of L-SVRG is much better than the one of GD, especially when m is large.

It is worth mentioning that the similar result holds for SVRG as well. Moreover, there exists another popular variance reduced method called SAGA. Instead of storing one vector, SAGA stores m vectors $\{\nabla f_i(w_i^k)\}_{i=1}^m$ corresponding to the gradients of summands at the points where they were computed last time before the iteration k. More formally, SAGA has the same update rule as SGD (3.3) with

$$g^k = \nabla f_{\xi^k}(x^k) - \nabla f_{\xi^k}(w_i^k) + \frac{1}{m}\sum_{i=1}^m \nabla f_i(w_i^k), \tag{3.28}$$

$$w_i^0 = x^0, \quad w_i^{k+1} = \begin{cases} x^k, & \text{if } i = \xi^k, \\ w_i^k, & \text{if } i \neq \xi^k, \end{cases} \quad i = 1, \ldots, m, \tag{3.29}$$

where ξ^k is sampled uniformly at random from $\{1, \ldots, m\}$ independently from previous steps. One can easily show that $\mathbb{E}_k[g^k] = \nabla f(x^k)$ in this case as well. Moreover, the convergence guarantees for SAGA are identical to the ones for L-SVRG and even can be proven using similar steps. However, SAGA and L-SVRG have some important difference. Unlike SVRG/L-SVRG, SAGA does not require the computation of the full gradient during its work (although one needs to compute initial vectors $\{\nabla f_i(w_i^0)\}_{i=1}^m$). In this aspect, SAGA is better than SVRG/L-SVRG but it has a disadvantage as well: one needs to store m vectors $\{\nabla f_i(w_i^k)\}_{i=1}^m$ to use the method. When m and n are large, the storage of m vectors might be infeasible.

3.1.4 Unified Analysis of SGD

The proofs from the previous subsections have a lot of similarities: in all of them, we apply simple facts based on the recursion $x^{k+1} = x^k - \gamma g^k$, derive upper bounds for $\mathbb{E}_k[\|g^k\|^2]$, and in some cases, derive another auxiliary inequality to handle the variance terms (like in the proof for L-SVRG). Therefore, for the ease of further discussion and establishing the connections between different methods, in this subsection, we focus on an unified analysis of SGD.

First of all, we will consider a more general class of problems – composite/regularized optimization problems:

$$\min_{x \in \mathbb{R}^n} \{f_h(x) = f(x) + h(x)\}, \tag{3.30}$$

where function $f(x) : \mathbb{R}^n \to \mathbb{R}$ is (μ, x^*)-quasi strongly convex and L-smooth, x^* is a minimizer of f_h, and function $h(x) : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a proper lower semicontinuous convex function called regularization/composite term. Moreover, function h is assumed to have simple structure, i.e., the proximal operator associated with h

$$\text{prox}_h(x) = \operatorname*{argmin}_{y \in \mathbb{R}^n} \left\{ h(y) + \frac{1}{2}\|y - x\|^2 \right\} \tag{3.31}$$

should be easy to compute for any $x \in \mathbb{R}^n$. The problems of this type are quite general, e.g., they include constrained optimization problems and problems with sparsity-inducing penalties (see the details in Appendix TODO: add the reference to the necessary section).

TODO: we should add a section in the appendix about composite minimization problems and proximal operators.

As before, we assume that function f has either an expectation

$$f(x) = \mathbb{E}_{\xi \sim \mathcal{D}}[f_\xi(x)] \tag{3.32}$$

or a finite-sum form

$$f(x) = \frac{1}{m} \sum_{i=1}^{m} f_i(x), \tag{3.33}$$

and is accessible through the unbiased estimate g^k of the full gradient: $\mathbb{E}_k[g^k] = \nabla f(x^k)$. Next, SGD in this case has slightly different update rule:

$$x^{k+1} = \text{prox}_{\gamma_k h}\left(x^k - \gamma_k g^k\right). \tag{3.34}$$

The above update rule takes into account the information about function f through the usage of g^k and about function h – through the computing proximal operator from the right-hand side of the update rule (3.3) considered previously. The method from (3.34) is usually called Prox-SGD, but for simplicity we will refer to this method as SGD.

Now, we are ready to introduce a general parametric assumption for the unified analysis of SGD.

Assumption 3.3 (Unified parametric assumption) Let $\{x^k\}_{k\geq 0}$ be the iterates generated by SGD from (3.34). We assume that for all $k \geq 0$ estimator g^k is unbiased estimate of the gradient: $\mathbb{E}_k[g^k] = \nabla f(x^k)$. Moreover, we assume that there exist non-negative constants $A, B, C, D_1, D_2 \geq 0$, $\rho \in (0, 1]$, and a (possibly) random sequence $\{\sigma_k^2\}_{k\geq 0}$ such that σ_k can only depend on the randomness from iterations $0, \ldots, k-1$ and the following two inequalities hold for all $k \geq 0$:

$$\mathbb{E}_k\left[\|g^k - \nabla f(x^*)\|^2\right] \leq 2AV_f(x^k, x^*) + B\sigma_k^2 + D_1, \tag{3.35}$$

$$\mathbb{E}_k\left[\sigma_{k+1}^2\right] \leq (1-\rho)\sigma_k^2 + 2CV_f(x^k, x^*) + D_2, \tag{3.36}$$

where $V_f(x, y) = f(x) - f(y) - \langle \nabla f(y), x - y\rangle$ is a Bregman divergence associated with function f.

As we will see next, this assumptions covers all the setups considered above in this chapter and even more. First of all, let us discuss inequality (3.35). It can be seen as a unification of the typical bounds for $\mathbb{E}_k[\|g^k\|^2]$ arising in the analysis of standard SGD (3.8), (3.14) and L-SVRG (3.21) for the special case of (3.30) with $h(x) \equiv 0$. However, in the general case, one has to take into account that $\nabla f(x^*) \neq 0$, which is reflected in the both sides of inequality (3.35). Next, inequality (3.36) describes a variance reduction process similarly to (3.24) for L-SVRG.

From these connections with the inequalities derived earlier in this chapter one can conjecture that constants A and C are related to the smoothness properties of the problem, D_1 and D_2 are some noises that are not handed by variance reduction mechanism, B is typically a numerical constant, and ρ is the "rate" of the variance reduction. In fact, this is a quite tight description of the parameters from Assumption 3.3. Before we move on to the analysis under this assumption, we provide formal proofs that SGD under Assumptions 3.1, 3.2 and L-SVRG fit it.

SGD under Uniformly Bounded Variance Assumption

Consider the case when f is convex, L-smooth, and Assumption 3.1 holds. Then, following exactly the same steps as in the proof of (3.8), we derive

$$\begin{aligned}\mathbb{E}_k\left[\|g^k - \nabla f(x^*)\|^2\right] &= \|\nabla f(x^k) - \nabla f(x^*)\|^2 + \mathbb{E}_k\left[\|g^k - \nabla f(x^k)\|^2\right] \\ &\overset{(3.5)}{\leq} \|\nabla f(x^k) - \nabla f(x^*)\|^2 + \sigma^2 \\ &\overset{(B.10)}{\leq} 2LV_f(x^k, x^*) + \sigma^2.\end{aligned}$$

That is, in this case, SGD satisfies Assumption 3.3 with the following parameters:

$$A = L, \quad B = 0, \quad C = 0, \quad D_1 = \sigma^2, \quad \rho = 1, \quad \sigma_k^2 \equiv 0, \quad D_2 = 0. \tag{3.37}$$

SGD under Convex Smooth Stochastic Trajectories Assumption

Let f be convex, L-smooth, and let Assumption 3.2 hold. Then, following exactly the same steps as in the proof of (3.14), we derive

$$\begin{aligned}
\mathbb{E}_k\left[\|g^k - \nabla f(x^*)\|^2\right] &= \mathbb{E}_{\xi^k}\left[\|\nabla f_{\xi^k}(x^k) - \nabla f_{\xi^k}(x^*) + \nabla f_{\xi^k}(x^*) - \nabla f(x^*)\|^2\right] \\
&\leq 2\mathbb{E}_{\xi^k}\left[\|\nabla f_{\xi^k}(x^k) - \nabla f_{\xi^k}(x^*)\|^2\right] \\
&\quad +2\mathbb{E}_{\xi^k}\left[\|\nabla f_{\xi^k}(x^*) - \nabla f(x^*)\|^2\right] \\
&\overset{(B.10)}{\leq} 4L_{\max}\mathbb{E}_{\xi^k}\left[f_{\xi^k}(x^k) - f_{\xi^k}(x^*) - \langle \nabla f_{\xi^k}(x^*), x^k - x^*\rangle\right] \\
&\quad +2\mathbb{E}_{\xi}\left[\|\nabla f_{\xi}(x^*) - \nabla f(x^*)\|^2\right] \\
&= 4L_{\max}\left(f(x^k) - f(x^*) - \langle \nabla f(x^*), x^k - x^*\rangle\right) + 2\sigma_*^2 \\
&= 4L_{\max}V_f(x^k, x^*) + 2\sigma_*^2,
\end{aligned}$$

where $\sigma_*^2 = \mathbb{E}_{\xi}[\|\nabla f_{\xi}(x^*) - \nabla f(x^*)\|^2]$. That is, in this case, SGD satisfies Assumption 3.3 with the following parameters:

$$A = 2L_{\max}, \quad B = 0, \quad C = 0, \quad D_1 = 2\sigma_*^2, \quad \rho = 1, \quad \sigma_k^2 \equiv 0, \quad D_2 = 0. \tag{3.38}$$

L-SVRG for the Finite Sums of Smooth Convex Functions

Assume that f has a finite sum structure (3.33) and f_i is convex and L_i-smooth for all $i = 1, \ldots, m$. Then, to show that inequalities (3.35) and (3.36) hold for L-SVRG in this case one can apply almost the same sequence of steps as in the proofs of Lemmas 3.3 and 3.4. In particular, estimating $\mathbb{E}_k[\|g^k - \nabla f(x^*)\|^2]$ instead of $\mathbb{E}_k[\|g^k\|^2]$ one will get $V_f(x^k, x^*)$ instead of $f(x^k) - f(x^*)$ in (3.22) (since $\frac{1}{m}\sum_{i=1}^m \nabla f_i(x^*) = \nabla f(x^*) \neq 0$) and $\frac{2}{m}\sum_{i=1}^m \|\nabla f_i(w^k) - \nabla f_i(x^*) - (\nabla f(w^k) - \nabla f(x^*))\|^2$ instead of $\frac{2}{m}\sum_{i=1}^m \|\nabla f_i(w^k) - \nabla f_i(x^*) - \nabla f(w^k)\|^2$ in (3.23). Next, the proof of inequality

$$\mathbb{E}_k\left[\sigma_{k+1}^2\right] \leq (1-p)\sigma_k^2 + 2pL_{\max}V_f(x^k, x^*)$$

is identical to the proof of (3.24) up to the following change: in (3.25), one should take into account that $\frac{1}{m}\sum_{i=1}^m \nabla f_i(x^*) = \nabla f(x^*) \neq 0$.

Therefore, in this setup, L-SVRG satisfies Assumption 3.3 with the following parameters:

$$\begin{gathered}
A = 2L_{\max}, \quad B = 2, \quad \sigma_k^2 = \frac{1}{m}\sum_{i=1}^m \|\nabla f_i(w^k) - \nabla f_i(x^*)\|^2, \quad D_1 = 0, \\
C = pL_{\max}, \quad \rho = p, \quad D_2 = 0.
\end{gathered} \tag{3.39}$$

Under Assumption 3.3 and μ-quasi strong convexity of f the following result holds.

Theorem 3.5 *Let the objective function f be (μ, x^*)-quasi strongly convex with $\mu > 0$, $V_f(x, x^*) \geq 0$ for all[1] $x \in \mathbb{R}^n$, and Assumption 3.3 hold. Assume that the stepsize $\gamma_k = \gamma$ satisfies*

$$0 < \gamma \leq \min\left\{\frac{1}{\mu}, \frac{1}{A + CM}\right\}, \tag{3.40}$$

where $M > B/\rho$. Then, for each $k \geq 0$ the iterations of the method from (3.34) *satisfy*

$$\mathbb{E}[V_k] \leq \left(1 - \min\left\{\gamma\mu, \rho - \frac{B}{M}\right\}\right)^k \mathbb{E}[V_0] + \frac{\gamma^2(D_1 + MD_2)}{\min\left\{\gamma\mu, \rho - \frac{B}{M}\right\}}, \tag{3.41}$$

where $V_k = \|x^k - x^*\|^2 + M\gamma^2\sigma_k^2$.

Proof Due to non-expansiveness of proximal operator and fixed point property $x^* = \text{prox}_{\gamma h}(x^* - \gamma\nabla f(x^*))$ we have TODO: add here the reference to the appendix with properties of prox-operator

$$\begin{aligned}
\|x^{k+1} - x^*\|^2 &= \left\|\text{prox}_{\gamma h}(x^k - \gamma g^k) - \text{prox}_{\gamma h}(x^* - \gamma\nabla f(x^*))\right\|^2 \\
&\leq \|x^k - \gamma g^k - (x^* - \gamma\nabla f(x^*))\|^2 \\
&= \|x^k - x^*\|^2 - 2\gamma\langle x^k - x^*, g^k - \nabla f(x^*)\rangle + \gamma^2\|g^k - \nabla f(x^*)\|^2.
\end{aligned}$$

Next, we take conditional expectation from the above inequality and apply (3.35) from Assumption 3.3:

$$\begin{aligned}
\mathbb{E}_k\left[\|x^{k+1} - x^*\|^2\right] &= \|x^k - x^*\|^2 - 2\gamma\mathbb{E}_k\left[\langle x^k - x^*, g^k - \nabla f(x^*)\rangle\right] \\
&\quad + \gamma^2\mathbb{E}_k\left[\|g^k - \nabla f(x^*)\|^2\right] \\
&= \|x^k - x^*\|^2 - 2\gamma\langle x^k - x^*, \nabla f(x^k) - \nabla f(x^*)\rangle \\
&\quad + \gamma^2\mathbb{E}_k\left[\|g^k - \nabla f(x^*)\|^2\right] \\
&\overset{(3.35)}{\leq} \|x^k - x^*\|^2 - 2\gamma\langle x^k - x^*, \nabla f(x^k) - \nabla f(x^*)\rangle \\
&\quad + \gamma^2\left(2AV_f(x^k, x^*) + B\sigma_k^2 + D_1\right).
\end{aligned}$$

Function f is (μ, x^*)-quasi strongly convex. This implies that

$$\langle x^k - x^*, \nabla f(x^k) - \nabla f(x^*)\rangle \geq V_f(x^k, x^*) + \frac{\mu}{2}\|x^k - x^*\|^2.$$

Therefore, we have

[1] To guarantee $V_f(x, x^*) \geq 0$ it is sufficient to assume convexity of f.

$$\mathbb{E}_k\left[\|x^{k+1}-x^*\|^2\right] \leq (1-\gamma\mu)\|x^k-x^*\|^2 - 2\gamma(1-A\gamma)V_f(x^k,x^*) + \gamma^2 B\sigma_k^2 + \gamma^2 D_1.$$

Summing up the above inequality with $M\gamma^2$-multiple of (3.36) and using $V_k = \|x^k-x^*\|^2 + M\gamma^2\sigma_k^2$, we derive

$$\begin{aligned}
\mathbb{E}_k[V_{k+1}] &= \mathbb{E}_k\left[\|x^{k+1}-x^*\|^2 + M\gamma^2\sigma_{k+1}^2\right] \\
&\overset{(3.36)}{\leq} (1-\gamma\mu)\|x^k-x^*\|^2 - 2\gamma(1-A\gamma)V_f(x^k,x^*) \\
&\quad + \gamma^2 B\sigma_k^2 + \gamma^2 D_1 + M\gamma^2\left((1-\rho)\sigma_k^2 + 2CV_f(x^k,x^*) + D_2\right) \\
&= (1-\gamma\mu)\|x^k-x^*\|^2 + M\gamma^2\left(1-\rho+\frac{B}{M}\right)\sigma_k^2 \\
&\quad - 2\gamma(1-\gamma(A+CM))V_f(x^k,x^*) + \gamma^2(D_1+MD_2) \qquad (3.42) \\
&\leq \left(1-\min\left\{\gamma\mu, \rho-\frac{B}{M}\right\}\right)\underbrace{\left(\|x^k-x^*\|^2 + M\gamma^2\sigma_k^2\right)}_{V_k} + \gamma^2(D_1+MD_2),
\end{aligned}$$

where in the last step we use $0 < \gamma \leq 1/(A+CM)$ and $V_f(x^k,x^*) \geq 0$. Next, we take the full expectation

$$\mathbb{E}[V_{k+1}] \leq \left(1-\min\left\{\gamma\mu, \rho-\frac{B}{M}\right\}\right)\mathbb{E}[V_k] + \gamma^2(D_1+MD_2) \qquad (3.43)$$

and unroll the obtained recurrence

$$\begin{aligned}
\mathbb{E}[V_k] &\leq \left(1-\min\left\{\gamma\mu, \rho-\frac{B}{M}\right\}\right)^k \mathbb{E}[V_0] \\
&\quad + \gamma^2(D_1+MD_2)\sum_{t=0}^{k-1}\left(1-\min\left\{\gamma\mu, \rho-\frac{B}{M}\right\}\right)^t \\
&\leq \left(1-\min\left\{\gamma\mu, \rho-\frac{B}{M}\right\}\right)^k \mathbb{E}[V_0] \\
&\quad + \gamma^2(D_1+MD_2)\sum_{t=0}^{\infty}\left(1-\min\left\{\gamma\mu, \rho-\frac{B}{M}\right\}\right)^t \\
&= \left(1-\min\left\{\gamma\mu, \rho-\frac{B}{M}\right\}\right)^k \mathbb{E}[V_0] + \frac{\gamma^2(D_1+MD_2)}{\min\left\{\gamma\mu, \rho-\frac{B}{M}\right\}}.
\end{aligned}$$

which finishes the proof. □

As in the previous results of this Chapter, the above theorem establishes linear convergence of the method to some neighborhood that depends on the parameters from Assumption 3.3, parameter μ, and stepsize γ. Taking into account the examples

given before Theorem 3.5, one can easily notice that the above theorem recovers the results for SGD and L-SVRG derived in Sections 3.1.1-3.1.3. Moreover, as we will see further in this Chapter, this approach recovers many other results for SGD-like methods.

However, before we start our consideration of special cases, it is useful to derive another general result – for the case when $\mu = 0$. Although it is possible to derive the result in the same generality as we have for the case when $\mu > 0$, in the next theorem, for simplicity we assume that the composite term is zero, i.e., $R(x) \equiv 0$.

Theorem 3.6 *Let the objective function f be (μ, x^*)-quasi strongly convex with $\mu \geq 0$, $R(x) \equiv 0$, and Assumption 3.3 hold. Assume that the stepsize $\gamma_k = \gamma$ satisfies*

$$0 < \gamma \leq \min\left\{\frac{1}{\mu}, \frac{1}{2(A+CM)}\right\}, \tag{3.44}$$

where $M > B/\rho$. Then, for each $k \geq 0$ the iterations of the method from (3.34) *satisfy*

$$\mathbb{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \frac{\mathbb{E}[V_0]}{\gamma W_K} + \gamma(D_1 + MD_2), \tag{3.45}$$

where $\overline{x}^K = \frac{1}{W_K}\sum_{k=0}^K w_k x^k$, $w_k = (1-\eta)^{-(k+1)}$, $W_K = \sum_{k=0}^K w_k$, $\eta = \min\left\{\gamma\mu, \rho - \frac{B}{M}\right\}$, $V_k = \|x^k - x^*\|^2 + M\gamma^2\sigma_k^2$. In particular,

$$\mathbb{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq (1-\eta)^{K+1}\frac{\mathbb{E}[V_0]}{\gamma} + \gamma(D_1 + MD_2), \text{ when } \mu > 0, \tag{3.46}$$

$$\mathbb{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \frac{\mathbb{E}[V_0]}{\gamma(K+1)} + \gamma(D_1 + MD_2), \text{ when } \mu = 0. \tag{3.47}$$

Proof We notice that in the proof of Theorem 3.5, we derived inequality (3.42) that holds in the setting of Theorem 3.6 as well. Taking into account that $V_f(x^k, x^*) = f(x^k) - f(x^*)$ because of $R(x) \equiv 0$, we have

$$\begin{aligned}
\mathbb{E}_k[V_{k+1}] &\leq \left(1 - \min\left\{\gamma\mu, \rho - \frac{B}{M}\right\}\right)V_k + \gamma^2(D_1 + MD_2) \\
&\quad -2\gamma\left(1 - \gamma(A+CM)\right)\left(f(x^k) - f(x^*)\right) \\
&\overset{(3.44)}{\leq} (1-\eta)V_k + \gamma^2(D_1 + MD_2) - \gamma\left(f(x^k) - f(x^*)\right),
\end{aligned}$$

where $\eta = \min\left\{\gamma\mu, \rho - \frac{B}{M}\right\}$. Taking the full expectation and rearranging the terms, we get

$$\gamma\mathbb{E}\left[f(x^k) - f(x^*)\right] \leq (1-\eta)\mathbb{E}[V_k] - \mathbb{E}[V_{k+1}] + \gamma^2(D_1 + MD_2). \tag{3.48}$$

Next, we sum up the above inequality for $k = 0, \ldots, K$ with weights w_k and divide the result by γW_K:

$$\begin{aligned}\mathbb{E}\left[\sum_{k=0}^{K}\frac{w_k}{W_K}(f(x^k)-f(x^*))\right] &\leq \frac{1}{\gamma W_K}\sum_{k=0}^{K}((1-\eta)w_k\mathbb{E}[V_k]-w_k\mathbb{E}[V_{k+1}])\\ &\quad+\frac{\gamma(D_1+MD_2)}{W_K}\sum_{k=0}^{K}w_k\\ &= \frac{1}{\gamma W_K}\sum_{k=0}^{K}(w_{k-1}\mathbb{E}[V_k]-w_k\mathbb{E}[V_{k+1}])\\ &\quad+\gamma(D_1+MD_2)\\ &\leq \frac{\mathbb{E}[V_0]}{\gamma W_K}+\gamma(D_1+MD_2),\end{aligned}$$

where in the last step we use $w_{-1}=1$ and $w_K\mathbb{E}[V_{K+1}]\geq 0$. Applying Jensen's inequality, we derive (3.45). When $\mu>0$, we have $W_K\geq w_K=(1-\eta)^{-(K+1)}$ implying (3.46). In the case of $\mu=0$, we have $W_K=K+1$ implying (3.47). □

The above result is derived for the case of $\mu>0$ and $\mu=0$ simultaneously. When $\mu>0$, the above theorem implies that in terms of the expected functional suboptimality SGD converges exponentially fast to some accuracy depending on the stepsize γ and noises D_1, D_2. In view of Theorem 3.5, such a behavior is expected and very similar to the convergence of expected squared distance to the solution, when $\mu>0$. Since $\mathbb{E}[\|x^k-x^*\|^2]$ is not a valid metric of the convergence when $\mu=0$, we consider the expected functional suboptimality in this case. In this case, SGD converges with rate $O(1/K)$ to the accuracy $\gamma(D_1+MD_2)$ depending on the stepsize γ and noises D_1, D_2.

Theorems 3.5 and 3.6 provide the convergence guarantees for SGD under quite general assumptions. These guarantees establish the convergence only to some neighborhood/accuracy of the solution. To achieve any predefined accuracy $\varepsilon>0$ one needs to adjust stepsize γ accordingly. In the following subsection, we focus on thus aspect.

3.1.4.1 On the Choice of Stepsizes

We start with the general lemma for handling typical recurrences appearing in the convergence analysis of SGD-like methods.

Lemma 3.5 (Lemma 3 from [124]) *Assume that the non-negative sequences* $\{r_k\}_{k\geq 0}, \{s_k\}_{k\geq 0}$ *satisfy the recursion*

$$r_{k+1}\leq(1-a\gamma_k)r_k-b\gamma_k s_k+c\gamma_k^2 \tag{3.49}$$

for $k\geq 0$*, where* $a>0$*,* $c\geq 0$ *and* $\{\gamma_k\}_{k\geq 0}$ *is a non-negative sequence satisfying* $\gamma_k\leq 1/t$ *for some* $t\geq a$*. Then, for any* $K\geq 2$ *one can choose* $\{\gamma_k\}_{k\geq 0}$ *and sequence* $\{w_k\}_{k\geq 0}$ *as follows:*

$$\text{if } K \le \frac{t}{a}, \qquad \gamma_k = \tfrac{1}{t}, \qquad w_k = \left(1 - \frac{a}{t}\right)^{-(k+1)},$$
$$\text{if } K > \frac{t}{a} \text{ and } k < k_0, \qquad \gamma_k = \tfrac{1}{t}, \qquad w_k = 0,$$
$$\text{if } K > \frac{t}{a} \text{ and } k \ge k_0, \ \gamma_k = \tfrac{2}{a(\kappa+k-k_0)}, \ w_k = \kappa + k - k_0,$$

where $\kappa = 2t/a$ *and* $k_0 = \lceil K/2 \rceil$. *For this choice of* γ_k *the following inequality holds:*

$$\frac{b}{W_K} \sum_{k=0}^{K} w_k s_k + a r_{K+1} \le 128 t r_0 \exp\left(-\frac{aK}{2t}\right) + \frac{72c}{aK},$$

where $W_K = \sum_{k=0}^{K} w_k$.

Proof Consider the first case: $K \le t/a$. In this case, $\gamma_k = \gamma = 1/t$ and $w_k = (1 - a/t)^{-(k+1)} = (1 - a\gamma)^{-(k+1)}$. Multiplying (3.49) by w_k/γ and rearranging the terms, we get

$$b w_k s_k \le \frac{w_k(1 - a\gamma) r_k}{\gamma} - \frac{w_k r_{k+1}}{\gamma} + c\gamma w_k = \frac{w_{k-1} r_k}{\gamma} - \frac{w_k r_{k+1}}{\gamma} + c\gamma w_k,$$

where for $k = 0$ the weight w_{k-1} is defined as $w_{-1} = 1$. Next, we sum up the above inequality from $k = 0$ to $k = K$ and obtain

$$\begin{aligned} \frac{b}{W_K} \sum_{k=0}^{K} w_k s_k &\le \frac{1}{W_K} \sum_{k=0}^{K} \left(\frac{w_{k-1} r_k}{\gamma} - \frac{w_k r_{k+1}}{\gamma} \right) + c\gamma \frac{1}{W_K} \sum_{k=0}^{K} w_k \\ &= \frac{r_0}{\gamma W_K} - \frac{w_K r_{K+1}}{\gamma W_K} + c\gamma. \end{aligned}$$

Since $W_K = \sum_{k=0}^{K} (1 - a\gamma)^{-(k+1)} = (1 - a\gamma)^{-(K+1)} \sum_{k=0}^{K} (1 - a\gamma)^k \le \frac{1}{a\gamma(1-a\gamma)^{K+1}} \le \frac{1}{a\gamma}$ and $W_K \ge w_K = (1 - a\gamma)^{-(K+1)} \le \exp(-a\gamma(K+1))$, we can continue the above derivation as follows:

$$\begin{aligned} \frac{b}{W_K} \sum_{k=0}^{K} w_k s_k + a r_{K+1} &\le \frac{b}{W_K} \sum_{k=0}^{K} w_k s_k + \frac{w_K r_{K+1}}{\gamma W_K} \le \frac{r_0}{\gamma W_K} + c\gamma \\ &\overset{\gamma = 1/t \le 1/K}{\le} t r_0 \exp\left(-\frac{aK}{2t}\right) + \frac{c}{aK}. \end{aligned} \tag{3.50}$$

In the second case, we have $K > t/a$. For all $k \ge k_0$ we have $\gamma_k = 2/a(\kappa+k-k_0)$, $w_k = (\kappa + k - k_0)^2$, and

$$\begin{aligned} bw_k s_k &\overset{(3.49)}{\leq} \frac{w_k(1-a\gamma_k)r_k}{\gamma_k} - \frac{w_k r_{k+1}}{\gamma_k} + c\gamma_k w_k \\ &= a(\kappa+k-k_0)(\kappa+k-k_0-2)r_k - a(\kappa+k-k_0)^2 r_{k+1} + \frac{c}{a} \\ &\leq a(\kappa+k-k_0-1)^2 r_k - a(\kappa+k-k_0)^2 r_{k+1} + \frac{c}{a}, \end{aligned}$$

where in the last step we use $(\kappa+k-k_0)(\kappa+k-k_0-2) = (\kappa+k-k_0-1)^2 - 1 \leq (\kappa+k-k_0-1)^2$. Summing up the obtained inequality from $k = k_0$ to $k = K$, we obtain a telescoping sum and after small rearrangements we derive

$$\frac{b}{W_K}\sum_{k=k_0}^{K} w_k s_k + \frac{a(\kappa+K-k_0)^2 r_{K+1}^2}{W_K} \leq \frac{a\kappa^2 r_{k_0}}{W_K} + \frac{c(K+1-k_0)}{aW_K}.$$

Next, we provide lower and upper bounds for W_K:

$$\begin{aligned} W_K &= \sum_{k=0}^{K} w_k = \sum_{k=k_0}^{K} w_k = \sum_{k=k_0}^{K} (\kappa+k-k_0) = \frac{(2\kappa+K-k_0)(K-k_0+1)}{2} \\ &\geq \frac{(K-k_0)(K-k_0+1)}{2} \geq \frac{(K-k_0)^2}{2}, \\ W_K &= \frac{(2\kappa+K-k_0)(K-k_0+1)}{2} \leq \frac{2(\kappa+K-k_0)(K-k_0+1)}{2} \\ &\leq (\kappa+K-k_0)^2, \end{aligned}$$

where in the last step we use $t \geq a$. Using these relations and $w_k = 0$ for $k \leq K_0$, we obtain

$$\begin{aligned} \frac{b}{W_K}\sum_{k=0}^{K} w_k s_k + a r_{K+1} &\leq \frac{b}{W_K}\sum_{k=k_0}^{K} w_k s_k + \frac{a(\kappa+K-k_0)^2 r_{K+1}^2}{W_K} \\ &\leq \frac{a\kappa^2 r_{k_0}}{W_K} + \frac{c(K+1-k_0)}{aW_K} \\ &\leq \frac{2a\kappa^2 r_{k_0}}{(K-k_0)^2} + \frac{2c}{a(K-k_0)} \\ &\leq \frac{32a\kappa^2 r_{k_0}}{K^2} + \frac{8c}{aK}, \end{aligned} \tag{3.51}$$

where in the last step we use $k_0 = \lceil K/2 \rceil \leq {}^{(K+1)}/_2$ and $K - 1 \geq {}^{K}/_2$. To estimate r_{k_0} we unroll the recurrence (3.49):

$$\begin{aligned} r_{k_0} &\leq (1-a\gamma_k)r_k - b\gamma_k s_k + c\gamma_k^2 \overset{\gamma_k=\gamma}{\leq} (1-a\gamma)^{k_0} r_0 + c\gamma^2 \sum_{k=0}^{k_0} (1-a\gamma)^k \\ &\leq r_0 \exp\left(-\frac{a}{t}k_0\right) + \frac{c}{at} \overset{k_0 \geq K/2}{\leq} r_0 \exp\left(-\frac{aK}{2t}\right) + \frac{c}{at}. \end{aligned} \tag{3.52}$$

Plugging (3.52) in (3.51) and applying $K \geq t/a$, we get

$$\frac{b}{W_K}\sum_{k=0}^{K} w_k s_k + a r_{K+1} \leq \frac{32a\kappa^2 r_0 \exp\left(-\frac{aK}{2t}\right)}{K^2} + \frac{32c\kappa^2}{tK^2} + \frac{8c}{aK}$$
$$= 128 a r_0 \exp\left(-\frac{aK}{2t}\right) + \frac{72c}{aK}.$$

Combining the above upper bound with (3.50) we get the result. □

Using the above lemma, we derive several general results allowing to transform the upper bounds from Theorems 3.5 and 3.6 to the convergence rates.

Corollary 3.2 *Let the assumptions of Theorem 3.5 hold and $D_1 + MD_2 \neq 0$. Then, for any $K \geq 2$ one can choose $\{\gamma_k\}_{k\geq 0}$ as follows:*

$$\text{if } K \leq \frac{t}{\mu}, \qquad \gamma_k = \frac{1}{t},$$
$$\text{if } K > \frac{t}{\mu} \text{ and } k < k_0, \qquad \gamma_k = \frac{1}{t},$$
$$\text{if } K > \frac{t}{\mu} \text{ and } k \geq k_0, \qquad \gamma_k = \frac{2}{\mu(\kappa + k - k_0)},$$

where $k_0 = \lceil K/2 \rceil$, $\kappa = 2t/\mu$, and

$$t = \max\left\{\frac{2\mu}{\rho}, A + \frac{2BC}{\rho}\right\}. \tag{3.53}$$

For this choice of γ_k the iterates produced by SGD *satisfy*

$$\mathbb{E}\left[\|x^{K+1} - x^*\|^2\right] \leq 128\Omega_0^2 \exp\left(-\min\left\{\frac{\rho}{4}, \frac{\mu}{2A + 4BC/\rho}\right\} K\right) + \frac{72(D_1 + 2BD_2/\rho)}{\mu^2 K},$$

where $\Omega_0^2 = \|x^0 - x^*\|^2 + \mathbb{E}[\sigma_0^2]2B/\rho t^2$.

Proof In the proof of Theorem 3.5, we derived (3.43) stating that for all $k \geq 0$ and any stepsize

$$0 < \gamma \leq \min\left\{\frac{1}{\mu}, \frac{1}{A + CM}\right\}$$

we have

$$\mathbb{E}[V_{k+1}] \leq \left(1 - \min\left\{\gamma\mu, \rho - \frac{B}{M}\right\}\right)\mathbb{E}[V_k] + \gamma^2(D_1 + MD_2),$$

where $V_k = \|x^k - x^*\|^2 + M\gamma^2\sigma_k^2$ and $M > B/\rho$. To derive this inequality, we consider one iteration of SGD with constant stepsize $\gamma_k = \gamma$. However, the same inequality holds for SGD with non-constant stepsizes, if γ_k satisfies

$$0 < \gamma_k \leq \min\left\{\frac{1}{\mu}, \frac{1}{A + CM}\right\}$$

for all $k \geq 0$. Since our choice of γ_k satisfies this condition for $M = 2B/\rho$, we have

$$\begin{aligned}\mathbb{E}[V_{k+1}] &\leq \left(1 - \min\left\{\gamma_k\mu, \rho - \frac{B}{M}\right\}\right)\mathbb{E}[V_k] + \gamma_k^2(D_1 + MD_2) \\ &= (1 - \gamma_k\mu)\,\mathbb{E}[V_k] + \gamma_k^2(D_1 + 2BD_2/\rho), \end{aligned} \tag{3.54}$$

where in the second step we apply

$$\gamma_k\mu \leq \frac{\rho}{2} = \rho - \frac{B}{M}.$$

Inequality (3.54) implies (3.49) from Lemma 3.5 with $r_k = \mathbb{E}[V_k]$, $a = \mu$, $b = 0$, $s_k \equiv 0$, $c = D_1 + 2BD_2/\rho$, and t defined in (3.53). Plugging these parameters in Lemma 3.5 and using $V_k \geq \|x^k - x^*\|^2$, we get the result. □

Corollary 3.3 *Let the assumptions of Theorem 3.5 hold and $D_1 + MD_2 = 0$. Let the stepsize γ_k be*

$$\gamma_k = \gamma = \min\left\{\frac{1}{\mu}, \frac{1}{A + 2BC/\rho}\right\}$$

and $M = 2B/\rho$. For this choice of γ_k and for any $K \geq 0$ the iterates produced by SGD *satisfy*

$$\mathbb{E}\left[\|x^K - x^*\|^2\right] \leq \Omega_0^2 \exp\left(-\min\left\{\frac{\mu}{A + 2BC/\rho}, \frac{\rho}{2}\right\}K\right),$$

where $\Omega_0^2 = \|x^0 - x^*\|^2 + \mathbb{E}[\sigma_0^2]2B\gamma^2/\rho$.

Proof Since $D_1 + MD_2 = 0$, $M = 2B/\rho$, Theorem 3.5 implies for any $K \geq 0$ that

$$\mathbb{E}\left[V_K\right] \leq \left(1 - \min\left\{\gamma\mu, \frac{\rho}{2}\right\}\right)^K \mathbb{E}[V_0],$$

where $V_K = \|x^K - x^*\|^2 + M\gamma^2\sigma_K^2$. Taking into account that $V_K \geq \|x^K - x^*\|^2$ and $(1 - a)^t \leq \exp(-at)$ for any $a \in (0, 1)$, we derive

$$\begin{aligned}\mathbb{E}\left[\|x^K - x^*\|^2\right] &\leq \mathbb{E}[V_0]\exp\left(-\min\left\{\gamma\mu, \frac{\rho}{2}\right\}K\right) \\ &= \Omega_0^2 \exp\left(-\min\left\{\frac{\mu}{A + 2BC/\rho}, \frac{\rho}{2}\right\}K\right),\end{aligned}$$

which concludes the proof. □

Corollary 3.4 *Let the assumptions of Theorem 3.6 hold, $\mu > 0$, and $D_1 + MD_2 \neq 0$. hen, for any $K \geq 2$ one can choose $\{\gamma_k\}_{k\geq 0}$ and sequence $\{w_k\}_{k\geq 0}$ as follows:*

$$\begin{aligned}
&\text{if } K \le \frac{t}{\mu}, \qquad \gamma_k = \tfrac{1}{t}, \qquad w_k = \left(1 - \frac{\mu}{t}\right)^{-(k+1)},\\
&\text{if } K > \frac{t}{\mu} \text{ and } k < k_0, \qquad \gamma_k = \tfrac{1}{t}, \qquad w_k = 0,\\
&\text{if } K > \frac{t}{\mu} \text{ and } k \ge k_0,\ \gamma_k = \tfrac{2}{\mu(\kappa+k-k_0)},\ w_k = \kappa + k - k_0,
\end{aligned}$$

where $k_0 = \lceil K/2 \rceil$, $\kappa = 2t/\mu$, *and*

$$t = \max\left\{\frac{2\mu}{\rho}, 2A + \frac{4BC}{\rho}\right\}. \tag{3.55}$$

For this choice of γ_k *the iterates produced by* SGD *satisfy*

$$\mathbb{E}\left[f(\overline{x}^K) - f(x^*)\right] \le 128 t \Omega_0^2 \exp\left(-\min\left\{\frac{\rho}{4}, \frac{\mu}{4A + 8BC/\rho}\right\} K\right) + \frac{72(D_1 + 2BD_2/\rho)}{\mu K},$$

where $\overline{x}^K = \frac{1}{W_K}\sum_{k=0}^K w_k x^k$ *and* $\Omega_0^2 = \|x^0 - x^*\|^2 + \mathbb{E}[\sigma_0^2]2B/\rho t^2$.

Proof In the proof of Theorem 3.6, we derived (3.48) stating that for all $k \ge 0$ and any stepsize

$$0 < \gamma \le \min\left\{\frac{1}{\mu}, \frac{1}{2(A + CM)}\right\}$$

we have

$$\mathbb{E}[V_{k+1}] \le (1 - \eta)\,\mathbb{E}[V_k] - \gamma \mathbb{E}\left[f(x^k) - f(x^*)\right] + \gamma^2(D_1 + MD_2),$$

where $V_k = \|x^k - x^*\|^2 + M\gamma^2\sigma_k^2$, $\eta = \min\{\gamma\mu, \rho - B/M\}$, and $M > B/\rho$. To derive this inequality, we consider one iteration of SGD with constant stepsize $\gamma_k = \gamma$. However, the same inequality holds for SGD with non-constant stepsizes, if γ_k satisfies

$$0 < \gamma_k \le \min\left\{\frac{1}{\mu}, \frac{1}{2(A + CM)}\right\}$$

for all $k \ge 0$. Since our choice of γ_k satisfies this condition for $M = 2B/\rho$, we have

$$\begin{aligned}
\mathbb{E}[V_{k+1}] &\le (1 - \eta)\,\mathbb{E}[V_k] - \gamma_k \mathbb{E}\left[f(x^k) - f(x^*)\right] + \gamma_k^2(D_1 + MD_2)\\
&= (1 - \gamma_k\mu)\,\mathbb{E}[V_k] - \gamma_k \mathbb{E}\left[f(x^k) - f(x^*)\right]\\
&\quad + \gamma_k^2(D_1 + 2BD_2/\rho),
\end{aligned} \tag{3.56}$$

where in the second step we apply

$$\gamma_k\mu \le \frac{\rho}{2} = \rho - \frac{B}{M}.$$

Inequality (3.56) implies (3.49) from Lemma 3.5 with $r_k = \mathbb{E}[V_k]$, $a = \mu$, $b = 1$, $s_k = \mathbb{E}[f(x^k) - f(x^*)]$, $c = D_1 + 2BD_2/\rho$, and t defined in (3.55). Plugging these

parameters in Lemma 3.5 and using $V_k \geq 0$ and Jensen's inequality for the function f, i.e., using

$$f(\overline{x}^K) \leq \frac{1}{W_K} \sum_{k=0}^{K} w_k f(x^k)$$

for $\overline{x}^K = \frac{1}{W_K} \sum_{k=0}^{K} w_k$, we get the result. □

Corollary 3.5 *Let the assumptions of Theorem 3.6 hold, $\mu > 0$, and $D_1 + MD_2 = 0$. Let the stepsize γ_k be*

$$\gamma_k = \gamma = \min\left\{\frac{1}{\mu}, \frac{1}{2A + {}^{4BC}/_{\rho}}\right\}$$

and $M = {}^{2B}/_{\rho}$. For this choice of γ_k and for any $K \geq 0$ the iterates produced by SGD *satisfy*

$$\mathbb{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \Omega_0^2 \exp\left(-\min\left\{\frac{\mu}{2A + {}^{4BC}/_{\rho}}, \frac{\rho}{2}\right\} K\right),$$

where $\overline{x}^K = \frac{1}{W_K} \sum_{k=0}^{K} w_k x^k$, $w_k = (1-\eta)^{-(k+1)}$, $\eta = \min\{{}^{\rho}/_2, \gamma\mu\}$, $W_K = \sum_{k=0}^{K} w_k$, and $\Omega_0^2 = {}^{\|x^0 - x^*\|^2}/_{\gamma} + \mathbb{E}[\sigma_0^2]{}^{2B\gamma}/_{\rho}$.

Proof Since $D_1 + MD_2 = 0$, $M = {}^{2B}/_{\rho}$, Theorem 3.6 implies for any $K \geq 0$ that

$$\mathbb{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \left(1 - \min\left\{\gamma\mu, \frac{\rho}{2}\right\}\right)^K \frac{\mathbb{E}[V_0]}{\gamma},$$

where $\overline{x}^K = \frac{1}{W_K} \sum_{k=0}^{K} w_k x^k$, $w_k = (1-\eta)^{-(k+1)}$, $\eta = \min\{{}^{\rho}/_2, \gamma\mu\}$, $W_K = \sum_{k=0}^{K} w_k$, and $V_K = \|x^K - x^*\|^2 + M\gamma^2\sigma_K^2$. Taking into account that $W_K \geq w_k = (1-\eta)^{-(K+1)} \geq \exp(\eta(K+1)) \geq \exp(\eta K)$, we derive

$$\begin{aligned}\mathbb{E}\left[f(\overline{x}^K) - f(x^*)\right] &\leq \frac{\mathbb{E}[V_0]}{\gamma} \exp\left(-\eta K\right) \\ &= \Omega_0^2 \exp\left(-\min\left\{\frac{\mu}{2A + {}^{4BC}/_{\rho}}, \frac{\rho}{2}\right\} K\right),\end{aligned}$$

which concludes the proof. □

Corollary 3.6 *Let the assumptions of Theorem 3.6 hold and $\mu = 0$. Let $K \geq 1$ and the stepsize γ_k be*

$$\gamma_k = \gamma = \min\left\{\frac{1}{2A + {}^{4BC}/_{\rho}}, \sqrt{\frac{\|x^0 - x^*\|^2}{(D_1 + {}^{2BD_2}/_{\rho})K + \mathbb{E}[\sigma_0^2]{}^{2B}/_{\rho}}}\right\}$$

and $M = {}^{2B}/_{\rho}$. For this choice of γ_k the iterates produced by SGD *satisfy*

$$\mathbb{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \frac{(2A + 4BC/\rho)R_0^2}{K} + \frac{2R_0\sqrt{2B\mathbb{E}[\sigma_0^2]}}{\sqrt{\rho}K} + \sqrt{\frac{4(D_1 + 2BD_2/\rho)R_0^2}{K}},$$

where $\overline{x}^K = \frac{1}{K+1}\sum_{k=0}^K x^k$ *and* $R_0 = \|x^0 - x^*\|$.

Proof Since $\mu = 0$, $V_0 = R_0^2 + M\gamma^2\sigma_0^2$, and $M = 2B/\rho$, Theorem 3.6 implies

$$\begin{aligned}\mathbb{E}\left[f(\overline{x}^K) - f(x^*)\right] &\leq \frac{\mathbb{E}[V_0]}{\gamma(K+1)} + \gamma(D_1 + MD_2)\\ &\leq \frac{R_0^2}{\gamma K} + \frac{2B\gamma\mathbb{E}[\sigma_0^2]}{\rho K} + \gamma(D_1 + 2BD_2/\rho). \qquad (3.57)\end{aligned}$$

It remains to estimate each term in the right-hand side of the above inequality. Using the choice of γ, we get

$$\begin{aligned}\frac{R_0^2}{\gamma K} &= \max\left\{\frac{(2A + 4BC/\rho)R_0^2}{K}, \frac{R_0\sqrt{(D_1 + 2BD_2/\rho)K + \mathbb{E}[\sigma_0^2]2B/\rho}}{K}\right\}\\ &\leq \frac{(2A + 4BC/\rho)R_0^2}{K} + \frac{R_0\sqrt{(D_1 + 2BD_2/\rho)K + \mathbb{E}[\sigma_0^2]2B/\rho}}{K}\\ &\leq \frac{(2A + 4BC/\rho)R_0^2}{K} + \frac{R_0\sqrt{D_1 + 2BD_2/\rho}}{\sqrt{K}} + \frac{R_0\sqrt{2B\mathbb{E}[\sigma_0^2]}}{\sqrt{\rho}K},\end{aligned}$$

$$\begin{aligned}\frac{2B\gamma\mathbb{E}[\sigma_0^2]}{\rho K} &\leq \frac{2B\mathbb{E}[\sigma_0^2]}{\rho K}\sqrt{\frac{\|x^0 - x^*\|^2}{(D_1 + 2BD_2/\rho)K + \mathbb{E}[\sigma_0^2]2B/\rho}}\\ &\leq \frac{R_0\sqrt{2B\mathbb{E}[\sigma_0^2]}}{\sqrt{\rho}K},\end{aligned}$$

and

$$\begin{aligned}\gamma(D_1 + 2BD_2/\rho) &\leq (D_1 + 2BD_2/\rho)\sqrt{\frac{\|x^0 - x^*\|^2}{(D_1 + 2BD_2/\rho)K + \mathbb{E}[\sigma_0^2]2B/\rho}}\\ &\leq \frac{R_0\sqrt{D_1 + 2BD_2/\rho}}{\sqrt{K}}.\end{aligned}$$

Plugging the above estimates in (3.57), we get the result. □

In the next few subsections, we will actively use Corollaries 3.2-3.6 to derive the convergence rates of SGD-like methods under different assumptions.

3.1.4.2 SGD Under the Bounded Variance Assumption

Consider the case when f is convex, L-smooth, and Assumption 3.1 holds. As we show earlier in this section (see (3.37)), Assumption 3.3 holds in this case with the following parameters:

$$A = L, \quad B = 0, \quad C = 0, \quad D_1 = \sigma^2, \quad \rho = 1, \quad \sigma_k^2 \equiv 0, \quad D_2 = 0.$$

We also highlight that in the case of $h(x) = 0$, one can derive the same fact without assuming the convexity of f. Plugging these parameters in Theorems 3.5, 3.6 and Corollaries 3.2, 3.4, 3.6 we obtain several convergence results summarized in the following theorems.

Theorem 3.7 *Let the objective function f be convex, (μ, x^*)-quasi strongly convex with $\mu > 0$, L-smooth, and Assumption 3.1 hold. Assume that the stepsize $\gamma_k = \gamma$ satisfies*

$$0 < \gamma \leq \frac{1}{L}. \tag{3.58}$$

Then, for each $k \geq 0$ the iterations of SGD *satisfy*

$$\mathbb{E}\left[\|x^k - x^*\|^2\right] \leq (1 - \gamma\mu)^k R_0^2 + \frac{\gamma\sigma^2}{\mu}, \tag{3.59}$$

where $R_0 = \|x^0 - x^*\|$. Moreover, for any $K \geq 2$ one can choose $\{\gamma_k\}_{k\geq 0}$ as follows:

$$\begin{aligned} \text{if } K \leq \frac{L}{\mu}, \quad & \gamma_k = \frac{1}{L}, \\ \text{if } K > \frac{L}{\mu} \text{ and } k < k_0, \quad & \gamma_k = \frac{1}{L}, \\ \text{if } K > \frac{L}{\mu} \text{ and } k \geq k_0, \quad & \gamma_k = \frac{2}{2L + \mu(k - k_0)}, \end{aligned}$$

where $k_0 = \lceil K/2 \rceil$. For this choice of γ_k the iterates produced by SGD *satisfy*

$$\mathbb{E}\left[\|x^{K+1} - x^*\|^2\right] \leq 128 R_0^2 \exp\left(-\frac{\mu}{2L}K\right) + \frac{72\sigma^2}{\mu^2 K}. \tag{3.60}$$

Proof The result follows from Theorem 3.5 and Corollary 3.2. □

In contrast to (3.59), upper bound (3.60) decreases with the growth of K. It consists of two terms: exponentially decaying one and $O(1/K)$ term. For large enough K the second term dominates the first one and the method converges relatively slow in this case. In contrast when σ^2 is sufficiently small, e.g., reduced via mini-batching, the first term can dominate the second one for quite large range of K, thus, the method converges linearly during the quite long initial stage.

In the next theorem, we provide the results in term of the functional values.

Theorem 3.8 *Let the objective function f be (μ, x^*)-quasi strongly convex with $\mu \geq 0$, L-smooth, $h(x) \equiv 0$, and Assumption 3.1 hold. Assume that the stepsize $\gamma_k = \gamma$ satisfies*

$$0 < \gamma \leq \frac{1}{2L}.$$

Then, for each $k \geq 0$ the iterations of the method from (3.34) *satisfy*

$$\mathbb{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq (1-\gamma\mu)^{K+1}\frac{R_0^2}{\gamma} + \gamma\sigma^2, \text{ when } \mu > 0, \tag{3.61}$$

$$\mathbb{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \frac{R_0^2}{\gamma(K+1)} + \gamma\sigma^2, \text{ when } \mu = 0, \tag{3.62}$$

where $\overline{x}^K = \frac{1}{W_K}\sum_{k=0}^K w_k x^k$, $w_k = (1-\gamma\mu)^{-(k+1)}$, $W_K = \sum_{k=0}^K w_k$, $R_0 = \|x^0 - x^*\|^2$. In particular, if $\mu > 0$, then for any $K \geq 2$ one can choose $\{\gamma_k\}_{k\geq 0}$ and sequence $\{w_k\}_{k\geq 0}$ as follows:

$$\begin{aligned}
&\text{if } K \leq \frac{2L}{\mu}, && \gamma_k = \tfrac{1}{2L}, && w_k = \left(1 - \frac{\mu}{2L}\right)^{-(k+1)},\\
&\text{if } K > \frac{2L}{\mu} \text{ and } k < k_0, && \gamma_k = \tfrac{1}{2L}, && w_k = 0,\\
&\text{if } K > \frac{2L}{\mu} \text{ and } k \geq k_0, && \gamma_k = \tfrac{2}{4L+\mu(k-k_0)}, && w_k = \frac{4L}{\mu} + k - k_0,
\end{aligned}$$

where $k_0 = \lceil K/2 \rceil$. For this choice of γ_k the iterates produced by SGD *satisfy*

$$\mathbb{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq 256LR_0^2\exp\left(-\frac{\mu}{8L}K\right) + \frac{72\sigma^2}{\mu K}, \tag{3.63}$$

where $\overline{x}^K = \frac{1}{W_K}\sum_{k=0}^K w_k x^k$. Finally, if $\mu = 0$, then for any $K \geq 1$ the iterates produced by SGD *with stepsize*

$$\gamma_k = \gamma = \min\left\{\frac{1}{4L}, \sqrt{\frac{R_0^2}{\sigma^2 K}}\right\}$$

satisfy

$$\mathbb{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \frac{4LR_0^2}{K} + \frac{2\sigma R_0}{\sqrt{K}}, \tag{3.64}$$

where $\overline{x}^K = \frac{1}{K+1}\sum_{k=0}^K x^k$.

Proof The result follows from Theorem 3.6 and Corollaries 3.4, 3.6. □

When $\mu > 0$, the above theorem establishes (3.63), which is very similar to (3.60). We also point out that neither (3.63) follows from (3.60) nor (3.60) follows from

(3.63). Indeed, let (3.60) holds and $h(x) \equiv 0$. Then from L-smoothness we get

$$\mathbb{E}\left[f(x^{K+1}) - f(x^*)\right] \leq 64LR_0^2 \exp\left(-\frac{\mu}{2L}K\right) + \frac{36L\sigma^2}{\mu^2 K},$$

which has an extra $O(L/\mu)$ factor in front to the second term if compared to (3.63). Let us now assume that (3.63) holds and f is μ-strongly convex in addition. Then, from the strong convexity we get

$$\mathbb{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq 512\frac{L}{\mu}R_0^2 \exp\left(-\frac{\mu}{8L}K\right) + \frac{144\sigma^2}{\mu^2 K},$$

which has an extra $O(L/\mu)$ factor in front to the first term if compared to (3.60). Therefore, the bounds (3.60) and (3.63) complement each other.

When $\mu = 0$, the above theorem gives bound (3.64), which has two terms. The first term decreases as $O(1/K)$, which corresponds to the convergence of gradient descent for convex smooth functions. The second term is due to the stochasticity of updates and it decreases much slower.

3.1.4.3 SGD and Arbitrary Sampling

As we already shown earlier (see (3.38)), when f is convex and Assumption 3.2 holds Assumption 3.3 is satisfied with the following parameters:

$$A = 2L_{\max}, \quad B = 0, \quad C = 0, \quad D_1 = 2\sigma_*^2, \quad \rho = 1, \quad \sigma_k^2 \equiv 0, \quad D_2 = 0. \tag{3.65}$$

Let us remind here that $L_{\max}$ is the worst smoothness constant of the stochastic realization $f_\xi(x)$. This constant can be much larger than the smoothness constant of f, e.g., in the finite-sum case, it can be m times larger. Since the rates from Corollaries 3.2-3.6 becomes worse when A increases, the question of reducing A is important. To address this question we consider a modification of Assumption 3.2.

Assumption 3.4 (Expected Smoothness) Stochastic gradient g^k is computed as $g^k = \nabla f_{\xi^k}(x^k)$, where ξ^k is sampled from some distribution $\mathcal{D}$ independently from previous iterations. Moreover, there exists a positive constant $\mathcal{L} > 0$ such that for all $x \in \mathbb{R}^n$ and x^* being the solution of (3.30) the following inequality holds:

$$\mathbb{E}_\xi\left[\|\nabla f_\xi(x) - \nabla f_\xi(x^*)\|^2\right] \leq 2\mathcal{L}V_f(x, x^*). \tag{3.66}$$

As we will see further, this assumption can replace Assumption 3.2 to achieve similar convergence guarantees as in Theorem 3.2, i.e., Assumption 3.4 is a weaker version of Assumption 3.2 allowing to achieve similar results. To illustrate the generality of the introduced assumption, we consider finite-sum case, i.e., we focus on problem (3.30) with

$$f(x) = \frac{1}{m}\sum_{i=1}^m f_i(x).$$

Classical approach to sample stochastic gradients in this case is a uniform sampling, i.e., $g^k = \nabla f_{\xi^k}(x^k)$, where ξ^k is sampled from the uniform distribution on $\{1, \dots, m\}$ independently from previous steps. There exist other sampling strategies that we can describe as follows. Let $\xi = (\xi_1, \dots, \xi_m)^\top \in \mathbb{R}^m$ be m-dimensional random vector with distribution $\mathcal{D}$ such that

$$\mathbb{E}_{\xi \sim \mathcal{D}}[\xi_i] = 1 \quad \text{for all } i \in \{1, \dots, m\}. \tag{3.67}$$

Next, for any $\xi \sim \mathcal{D}$ we consider function

$$f_\xi(x) = \frac{1}{m} \sum_{i=1}^m \xi_i f_i(x). \tag{3.68}$$

Since $\mathbb{E}[\xi_i] = 1$, we have $\mathbb{E}_{\xi \sim \mathcal{D}}[f_\xi(x)] = f(x)$ and $\mathbb{E}_{\xi \sim \mathcal{D}}[\nabla f_\xi(x)] = \nabla f(x)$. Such a *stochastic reformulation* of the initial problem is convenient for dealing with a broad class of samplings of stochastic gradients. Indeed, $g^k = \nabla f_{\xi^k}(x^k)$ can be seen as a stochastic gradient and the sampling strategy is determined by distribution $\mathcal{D}$ of *sampling vector* ξ. Below we provide several examples of samplings that fit the described framework. We start with the standard uniform sampling.

Uniform Sampling

Let $\mathbb{P}\{\xi = m \cdot e_i\} = 1/m$, where e_i is the i-th basis vector in $\mathbb{R}^m$, $i \in \{1, \dots, m\}$, i.e., i-th component of e_i equals 1 and all remaining components of e_i equal 0. Then, $\mathbb{E}[\xi_i] = m \cdot (1/m) + 0 \cdot (1 - 1/m) = 1$ for all $i \in \{1, \dots, m\}$ and $\nabla f_\xi(x) = \nabla f_j(x)$, where j is the random integer uniformly distributed on $\{1, \dots, m\}$.

Next, assume that f_i is convex and L_i-smooth for all $i \in \{1, \dots, m\}$. Then,

$$\begin{aligned}
\mathbb{E}_\xi\left[\|\nabla f_\xi(x) - \nabla f_\xi(x^*)\|^2\right] &= \frac{1}{m}\sum_{i=1}^m \|\nabla f_i(x) - \nabla f_i(x^*)\|^2 \\
&\overset{\text{(B.10)}}{\leq} \frac{1}{m}\sum_{i=1}^m 2L_i V_{f_i}(x, x^*) \\
&\leq 2L_{\max} \cdot \frac{1}{m}\sum_{i=1}^m V_{f_i}(x, x^*) = 2L_{\max} V_f(x, x^*),
\end{aligned}$$

where $L_{\max} = \max_{i \in [m]} L_i$. That is, Assumption 3.4 holds with $\mathcal{L} = \mathcal{L}_{\text{US}} \overset{\text{def}}{=} L_{\max}$.

As expected, $\mathcal{L}$ equals the worst smoothness constant $L_{\max}$ in the case of uniform sampling. When smoothness constants $L_1, \dots, L_m$ are known or can be estimated in advance as in logistic/linear regression, it is possible to improve $\mathcal{L}$.

Importance Sampling

Let $\mathbb{P}\{\xi = (m\overline{L}/L_i) \cdot e_i\} = L_i/(m\overline{L})$ for all $i \in \{1, \ldots, m\}$, where $\overline{L} = \frac{1}{m}\sum_{i=1}^m L_i$. Then, for all $i \in \{1, \ldots, m\}$

$$\mathbb{E}[\xi_i] = \frac{m\overline{L}}{L_i} \cdot \frac{L_i}{m\overline{L}} + 0 \cdot \left(1 - \frac{L_i}{m\overline{L}}\right) = 1, \quad \nabla f_\xi(x) = \frac{\overline{L}}{L_j}\nabla f_j(x),$$

where j is the random integer such that $\mathbb{P}\{j = i\} = L_i/(m\overline{L})$ for all $i \in \{1, \ldots, m\}$.

Next, assume that f_i is convex and L_i-smooth for all $i \in \{1, \ldots, m\}$. Then,

$$\begin{aligned}
\mathbb{E}_\xi\left[\|\nabla f_\xi(x) - \nabla f_\xi(x^*)\|^2\right] &= \sum_{i=1}^m \frac{L_i}{m\overline{L}}\left\|\frac{\overline{L}}{L_i}\nabla f_i(x) - \frac{\overline{L}}{L_i}\nabla f_i(x^*)\right\|^2 \\
&= \frac{1}{m}\sum_{i=1}^m \frac{\overline{L}}{L_i}\|\nabla f_i(x) - \nabla f_i(x^*)\|^2 \\
&\overset{\text{(B.10)}}{\leq} \frac{1}{m}\sum_{i=1}^m \frac{\overline{L}}{L_i} \cdot 2L_i V_{f_i}(x, x^*) \\
&= 2\overline{L} \cdot \frac{1}{m}\sum_{i=1}^m V_{f_i}(x, x^*) = 2\overline{L}V_f(x, x^*).
\end{aligned}$$

That is, Assumption 3.4 holds with $\mathcal{L} = \mathcal{L}_{\text{IS}} \overset{\text{def}}{=} \overline{L}$.

We notice that $\overline{L} \leq L_{\max}$ and in some cases these smoothness constants can be significantly different. As we will see further, in such cases, SGD with importance sampling can significantly outperform standard SGD with uniform sampling.

However, in the worst case, both constants are m times larger than L – the smoothness constant of f. Indeed, consider the following function of m arguments:

$$f(x) = \frac{1}{m}\|x\|_2^2 = \frac{1}{m}\sum_{i=1}^m \underbrace{x_i^2}_{f_i(x)}. \tag{3.69}$$

Clearly, $f(x)$ is L-smooth with $L = 1/m$, while for all $i \in [m]$ function f_i is L_i-smooth with $L_i = 1$. Therefore, $\overline{L} = L_{\max} = 1 = mL$.

To some extent, one can handle this issue via mini-batching. Fortunately, mini-batching strategies also fit the formalism with sampling vectors ξ.

Sampling with Replacement

Let $\xi(1), \xi(2), \ldots, \xi(r)$ be i.i.d. samples from some distribution $\mathcal{D}$ satisfying (3.67). Then, distribution $\mathcal{D}_r$ of random vector

$$\xi = \frac{1}{r}\sum_{i=1}^r \xi(i)$$

also satisfies (3.67).

Next, assume that f is convex and L-smooth and Assumption 3.4 holds for distribution $\mathcal{D}$ with constant $\mathcal{L} = \mathcal{L}_{\mathcal{D}}$. To simplify the following derivation we introduce new notations: $\Delta(x, x^*) = \nabla f(x) - \nabla f(x^*)$ and $\Delta_{\xi(i)}(x, x^*) = \nabla f_{\xi(i)}(x) - \nabla f_{\xi(i)}(x^*)$, $i \in [r]$. Then, due to variance decomposition and independence of $\xi(1), \xi(2), \ldots, \xi(r)$ we have

$$\begin{aligned}
\mathbb{E}_{\xi}\left[\|\nabla f_{\xi}(x) - \nabla f_{\xi}(x^*)\|^2\right] &= \mathbb{E}_{\xi}\left[\left\|\frac{1}{r}\sum_{i=1}^{r}\Delta_{\xi(i)}(x, x^*)\right\|^2\right] \\
&= \|\Delta(x, x^*)\|^2 + \mathbb{E}_{\xi}\left[\left\|\frac{1}{r}\sum_{i=1}^{r}\left(\Delta_{\xi(i)}(x, x^*) - \Delta(x, x^*)\right)\right\|^2\right] \\
&= \|\Delta(x, x^*)\|^2 + \frac{1}{r^2}\sum_{i=1}^{r}\mathbb{E}_{\xi(i)}\left[\|\Delta_{\xi(i)}(x, x^*) - \Delta(x, x^*)\|^2\right] \\
&= \|\Delta(x, x^*)\|^2 + \frac{1}{r}\mathbb{E}_{\xi(1)}\left[\|\Delta_{\xi(1)}(x, x^*) - \Delta(x, x^*)\|^2\right] \\
&= \left(1 - \frac{1}{r}\right)\|\Delta(x, x^*)\|^2 + \frac{1}{r}\mathbb{E}_{\xi(1)}\left[\|\Delta_{\xi(1)}(x, x^*)\|^2\right] \\
&\overset{(B.10)}{\leq} 2\left(\left(1 - \frac{1}{r}\right)L + \frac{\mathcal{L}_{\mathcal{D}}}{r}\right)V_f(x, x^*).
\end{aligned}$$

That is, Assumption 3.4 holds with $\mathcal{L} = \mathcal{L}_{\mathcal{D},r} \overset{\text{def}}{=} \left(1 - \frac{1}{r}\right)L + \frac{\mathcal{L}_{\mathcal{D}}}{r}$. In particular, consider the situation when f_i are convex and L_i-smooth for $i \in [m]$. Then, for mini-batching with replacement using uniform sampling Assumption 3.4 holds with $\mathcal{L} = \mathcal{L}_{\text{US},r} \overset{\text{def}}{=} \left(1 - \frac{1}{r}\right)L + \frac{L_{\max}}{r}$ and for mini-batching with replacement using importance sampling Assumption 3.4 holds with $\mathcal{L} = \mathcal{L}_{\text{IS},r} \overset{\text{def}}{=} \left(1 - \frac{1}{r}\right)L + \frac{\overline{L}}{r}$, where $L_{\max} = \max_{i\in[m]} L_i$ and $\overline{L} = \frac{1}{m}\sum_{i=1}^{m} L_i$.

The above mini-batching strategy is classical. Next, we consider also two other approaches that form mini-batches *without replacement*.

Independent Sampling Without Replacement

Let random vector $\xi \in \mathbb{R}^m$ be such that

$$\xi_i = \begin{cases} \frac{1}{p_i}, & \text{with probability } p_i > 0, \\ 0, & \text{with probability } 1 - p_i, \end{cases} \qquad i \in [m]$$

and $\xi_1, \ldots, \xi_m$ are independent. In other words, each summand is taken in the mini-batch with probability p_i independently from other summands. By definition we

have $\mathbb{E}[\xi_i] = (1/p_i) \cdot p_i + 0 \cdot (1 - p_i) = 1$, so, (3.67) holds. The number of sampled summands is a random variable having the expectation equal $r \stackrel{\text{def}}{=} \sum_{i=1}^m p_i$.

Next, assume that f_i is convex and L_i-smooth for all $i \in \{1, \ldots, m\}$, f is L-smooth. To simplify the following derivation we introduce new notations: $\Delta(x, x^*) = \nabla f(x) - \nabla f(x^*)$, $\Delta_i(x, x^*) = \nabla f_i(x) - \nabla f_i(x^*)$, $i \in [m]$. Then, due to variance decomposition and independence of $\xi_1, \xi_2, \ldots, \xi_r$ we have

$$
\begin{aligned}
\mathbb{E}_\xi\left[\|\nabla f_\xi(x) - \nabla f_\xi(x^*)\|^2\right] &= \|\Delta_f(x, x^*)\|^2 + \mathbb{E}_\xi\left[\left\|\frac{1}{m}\sum_{i=1}^m (\xi_i - 1)\Delta_i(x, x^*)\right\|^2\right] \\
&= \|\Delta_f(x, x^*)\|^2 + \frac{1}{m^2}\sum_{i=1}^m \mathbb{E}_\xi[(\xi_i - 1)^2]\|\Delta_i(x, x^*)\|^2 \\
&\overset{(B.10)}{\leq} 2LV_f(x, x^*) \\
&\quad + \frac{1}{m^2}\sum_{i=1}^m \left(\frac{(1-p_i)^2}{p_i} + 1 - p_i\right)\|\Delta_i(x, x^*)\|^2 \\
&\overset{(B.10)}{\leq} 2LV_f(x, x^*) \\
&\quad + \frac{1}{m^2}\sum_{i=1}^m \frac{2L_i(1-p_i)}{p_i} V_{f_i}(x, x^*) \\
&\leq 2\left(L + \max_{i\in[m]} \frac{L_i(1-p_i)}{mp_i}\right) V_f(x, x^*)
\end{aligned}
$$

That is, Assumption 3.4 holds with $\mathcal{L} = \mathcal{L}_{\{p_i\}_{i\in[m]}} \stackrel{\text{def}}{=} L + \max_{i\in[m]} \frac{L_i(1-p_i)}{mp_i}$. In particular, for independent uniform sampling with replacement with averaged batchsize $1 \leq r \leq m$ we have $p_i = r/m$ and Assumption 3.4 holds with $\mathcal{L} = L + \frac{L_{\max}(m-r)}{rm}$, where $L_{\max} = \max_{i\in[m]} L_i$. We notice that $\mathcal{L}$ in this case is better than $\mathcal{L}$ for the mini-batching with replacement.

The above approach has a random batchsize, which can be problematic in some applications. As an alternative, one can use another without replacement sampling with constant batchsize.

r-Nice Sampling

Let r be an integer from $[1, m]$ and random vector $\xi \in \mathbb{R}^m$ have a uniform distribution on the set of m-dimensional vectors from $\{0, m/r\}^m$ having r non-zero components. In other words, the set of indices chosen in the mini-batch is a random set from the uniform distribution on all r-element subsets of $[m]$. By definition we have $\mathbb{E}[\xi_i] = \frac{m}{r} \cdot \frac{r}{m} + 0 \cdot (1 - \frac{r}{m}) = 1$, so, (3.67) holds.

Next, assume that f_i is convex and L_i-smooth for all $i \in \{1, \ldots, m\}$, f is L-smooth. To simplify the following derivation we introduce new notations: $\Delta(x, x^*) =$

$\nabla f(x) - \nabla f(x^*)$, $\Delta_i(x, x^*) = \nabla f_i(x) - \nabla f_i(x^*)$, $i \in [m]$. Then,

$$
\begin{aligned}
\mathbb{E}_\xi \left[\|\nabla f_\xi(x) - \nabla f_\xi(x^*)\|^2\right] &= \frac{1}{\binom{m}{r}} \sum_{S\subseteq[m],|S|=r} \left\| \frac{1}{r} \sum_{i\in S} (\nabla f_i(x) - \nabla f_i(x^*)) \right\|^2 \\
&= \frac{1}{r^2\binom{m}{r}} \sum_{S\subseteq[m],|S|=r} \sum_{i\in S} \|\Delta_i(x,x^*)\|^2 \\
&\quad + \frac{2}{r^2\binom{m}{r}} \sum_{S\subseteq[m],|S|=r} \sum_{i,j\in S, i<j} \langle \Delta_i(x,x^*), \Delta_j(x,x^*)\rangle \\
&= \frac{1}{rm} \sum_{i=1}^m \|\Delta_i(x,x^*)\|^2 \\
&\quad + \frac{2(r-1)}{rm(m-1)} \sum_{1\le i<j\le m} \langle \Delta_i(x,x^*), \Delta_j(x,x^*)\rangle \\
&= \frac{1}{rm} \sum_{i=1}^m \|\Delta_i(x,x^*)\|^2 \\
&\quad + \frac{r-1}{r(m-1)} \sum_{i=1}^m \left\langle \Delta_i(x,x^*), \frac{1}{m} \sum_{j\neq i, j\in[m]} \Delta_j(x,x^*) \right\rangle \\
&= \frac{1}{rm} \sum_{i=1}^m \|\Delta_i(x,x^*)\|^2 - \frac{r-1}{rm(m-1)} \sum_{i=1}^m \|\Delta_i(x,x^*)\|^2 \\
&\quad + \frac{r-1}{r(m-1)} \sum_{i=1}^m \langle \Delta_i(x,x^*), \Delta(x,x^*)\rangle \\
&= \frac{m-r}{rm(m-1)} \sum_{i=1}^m \|\Delta_i(x,x^*)\|^2 + \frac{m(r-1)}{r(m-1)} \|\Delta(x,x^*)\|^2 \\
&\overset{(B.10)}{\le} \frac{2(m-r)}{rm(m-1)} \sum_{i=1}^m L_i V_{f_i}(x,x^*) + \frac{2Lm(r-1)}{r(m-1)} V_f(x,x^*) \\
&\le 2\left(\frac{m(r-1)}{r(m-1)} L + \frac{m-r}{rm(m-1)} L_{\max}\right) V_f(x,x^*),
\end{aligned}
$$

where $L_{\max} = \max_{i\in[m]} L_i$. That is, Assumption 3.4 holds with $\mathcal{L} = \mathcal{L}_{r\text{-nice}} \overset{\text{def}}{=} \frac{m(r-1)}{r(m-1)} L + \frac{m-r}{rm(m-1)} L_{\max}$. We notice that $\mathcal{L}_{r\text{-nice}}$ is smaller than $\mathcal{L}_{\text{US},r}$.

For more examples we refer to [**?**]. Overall, these examples illustrate the generality of the approach described above. Next, we show that this approach fits Assumption 3.3.

Lemma 3.6 *Let Assumption 3.4 hold. Then, the iterates of* SGD *with* $g^k = \nabla f_{\xi^k}(x^k)$ *satisfy*

$$\mathbb{E}_k\left[\|g^k - \nabla f(x^*)\|^2\right] \leq 4\mathcal{L}V_f(x^k, x^*) + 2\sigma_*^2,$$

where $\sigma_*^2 = \mathbb{E}_\xi\left[\|\nabla f_\xi(x^*) - \nabla f(x^*)\|^2\right]$.

Proof Using Young's inequality $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, $\forall a, b \in \mathbb{R}^n$, we derive

$$\begin{aligned}
\mathbb{E}_k\left[\|g^k - \nabla f(x^*)\|^2\right] &= \mathbb{E}_{\xi^k}\left[\|\nabla f_{\xi^k}(x^k) - \nabla f_{\xi^k}(x^*) + \nabla f_{\xi^k}(x^*) - \nabla f(x^*)\|^2\right] \\
&\leq 2\mathbb{E}_{\xi^k}\left[\|\nabla f_{\xi^k}(x^k) - \nabla f_{\xi^k}(x^*)\|^2\right] \\
&\quad +2\underbrace{\mathbb{E}_{\xi^k}\left[\|\nabla f_{\xi^k}(x^*) - \nabla f(x^*)\|^2\right]}_{\sigma_*^2} \\
&\overset{(3.66)}{\leq} 4\mathcal{L}V_f(x^k, x^*) + 2\sigma_*^2,
\end{aligned}$$

which concludes the proof. □

The above lemma proves that Assumption 3.4 implies that Assumption 3.3 holds with the following parameters:

$$A = 2\mathcal{L}, \quad B = 0, \quad C = 0, \quad D_1 = 2\sigma_*^2 \overset{\text{def}}{=} 2\mathbb{E}_\xi\left[\|\nabla f_\xi(x^*) - \nabla f(x^*)\|^2\right], \\ \rho = 1, \quad \sigma_k^2 \equiv 0, \quad D_2 = 0. \tag{3.70}$$

We notice that in the special case of uniform sampling the above parameters coincide with the ones from (3.65). Plugging the parameters from (3.70) in Theorems 3.5, 3.6 and Corollaries 3.2, 3.4, 3.6 we obtain several convergence results summarized in the following theorems.

Theorem 3.9 *Let the objective function f be (μ, x^*)-quasi strongly convex with $\mu > 0$, and Assumption 3.4 hold. Assume that the stepsize $\gamma_k = \gamma$ satisfies*

$$0 < \gamma \leq \frac{1}{2\mathcal{L}}. \tag{3.71}$$

Then, for each $k \geq 0$ the iterations of SGD *satisfy*

$$\mathbb{E}\left[\|x^k - x^*\|^2\right] \leq (1-\gamma\mu)^k R_0^2 + \frac{2\gamma\sigma_*^2}{\mu}, \tag{3.72}$$

where $R_0 = \|x^0 - x^*\|$ and $\sigma_*^2 = \mathbb{E}_\xi\left[\|\nabla f_\xi(x^*) - \nabla f(x^*)\|^2\right]$. Moreover, for any $K \geq 2$ one can choose $\{\gamma_k\}_{k\geq 0}$ as follows:

$$\begin{aligned}
\text{if } K \leq \frac{2\mathcal{L}}{\mu}, &\quad \gamma_k = \frac{1}{2\mathcal{L}}, \\
\text{if } K > \frac{2\mathcal{L}}{\mu} \text{ and } k < k_0, &\quad \gamma_k = \frac{1}{2\mathcal{L}}, \\
\text{if } K > \frac{2\mathcal{L}}{\mu} \text{ and } k \geq k_0, &\quad \gamma_k = \frac{2}{4\mathcal{L} + \mu(k - k_0)},
\end{aligned}$$

where $k_0 = \lceil K/2 \rceil$. *For this choice of* γ_k *the iterates produced by* SGD *satisfy*

$$\mathbb{E}\left[\|x^{K+1} - x^*\|^2\right] \leq 128R_0^2 \exp\left(-\frac{\mu}{4\mathcal{L}}K\right) + \frac{144\sigma_*^2}{\mu^2 K}. \tag{3.73}$$

Proof The result follows from Theorem 3.5 and Corollary 3.2. □

The result above is almost identical to Theorem 3.7. Neglecting numerical constants, the key difference is that instead of smoothness constant L of f and the uniform upper bound for the variance σ^2 the above result depends on the expected smoothness constant $\mathcal{L}$ and the variance at the solution σ_*^2. First, σ_*^2 can be much smaller than σ^2 and σ_*^2 is finite whenever the variance of the stochastic gradient exists at each $x \in \mathbb{R}^d$, e.g., for the finite-sum problems. In contrast, σ^2 can be infinite even for simple problems like linear regression.

However, $L \leq \mathcal{L}$ and the difference can be significant even for importance sampling (see the discussion of the example from (3.69)). This aspect affects the exponentially decaying term in the convergence rate. That is, via relaxing the assumption on the variance (and improving $O(1/K)$ term), we make the exponentially decaying term worse when we replace Assumption 3.1 with Assumption 3.4.

Next, we compare the rates for SGD with uniform and importance samplings. In the case of the uniform sampling, the bound from (3.73) is

$$\mathbb{E}\left[\|x^{K+1} - x^*\|^2\right] \leq 128R_0^2 \exp\left(-\frac{\mu}{4L_{\max}}K\right) + \frac{144\sigma_{\text{US},*}^2}{\mu^2 K}, \tag{3.74}$$

where $\sigma_{\text{US},*} = \frac{1}{m}\sum_{i=1}^m \|\nabla f_i(x^*) - \nabla f(x^*)\|^2$, and in the case of the importance sampling we have

$$\mathbb{E}\left[\|x^{K+1} - x^*\|^2\right] \leq 128R_0^2 \exp\left(-\frac{\mu}{4\overline{L}}K\right) + \frac{144\sigma_{\text{IS},*}^2}{\mu^2 K}, \tag{3.75}$$

where $\sigma_{\text{IS},*} = \frac{1}{m}\sum_{i=1}^m \frac{\overline{L}}{L_i}\|\nabla f_i(x^*) - \nabla f(x^*)\|^2$. Since $\overline{L} \leq L_{\max}$, importance sampling improves the exponentially decaying term. Regarding the variance term, one cannot compare them directly. However, assuming $\|\nabla f_i(x^*) - \nabla f(x^*)\| \sim L_i$ for all $i \in [m]$, we get that $\sigma_{\text{US},*}^2 \sim \frac{1}{m}\sum_{i=1}^m L_i^2 \overset{\text{def}}{=} \overline{L^2}$ and $\sigma_{\text{IS},*}^2 \sim \frac{1}{m}\overline{L}L_i = (\overline{L})^2$, which can be much smaller than $\overline{L^2}$. For example, when $\nabla f(x^*) = 0$, it is natural to expect that the norm of the gradient at x^* is larger for those summands that have larger smoothness constant. Therefore, in such cases, SGD with importance sampling is strictly better than SGD with uniform sampling.

Regarding the comparison of mini-batching strategies, we consider uniform sampling with replacement and r-nice sampling without replacement. As we already observed before, $\mathcal{L}_{r\text{-nice}} \leq \mathcal{L}_{\text{US},r}$. Moreover, one can verify that for the uniform sampling with replacement $\sigma_*^2 = \sigma_{\text{US},r,*}^2 \overset{\text{def}}{=} \frac{\sigma_{\text{US},*}^2}{r}$ and for r-nice sampling without replacement $\sigma_*^2 = \sigma_{r\text{-nice},*}^2 \overset{\text{def}}{=} \frac{(m-r)\sigma_{\text{US},*}^2}{r(m-1)}$, which is smaller than $\sigma_{\text{US},r,*}^2$. Therefore,

r-nice sampling without replacement is provably better than the standard uniform sampling with replacement.

In the next theorem, we provide the results in term of the functional values.

Theorem 3.10 *Let the objective function f be (μ, x^*)-quasi strongly convex with $\mu \geq 0$, L-smooth, $h(x) \equiv 0$, and Assumption 3.4 hold. Assume that the stepsize $\gamma_k = \gamma$ satisfies*

$$0 < \gamma \leq \frac{1}{4\mathcal{L}}.$$

Then, for each $k \geq 0$ the iterations of the method from (3.34) *satisfy*

$$\mathbb{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq (1-\gamma\mu)^{K+1}\frac{R_0^2}{\gamma} + 2\gamma\sigma_*^2, \text{ when } \mu > 0, \tag{3.76}$$

$$\mathbb{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \frac{R_0^2}{\gamma(K+1)} + 2\gamma\sigma_*^2, \text{ when } \mu = 0, \tag{3.77}$$

where $\overline{x}^K = \frac{1}{W_K}\sum_{k=0}^K w_k x^k$, $w_k = (1-\gamma\mu)^{-(k+1)}$, $W_K = \sum_{k=0}^K w_k$, $R_0 = \|x^0 - x^*\|^2$, $\sigma_*^2 = \mathbb{E}_\xi\left[\|\nabla f_\xi(x^*) - \nabla f(x^*)\|^2\right]$. In particular, if $\mu > 0$, then for any $K \geq 2$ one can choose $\{\gamma_k\}_{k\geq 0}$ and sequence $\{w_k\}_{k\geq 0}$ as follows:

$$\begin{aligned}
&\text{if } K \leq \frac{4\mathcal{L}}{\mu}, \qquad \gamma_k = \tfrac{1}{4\mathcal{L}}, \qquad w_k = \left(1 - \frac{\mu}{4\mathcal{L}}\right)^{-(k+1)},\\
&\text{if } K > \frac{4\mathcal{L}}{\mu} \text{ and } k < k_0, \qquad \gamma_k = \tfrac{1}{4\mathcal{L}}, \qquad w_k = 0,\\
&\text{if } K > \frac{4\mathcal{L}}{\mu} \text{ and } k \geq k_0,\ \gamma_k = \tfrac{2}{8\mathcal{L}+\mu(k-k_0)},\ w_k = \frac{8\mathcal{L}}{\mu} + k - k_0,
\end{aligned}$$

where $k_0 = \lceil K/2 \rceil$. For this choice of γ_k the iterates produced by SGD *satisfy*

$$\mathbb{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq 512\mathcal{L}R_0^2\exp\left(-\frac{\mu}{16\mathcal{L}}K\right) + \frac{144\sigma_*^2}{\mu K}, \tag{3.78}$$

where $\overline{x}^K = \frac{1}{W_K}\sum_{k=0}^K w_k x^k$. Finally, if $\mu = 0$, then for any $K \geq 1$ the iterates produced by SGD *with stepsize*

$$\gamma_k = \gamma = \min\left\{\frac{1}{8\mathcal{L}}, \sqrt{\frac{R_0^2}{\sigma_*^2 K}}\right\}$$

satisfy

$$\mathbb{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \frac{8\mathcal{L}R_0^2}{K} + \frac{2\sigma_* R_0}{\sqrt{K}}, \tag{3.79}$$

where $\overline{x}^K = \frac{1}{K+1}\sum_{k=0}^K x^k$.

Proof The result follows from Theorem 3.6 and Corollaries 3.4, 3.6. □

The result above is almost identical to Theorem 3.8 up to the differences discussed after Theorem 3.9.

3.1.4.4 SGD with Arbitrary Sampling and Variance Reduction

Shortly after formulating Assumption 3.3, we showed that L-SVRG for the finite-sum problems with f_i being convex and L_i-smooth satisfies Assumption 3.3 with the following parameters (see (3.39) and the discussion above):

$$A = 2L_{\max}, \quad B = 2, \quad \sigma_k^2 = \frac{1}{m}\sum_{i=1}^{m} \|\nabla f_i(w^k) - \nabla f_i(x^*)\|^2, \quad D_1 = 0,$$
$$C = pL_{\max}, \quad \rho = p, \quad D_2 = 0.$$

In particular, A and C are proportional to $L_{\max}$, which implies the convergence guarantees that depend on $L_{\max}$. This happens because standard L-SVRG uses uniform sampling. However, similarly to SGD, one can consider L-SVRG with arbitrary sampling. That is, consider the stochastic reformulation (3.68) and the following modification of L-SVRG:

$$\begin{aligned} g^k &= \nabla f_{\xi^k}(x^k) - \nabla f_{\xi^k}(w^k) + \nabla f(w^k), \\ w^{k+1} &= \begin{cases} x^k, & \text{with probability } p, \\ w^k, & \text{with probability } 1-p, \end{cases} \\ x^{k+1} &= \operatorname{prox}_{\gamma_k h}(x^k - \gamma_k g^k), \end{aligned} \tag{3.80}$$

where ξ^k is sampled independently from the previous iterations and the randomness in the update of w^{k+1} is independent from previous steps and ξ^k.

In the next two lemmas, we formally verify that Assumption 3.3 holds in this case as well.

Lemma 3.7 *Let Assumption 3.4 hold. Then, the iterates of* L-SVRG *with g^k defined in* (3.80) *satisfy*

$$\mathbb{E}_k\left[\|g^k - \nabla f(x^*)\|^2\right] \leq 4\mathcal{L}V_f(x^k, x^*) + 2\sigma_k^2,$$

where $\sigma_k^2 = 2\mathcal{L}V_f(w^k, x^*)$.

Proof Using Young's inequality $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, $\forall a, b \in \mathbb{R}^n$, we derive

$$\begin{aligned}\mathbb{E}_k\left[\|g^k - \nabla f(x^*)\|^2\right] &\leq 2\mathbb{E}_k\left[\|\nabla f_{\xi^k}(x^k) - \nabla f_{\xi^k}(x^*)\|^2\right] \\ &\quad +2\mathbb{E}_k\left[\|\nabla f_{\xi^k}(w^k) - \nabla f_{\xi^k}(x^*) - (\nabla f(w^k) - \nabla f(x^*))\|^2\right] \\ &\overset{(3.66)}{\leq} 4\mathcal{L}V_f(x^k, x^*) + 2\mathbb{E}_k\left[\|\nabla f_{\xi^k}(w^k) - \nabla f_{\xi^k}(x^*)\|^2\right] \\ &\overset{(3.66)}{\leq} 4\mathcal{L}V_f(x^k, x^*) + 2 \cdot \underbrace{2\mathcal{L}V_f(w^k, x^*)}_{\sigma_k^2}\end{aligned}$$

which concludes the proof. □

Lemma 3.8 *Let Assumption 3.4 hold. Then, the iterates of* L-SVRG *with g^k defined in* (3.80) *satisfy*

$$\mathbb{E}_k\left[\sigma_{k+1}^2\right] \leq (1-p)\sigma_k^2 + 2p\mathcal{L}V_f(x^k, x^*),$$

where $\sigma_k^2 = 2\mathcal{L}V_f(w^k, x^*)$.

Proof By definition of w^{k+1} we have

$$\begin{aligned}\mathbb{E}_k\left[\sigma_{k+1}^2\right] &= \mathbb{E}_k\left[2\mathcal{L}V_f(w^{k+1}, x^*)\right] \\ &= (1-p) \cdot \underbrace{2\mathcal{L}V_f(w^k, x^*)}_{\sigma_k^2} + 2p\mathcal{L}V_f(x^k, x^*),\end{aligned}$$

which concludes the proof. □

That is, Lemmas 3.7 and 3.8 imply that Assumption 3.3 holds with the following parameters

$$\begin{aligned}A = 2\mathcal{L}, \quad B = 2, \quad \sigma_k^2 = 2\mathcal{L}V_f(w^k, x^*), \quad D_1 = 0, \\ C = p\mathcal{L}, \quad \rho = p, \quad D_2 = 0.\end{aligned} \tag{3.81}$$

Plugging these parameters in Theorems 3.5, 3.6 and Corollaries 3.3, 3.5, 3.6 we obtain several convergence results summarized in the following theorems.

Theorem 3.11 *Let the objective function f be (μ, x^*)-quasi strongly convex with $\mu > 0$, and Assumption 3.4 hold. Assume that the stepsize $\gamma_k = \gamma$ satisfies*

$$0 < \gamma \leq \frac{1}{6\mathcal{L}}. \tag{3.82}$$

Then, for each $k \geq 0$ the iterations of L-SVRG *satisfy*

$$\mathbb{E}\left[\|x^k - x^*\|^2\right] \leq \left(1 - \min\left\{\gamma\mu, \frac{p}{2}\right\}\right)^k V_0, \tag{3.83}$$

where $V_0 = \|x^0 - x^*\|^2 + \sigma_0^2 4\gamma^2/p$. Moreover, if f is additionally L-smooth, then for any $K \geq 0$ and

$$\gamma = \frac{1}{6\mathcal{L}}$$

the iterates produced by L-SVRG *satisfy*

$$\mathbb{E}\left[\|x^{K+1} - x^*\|^2\right] \leq \left(1 + \frac{L}{9p\mathcal{L}}\right) R_0^2 \exp\left(-\min\left\{\frac{\mu}{6\mathcal{L}}, \frac{p}{2}\right\} K\right), \tag{3.84}$$

where $R_0 = \|x^0 - x^*\|$.

Proof The first part (bound (3.83)) follows from Theorem 3.5. Next, Corollary 3.3 implies that for $\gamma = 1/6\mathcal{L}$ the iterates produced by L-SVRG satisfy

$$\mathbb{E}\left[\|x^{K+1} - x^*\|^2\right] \leq \Omega_0^2 \exp\left(-\min\left\{\frac{\mu}{6\mathcal{L}}, \frac{p}{2}\right\} K\right),$$

where $\Omega_0^2 = \|x^0 - x^*\|^2 + \sigma_0^2/9p\mathcal{L}^2$. Using Assumption 3.4, we estimate $\mathbb{E}[\sigma_0^2]$ as follows:

$$\mathbb{E}[\sigma_0^2] = 2\mathcal{L}V_f(x^0, x^*) \leq \mathcal{L}L\|x^0 - x^*\|^2 = \mathcal{L}LR_0^2. \tag{3.85}$$

Plugging this upper bound in the definition of Ω_0^2, we get the result. □

As expected, the above result establishes linear convergence of L-SVRG. From (3.84) it is clear that the smaller $\mathcal{L}$ the better the rate. For example, consider single-element samplings: uniform and importance samplings. In this case, we take $p = 1/m$. Then, for L-SVRG with uniform sampling we have

$$\mathbb{E}\left[\|x^{K+1} - x^*\|^2\right] \leq \left(1 + \frac{L}{9pL_{\max}}\right) R_0^2 \exp\left(-\min\left\{\frac{\mu}{6L_{\max}}, \frac{1}{2m}\right\} K\right), \tag{3.86}$$

and for L-SVRG with importance sampling we have

$$\mathbb{E}\left[\|x^{K+1} - x^*\|^2\right] \leq \left(1 + \frac{L}{9p\overline{L}}\right) R_0^2 \exp\left(-\min\left\{\frac{\mu}{6\overline{L}}, \frac{1}{2m}\right\} K\right). \tag{3.87}$$

As expected, importance sampling improves the convergence rate of the standard L-SVRG. A similar conclusion is valid for the comparison of standard uniform sampling with replacement and r-nice sampling.

In the next theorem, we provide the results in term of the functional values.

Theorem 3.12 *Let the objective function f be (μ, x^*)-quasi strongly convex with $\mu \geq 0$, L-smooth, $h(x) \equiv 0$, and Assumption 3.4 hold. Assume that the stepsize $\gamma_k = \gamma$ satisfies*

$$0 < \gamma \leq \frac{1}{12\mathcal{L}}.$$

Then, for each $k \geq 0$ the iterations of the method from (3.34) *satisfy*

$$\mathbb{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \left(1 - \min\left\{\gamma\mu, \frac{p}{2}\right\}\right)^{K+1} \frac{R_0^2}{\gamma}, \text{ when } \mu > 0, \tag{3.88}$$

$$\mathbb{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \frac{R_0^2}{\gamma(K+1)}, \text{ when } \mu = 0, \tag{3.89}$$

where $\overline{x}^K = \frac{1}{W_K}\sum_{k=0}^K w_k x^k$, $w_k = (1 - \min\{\gamma\mu, p/2\})^{-(k+1)}$, $W_K = \sum_{k=0}^K w_k$, $R_0 = \|x^0 - x^*\|^2$. *In particular, if* $\mu > 0$, *then for any* $K \geq 0$ *and*

$$\gamma = \frac{1}{12\mathcal{L}}$$

the iterates produced by L-SVRG *satisfy*

$$\mathbb{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq (12\mathcal{L} + \frac{L}{3p})R_0^2 \exp\left(-\min\left\{\frac{\mu}{12\mathcal{L}}, \frac{p}{2}\right\}K\right), \tag{3.90}$$

where $R_0 = \|x^0 - x^*\|$. *Finally, if* $\mu = 0$, *then for any* $K \geq 1$ *the iterates produced by* L-SVRG *with stepsize*

$$\gamma_k = \gamma = \min\left\{\frac{1}{12\mathcal{L}}, \sqrt{\frac{pR_0^2}{4\sigma_0^2}}\right\}$$

satisfy

$$\mathbb{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \frac{\left(12\mathcal{L} + \frac{4\sqrt{\mathcal{L}L}}{\sqrt{p}}\right)R_0^2}{K}, \tag{3.91}$$

where $\overline{x}^K = \frac{1}{K+1}\sum_{k=0}^K x^k$.

Proof The first part (bounds (3.88) and (3.89)) follows from from Theorem 3.6. Next, Corollary 3.5 implies

$$\mathbb{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \Omega_0^2 \exp\left(-\min\left\{\frac{\mu}{12\mathcal{L}}, \frac{p}{2}\right\}K\right),$$

where $\Omega_0^2 = 12\mathcal{L}\|x^0 - x^*\|^2 + \sigma_0^2/3p\mathcal{L}$. Using the upper bound (3.85) for σ_0^2 we derive (3.90). Finally, Corollary 3.6 implies

$$\mathbb{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \frac{12\mathcal{L}R_0^2}{K} + \frac{4R_0\sqrt{\sigma_0^2}}{\sqrt{p}K},$$

Plugging the upper bound (3.85) for σ_0^2 we get (3.91). □

When $\mu > 0$ the result is very similar to Theorem 3.11. Let us discuss more the rate in the case when $\mu = 0$. Consider single-element samplings: uniform and importance samplings. In this case, we take $p = 1/m$. Then, for L-SVRG with uniform sampling we have

$$\mathbb{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \frac{\left(12L_{\max} + 4\sqrt{mL_{\max}\overline{L}}\right)R_0^2}{K},$$

and for L-SVRG with importance sampling we have

$$\mathbb{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \frac{\left(12\overline{L} + 4\sqrt{m\overline{L}L}\right)R_0^2}{K}.$$

Besides the superiority of importance sampling over uniform sampling in this case, we also notice that neglecting the differences between smoothness constants the rate of L-SVRG in the convex case is proportional to $\sqrt{m}$. When $L \approx \overline{L}$ or even $L \approx L_{\max}$, L-SVRG has $\sqrt{m}$ times better oracle complexity than GD. However, in the worst case, $\overline{L} = mL$ and the oracle complexity of L-SVRG with importance sampling coincides with the oracle complexity of GD. Finally, when comparing the rate of L-SVRG (see (3.91)) with the rate of SGD (see (3.79)), we observe that the rate of SGD has better $O(1/\kappa)$ term but it also has a slower $O(1/\sqrt{\kappa})$ term, which is dominating when K is sufficiently large. Therefore, when $\mu = 0$, SGD is preferable for finding relatively inaccurate solution in terms of the functional suboptimality, while L-SVRG is better for finding more accurate solutions.

3.1.4.5 SGD and Coordinate-Wise Randomization

In the simplest setup, the analysis of zeroth-order/coordinate methods is very similar to the analysis of stochastic first-order methods like SGD. Here we illustrate this aspect on Randomized Coordinate Descent (RCD) with uniform randomization of sampled components. Further details and examples are deferred to Chapter 5.

RCD can be seen as a special case of SGD (3.34) with g^k chosen as

$$g^k = n\langle \nabla f(x^k), e_{j_k}\rangle e_{j_k} = n[\nabla f(x^k)]_{j_k} e_{j_k}, \tag{3.92}$$

where e_{j_k} is j_k-th element of the standard basis in $\mathbb{R}^n$ and j_k is sampled uniformly at random from $[n]$ independently from the previous iterations. That is, at each iteration of RCD one needs to compute just one directional derivative $[\nabla f(x^k)]_{j_k}$ instead of the full gradient.

As we already noticed, despite the seeming differences between RCD and versions of SGD we consider before, RCD is in fact a special case of SGD. Indeed, estimator g^k defined in (3.92) is an unbiased estimator of $\nabla f(x^k)$:

$$\mathbb{E}_k\left[g^k\right] = \sum_{i=1} n[\nabla f(x^k)]_i e_i \cdot \frac{1}{n} = \sum_{i=1} [\nabla f(x^k)]_i e_i = \nabla f(x^k).$$

Moreover, for convex and L-smooth f we have the following result.

Lemma 3.9 *Let f be convex and L-smooth. Then, for all $k \geq 0$ the iterates produced by* RCD *satisfy*

$$\mathbb{E}_k\left[\|g^k - \nabla f(x^*)\|^2\right] \leq 4nLV_f(x^k, x^*) + 2n\|\nabla f(x^*)\|^2.$$

Proof Using Young's inequality $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, $\forall a, b \in \mathbb{R}^n$, we derive

$$\begin{aligned}
\mathbb{E}_k\left[\|g^k - \nabla f(x^*)\|^2\right] &= \mathbb{E}_k\left[\|n[\nabla f(x^k)]_{j_k} e_{j_k} \pm n[\nabla f(x^*)]_{j_k} e_{j_k} - \nabla f(x^*)\|^2\right] \\
&\leq 2n^2\mathbb{E}_k\left[\|[\nabla f(x^k)]_{j_k} e_{j_k} - [\nabla f(x^*)]_{j_k} e_{j_k}\|^2\right] \\
&\quad +2\mathbb{E}_k\left[\|n[\nabla f(x^*)]_{j_k} e_{j_k} - \nabla f(x^*)\|^2\right] \\
&\leq 2n\sum_{i=1}^{n}\left([\nabla f(x^k)]_i - [\nabla f(x^*)]_i\right)^2 \\
&\quad +2\mathbb{E}_k\left[\|n[\nabla f(x^*)]_{j_k} e_{j_k}\|^2\right] \\
&= 2n\|\nabla f(x^k) - \nabla f(x^*)\|^2 + 2n\sum_{i=1}^{n}[\nabla f(x^*)]_i^2 \\
&\overset{(B.10)}{\leq} 4nLV_f(x^k, x^*) + 2n\|\nabla f(x^*)\|^2,
\end{aligned}$$

which concludes the proof. □

That is, the above lemma implies that Assumption 3.3 holds with the following parameters:

$$\begin{aligned}
&A = 2nL, \quad B = 0, \quad \sigma_k^2 \equiv 0, \quad D_1 = 2n\|\nabla f(x^*)\|, \\
&\qquad C = 0, \quad \rho = 1, \quad D_2 = 0.
\end{aligned} \tag{3.93}$$

Plugging these parameters in Theorems 3.5, 3.6 and Corollaries 3.2, 3.4, 3.6 we obtain several convergence results summarized in the following theorems.

Theorem 3.13 *Let the objective function f be (μ, x^*)-quasi strongly convex with $\mu > 0$, convex, and L-smooth. Assume that the stepsize $\gamma_k = \gamma$ satisfies*

$$0 < \gamma \leq \frac{1}{2nL}. \tag{3.94}$$

Then, for each $k \geq 0$ the iterations of RCD *satisfy*

$$\mathbb{E}\left[\|x^k - x^*\|^2\right] \leq (1-\gamma\mu)^k R_0^2 + \frac{2\gamma n\|\nabla f(x^*)\|^2}{\mu}, \tag{3.95}$$

where $R_0 = \|x^0 - x^*\|$. Moreover, for any $K \geq 2$ one can choose $\{\gamma_k\}_{k\geq 0}$ as follows:

$$\begin{aligned}
&\text{if } K \leq \frac{2nL}{\mu}, && \gamma_k = \frac{1}{2nL}, \\
&\text{if } K > \frac{2nL}{\mu} \text{ and } k < k_0, && \gamma_k = \frac{1}{2nL}, \\
&\text{if } K > \frac{2nL}{\mu} \text{ and } k \geq k_0, && \gamma_k = \frac{2}{4nL + \mu(k - k_0)},
\end{aligned}$$

where $k_0 = \lceil K/2 \rceil$. *For this choice of* γ_k *the iterates produced by* RCD *satisfy*

$$\mathbb{E}\left[\|x^{K+1} - x^*\|^2\right] \leq 128R_0^2 \exp\left(-\frac{\mu}{4nL}K\right) + \frac{144n\|\nabla f(x^*)\|^2}{\mu^2 K}. \tag{3.96}$$

Proof The result follows from Theorem 3.5 and Corollary 3.2. □

Similarly to SGD, RCD with constant stepsize converges to the neighborhood of the solution and the radius of this neighborhood is proportional to $\sqrt{n}\|\nabla f(x^*)\|$. When $\nabla f(x^*) = 0$ the method converges linearly, but for general composite problems with $\nabla f(x^*) \neq 0$ it does not. Moreover, the bound from (3.96) shows that RCD has n times slower exponentially decaying term in comparison to the standard GD. This phenomenon is expected since RCD computes only one partial derivative at each iteration, while GD needs n partial derivatives per step, i.e., one gradient evaluation.

In the next theorem, we provide the results in term of the functional values.

Theorem 3.14 *Let the objective function* f *be* (μ, x^*)*-quasi strongly convex with* $\mu \geq 0$*, convex,* L*-smooth and* $h(x) \equiv 0$. *Assume that the stepsize* $\gamma_k = \gamma$ *satisfies*

$$0 < \gamma \leq \frac{1}{4nL}.$$

Then, for each $k \geq 0$ *the iterations of* RCD *satisfy*

$$\mathbb{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq (1 - \gamma\mu)^{K+1}\frac{R_0^2}{\gamma}, \text{ when } \mu > 0, \tag{3.97}$$

$$\mathbb{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \frac{R_0^2}{\gamma(K+1)}, \text{ when } \mu = 0, \tag{3.98}$$

where $\overline{x}^K = \frac{1}{W_K}\sum_{k=0}^K w_k x^k$, $w_k = (1-\gamma\mu)^{-(k+1)}$, $W_K = \sum_{k=0}^K w_k$, $R_0 = \|x^0 - x^*\|^2$. *In particular, if* $\mu > 0$*, then for any* $K \geq 0$ *and for* $\gamma_k = \gamma = 1/4nL$ *the iterates produced by* RCD *satisfy*

$$\mathbb{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq 4nLR_0^2 \exp\left(-\frac{\mu}{4nL}K\right), \tag{3.99}$$

where $\overline{x}^K = \frac{1}{W_K}\sum_{k=0}^K w_k x^k$. *Finally, if* $\mu = 0$*, then for any* $K \geq 1$ *the iterates produced by* RCD *with stepsize* $\gamma_k = \gamma = 1/4nL$ *satisfy*

$$\mathbb{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \frac{8nLR_0^2}{K}, \tag{3.100}$$

where $\overline{x}^K = \frac{1}{K+1}\sum_{k=0}^K x^k$.

Proof The result follows from Theorem 3.6 and Corollaries 3.4, 3.6. We also take into account that $h(x) \equiv 0$ implying $\nabla f(x^*) = 0$. □

The above theorem shows that RCD is n times slower rate than GD, when $h(x) \equiv 0$. In the discussion after Theorem 3.13, we explain why this is natural.

However, sub-linear convergence of RCD for general composite problems ($\nabla f(x^*) \neq 0$) is an undesired phenomenon. This is the case because the variance of the stochastic estimator at the solution is not zero. Indeed, let j be sampled uniformly at random from $[n]$. Then, we have

$$\begin{aligned}\mathbb{E}\left[\|n[\nabla f(x^*)]_j e_j - \nabla f(x^*)\|^2\right] &= \mathbb{E}\left[\|n[\nabla f(x^*)]_j e_j\|^2\right] - \|\nabla f(x^*)\|^2 \\ &= (n-1)\|\nabla f(x^*)\|^2,\end{aligned}$$

which is not zero when $n > 1$ (otherwise RCD reduces to GD) and $\nabla f(x^*) \neq 0$.

We notice that SGD has a similar issue: the variance of the plain stochastic gradient estimator is not zero at the solution. This issue is addressed via using variance reduction mechanism. Therefore, it is natural to apply a similar technique to RCD. In particular, we consider the following estimator:

$$g^k = n\left([\nabla f(x^k)]_{j_k} - [\nabla f(w^k)]_{j_k}\right) e_{j_k} + \nabla f(w^k), \tag{3.101}$$

$$w^{k+1} = \begin{cases} x^k, & \text{with probability } p, \\ w^k, & \text{with probability } 1-p, \end{cases}$$

where j_k is sampled uniformly at random from $[n]$ independently from the previous iterations. SGD (3.34) with estimator g^k from (3.101) is called Loopless Variance Reduced Coordinate Descent (L-VRCD). L-VRCD is very similar to L-SVRG: instead of sampling the random function from the sum, L-VRCD samples random component. Up to this difference L-VRCD and L-SVRG are identical. In particular, w^k in L-VRCD is a point where full gradient computations happen with probability p. Choosing $p = 1/n$, we get that the expected cost of one iteration of L-VRCD is 2 partial derivatives computations, which is of the same order as for RCD. Moreover, L-VRCD also fits Assumption 3.3. The two following lemmas show this.

Lemma 3.10 *Let f be convex and L-smooth. Then, for all $k \geq 0$ the iterates produced by* L-VRCD *satisfy*

$$\mathbb{E}_k\left[\|g^k - \nabla f(x^*)\|^2\right] \leq 4nLV_f(x^k, x^*) + 2n\sigma_k^2,$$

where $\sigma_k^2 = 2LV_f(w^k, x^*)$.

Proof Using Young's inequality $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, $\forall a, b \in \mathbb{R}^n$, we derive

$$\begin{aligned}
\mathbb{E}_k\left[\|g^k - \nabla f(x^*)\|^2\right] &= \mathbb{E}_k\left[\|n([\nabla f(x^k)]_{j_k} - [\nabla f(w^k)]_{j_k})e_{j_k} - \nabla f(x^*)\|^2\right] \\
&\leq 2n^2\mathbb{E}_k\left[\|[\nabla f(x^k)]_{j_k}e_{j_k} - [\nabla f(x^*)]_{j_k}e_{j_k}\|^2\right] \\
&\quad +2\mathbb{E}_k\left[\|n[\nabla f(w^k) - \nabla f(x^*)]_{j_k}e_{j_k} - (\nabla f(w^k) - \nabla f(x^*))\|^2\right] \\
&\leq 2n\sum_{i=1}^{n}[\nabla f(x^k) - \nabla f(x^*)]_{j_k}^2 \\
&\quad +2\mathbb{E}_k\left[\|n[\nabla f(w^k) - \nabla f(x^*)]_{j_k}e_{j_k}\|^2\right] \\
&= 2n\|\nabla f(x^k) - \nabla f(x^*)\|^2 + 2n\|\nabla f(w^k) - \nabla f(x^*)\|^2 \\
&\overset{(B.10)}{\leq} 4nLV_f(x^k, x^*) + 2n\cdot\underbrace{2LV_f(w^k, x^*)}_{\sigma_k^2},
\end{aligned}$$

which concludes the proof. □

Lemma 3.11 *Let f be convex and L-smooth. Then, for all $k \geq 0$ the iterates produced by* L-VRCD *satisfy*

$$\mathbb{E}_k\left[\sigma_{k+1}^2\right] \leq (1-p)\sigma_k^2 + 2pLV_f(x^k, x^*),$$

where $\sigma_k^2 = 2LV_f(w^k, x^*)$.

Proof By definition of w^{k+1} we have

$$\begin{aligned}
\mathbb{E}_k\left[\sigma_{k+1}^2\right] &= \mathbb{E}_k\left[2LV_f(w^{k+1}, x^*)\right] \\
&= (1-p)\cdot\underbrace{2LV_f(w^k, x^*)}_{\sigma_k^2} + p\cdot 2LV_f(x^k, x^*),
\end{aligned}$$

which concludes the proof. □

That is, Lemmas 3.10 and 3.11 imply that Assumption 3.3 holds with the following parameters

$$\begin{aligned}
A = 2nL, \quad B = 2n, \quad \sigma_k^2 = 2LV_f(w^k, x^*), \quad D_1 = 0, \\
C = pL, \quad \rho = p, \quad D_2 = 0.
\end{aligned} \tag{3.102}$$

Plugging these parameters in Theorems 3.5, 3.6 and Corollaries 3.3, 3.5, 3.6 we obtain several convergence results summarized in the following theorems.

Theorem 3.15 *Let the objective function f be convex, (μ, x^*)-quasi strongly convex with $\mu > 0$, and L-smooth. Assume that the stepsize $\gamma_k = \gamma$ satisfies*

$$0 < \gamma \leq \frac{1}{6nL}. \tag{3.103}$$

Then, for each $k \geq 0$ the iterations of L-VRCD *satisfy*

$$\mathbb{E}\left[\|x^k - x^*\|^2\right] \leq \left(1 - \min\left\{\gamma\mu, \frac{p}{2}\right\}\right)^k V_0, \tag{3.104}$$

where $V_0 = \|x^0 - x^*\|^2 + \sigma_0^2 4n\gamma^2/p$. *Moreover, for any* $K \geq 0$ *and*

$$\gamma = \frac{1}{6nL}$$

the iterates produced by L-VRCD *satisfy*

$$\mathbb{E}\left[\|x^{K+1} - x^*\|^2\right] \leq \left(1 + \frac{1}{9pn}\right) R_0^2 \exp\left(-\min\left\{\frac{\mu}{6nL}, \frac{p}{2}\right\} K\right), \tag{3.105}$$

where $R_0 = \|x^0 - x^*\|$.

Proof The first part (bound (3.104)) follows from Theorem 3.5. Next, Corollary 3.3 implies that for $\gamma = 1/6nL$ the iterates produced by L-VRCD satisfy

$$\mathbb{E}\left[\|x^{K+1} - x^*\|^2\right] \leq \Omega_0^2 \exp\left(-\min\left\{\frac{\mu}{6nL}, \frac{p}{2}\right\} K\right),$$

where $\Omega_0^2 = \|x^0 - x^*\|^2 + \sigma_0^2/9pnL^2$. Using L-smoothness of f, we estimate $\mathbb{E}[\sigma_0^2]$ as follows:

$$\mathbb{E}[\sigma_0^2] = 2LV_f(x^0, x^*) \leq L^2\|x^0 - x^*\|^2 = L^2 R_0^2. \tag{3.106}$$

Plugging this upper bound in the definition of Ω_0^2, we get the result. □

In contrast to RCD, L-VRCD converges linearly even when $\nabla f(x^*) \neq 0$. The rate of convergence of L-VRCD is n times slower (if we take $p = 1/n$, then $\min\{\mu/6nL, p/2\} = \mu/6nL$) than in the case of GD, while L-VRCD requires $O(n)$ times fewer partial derivative computations per step (in expectation, when $p \sim 1/n$) than GD.

In the next theorem, we provide the results in term of the functional values.

Theorem 3.16 *Let the objective function* f *be convex,* (μ, x^*)*-quasi strongly convex with* $\mu > 0$*, and* L*-smooth, and* $h(x) \equiv 0$*. Assume that the stepsize* $\gamma_k = \gamma$ *satisfies*

$$0 < \gamma \leq \frac{1}{12nL}.$$

Then, for each $k \geq 0$ *the iterations of* L-VRCD *satisfy*

$$\mathbb{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \left(1 - \min\left\{\gamma\mu, \frac{p}{2}\right\}\right)^{K+1} \frac{R_0^2}{\gamma}, \text{ when } \mu > 0, \tag{3.107}$$

$$\mathbb{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \frac{R_0^2}{\gamma(K+1)}, \text{ when } \mu = 0, \tag{3.108}$$

where $\overline{x}^K = \frac{1}{W_K}\sum_{k=0}^K w_k x^k$, $w_k = (1 - \min\{\gamma\mu, p/2\})^{-(k+1)}$, $W_K = \sum_{k=0}^K w_k$, $R_0 = \|x^0 - x^*\|^2$. *In particular, if* $\mu > 0$*, then for any* $K \geq 0$ *and*

$$\gamma = \frac{1}{12nL}$$

the iterates produced by L-VRCD *satisfy*

$$\mathbb{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \left(12nL + \frac{L}{3p}\right) R_0^2 \exp\left(-\min\left\{\frac{\mu}{12nL}, \frac{p}{2}\right\} K\right), \quad (3.109)$$

where $R_0 = \|x^0 - x^*\|$. *Finally, if* $\mu = 0$, *then for any* $K \geq 1$ *the iterates produced by* L-VRCD *with stepsize*

$$\gamma_k = \gamma = \min\left\{\frac{1}{12nL}, \sqrt{\frac{pR_0^2}{4\sigma_0^2}}\right\}$$

satisfy

$$\mathbb{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \frac{\left(12nL + \frac{4L\sqrt{n}}{\sqrt{p}}\right) R_0^2}{K}, \quad (3.110)$$

where $\overline{x}^K = \frac{1}{K+1}\sum_{k=0}^K x^k$.

Proof The first part (bounds (3.107) and (3.108)) follows from from Theorem 3.6. Next, Corollary 3.5 implies

$$\mathbb{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \Omega_0^2 \exp\left(-\min\left\{\frac{\mu}{12nL}, \frac{p}{2}\right\} K\right),$$

where $\Omega_0^2 = 12nL\|x^0 - x^*\|^2 + \sigma_0^2/3pL$. Using the upper bound (3.85) for σ_0^2 we derive (3.109). Finally, Corollary 3.6 implies

$$\mathbb{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \frac{12nLR_0^2}{K} + \frac{4R_0\sqrt{\sigma_0^2}}{\sqrt{p}K},$$

Plugging the upper bound (3.106) for σ_0^2 we get (3.110). □

The implications from the above result are almost the same as for RCD in the same setup: since $\nabla f(x^*) = 0$, variance reduction does not change the rate in this case.

3.1.5 Convergence of SGD for Over-Parameterized Models

Recent advances in Deep Learning are associated with training of very large neural networks having billions of parameters. In such cases, the resulting model can be such powerful that it can perfectly fit the training dataset, i.e., one can achieve a zero

loss on the training data. Surprisingly, but this does not make the generalization of such models worse, e.g., see [**?**, **?**, **?**]. Moreover, under certain assumptions on the model, the corresponding loss function becomes convex-like, i.e., the loss function is non-convex but satisfies some condition that holds for (strongly) convex functions, for example, PŁ*-condition [**?**]. Typically, under such assumptions standard methods like SGD converge as in the (strongly) convex case. Therefore, assuming (strong) convexity of the objective function is a meaningful simplification for the analysis of stochastic methods for such models that are usually called *over-parameterized*.

That is, consider the finite-sum minimization problem (3.16), where functions f_i satisfy so-called *interpolation condition*: there exists $x^* \in \mathbb{R}^n$ such that

$$f_i(x^*) = \min_{x \in \mathbb{R}^n} f_i(x) \quad \forall i \in [m]. \tag{3.111}$$

In other words, we assume that there exists a common optimum for all terms in the finite-sum. Taking into account that in the case of empirical risk minimization $f_i(x)$ corresponds to the loss of the model for the i-th data point, one can see that the interpolation condition implies that the model interpolates the training data perfectly.

Interpolation condition (3.111) implies useful properties for the analysis of the methods. In particular, if f_i is L_i-smooth for all $i \in [m]$, then for j being sampled uniformly at random from $[m]$ we have

$$\begin{aligned}\mathbb{E}_j\left[\|\nabla f_j(x)\|^2\right] &= \frac{1}{m}\sum_{i=1}^m \|\nabla f_i(x)\|^2 \leq \frac{2}{m}\sum_{i=1}^m L_i\left(f_i(x) - f_i(x^*)\right) \\ &\leq 2L_{\max}\left(f(x) - f(x^*)\right). \end{aligned} \tag{3.112}$$

This means that in this setup SGD with uniform sampling satisfies Assumption 3.3 with the following parameters:

$$A = L_{\max}, \quad B = 0, \quad \sigma_k^2 \equiv 0, \quad D_1 = 0, \quad C = 0, \quad \rho = 1, \quad D_2 = 0. \tag{3.113}$$

We notice here that $D_1 = 0$ and $D_2 = 0$ meaning that SGD *with constant stepsize* converges to the exact solution asymptotically in expectation (see Theorems 3.5 and 3.6). In particular, when f is quasi-strongly convex SGD converges linearly in this setup. Previously, we considered the variance reduction mechanism as a tool to achieve this property. Therefore, from this perspective variance reduction is not needed for over-parameterized models. This observation is also supported by the fact that variance reduced methods do not behave well when they are combined with some tricks popular in deep learning [**?**].

Going back to the analysis of SGD under over-parameterization, we introduce an assumption relaxing (3.112).

Assumption 3.5 (Relaxed Weak Growth Condition (R-WGC)) There exist constants $\mathcal{L} > 0$, $\eta > 0$, and $\sigma_* \geq 0$ such that for all $x \in \mathbb{R}^n$

$$\mathbb{E}_j\left[\|\nabla f_j(x)\|^2\right] \leq 2\mathcal{L}\eta\left(f(x) - f(x^*)\right) + \sigma_*^2, \tag{3.114}$$

where j is sampled uniformly at random from $[m]$.

If $\sigma_* = 0$, then condition (3.114) is called Weak Growth Condition (WGC). If all functions f_i are convex, then WGC implies interpolation condition (3.111): indeed, we have

$$\mathbb{E}_j \left[\|\nabla f_j(x^*)\|^2 \right] \leq 2\mathcal{L}\eta \left(f(x^*) - f(x^*) \right) = 0,$$

implying $\nabla f_i(x^*) = 0$ for all $i \in [m]$, i.e., $f_i(x^*) = \min_{x \in \mathbb{R}^n} f_i(x)$ for all $i \in [m]$ due to convexity.

Therefore, constant σ_*^2 from (3.114) can be seen as a measure of how over-parameterized the model is: the smaller σ_*^2, the more over-parameterized the model. Moreover, when each f_i is L_i-smooth and has a minimizer x_i^*, i.e., $x_i^* = \min_{x \in \mathbb{R}^n} f_i(x)$, we have

$$\begin{aligned}
\mathbb{E}_j \left[\|\nabla f_j(x)\|^2 \right] &= \frac{1}{m} \sum_{i=1}^m \|\nabla f_i(x)\|^2 \leq \frac{2}{m} \sum_{i=1}^m L_i \left(f_i(x) - f_i(x_i^*) \right) \\
&= \frac{2}{m} \sum_{i=1}^m L_i \left(f_i(x) - f_i(x^*) \right) + \frac{2}{m} \sum_{i=1}^m L_i \left(f_i(x^*) - f_i(x_i^*) \right) \\
&\leq 2L_{\max} \left(f(x) - f(x^*) \right) + \frac{2}{m} \sum_{i=1}^m L_i \left(f_i(x^*) - f_i(x_i^*) \right),
\end{aligned}$$

i.e., R-WGC holds with $\mathcal{L} = L_{\max}$, $\eta = 1$, $\sigma_*^2 = \frac{2}{m} \sum_{i=1}^m L_i (f_i(x^*) - f_i(x_i^*))$. In this case, $\sigma_*^2 = 0$ iff $f_i(x^*) = \min_{x \in \mathbb{R}^n} f_i(x)$ for all $i \in [m]$. One can also note that up to the smoothness constants σ_*^2 measures functional sub-optimality of x^* for all summands on average.

The discussion above explains why Assumption 3.5 is adequate for describing over-parameterized or almost over-parameterized models. Moreover, this assumption also perfectly fits the general framework we consider earlier in this Chapter. That is, Assumption 3.5 implies that Assumption 3.3 holds the following parameters:

$$A = \mathcal{L}\eta, \quad B = 0, \quad \sigma_k^2 \equiv 0, \quad D_1 = \sigma_*^2, \quad C = 0, \quad \rho = 1, \quad D_2 = 0. \quad (3.115)$$

Eduard: formulate the results, say few words about non-smoothness (compositional one)

3.1.6 Convergence of SGD with and without Averaging

Eduard: mention here that SGD has different rates for the last iterate and for the averaged iterate (paper by Taylor and Bach), say there that one can construct potentials via computer-assisted analysis; say that averaging makes the limit distribution normal, mention the results about asymptotic normality and non-normality. I guess, here we will not have much proofs.

3.1.7 Convergence of SGD under Structured Non-Convexity

Eduard: I am not sure that we need much details here. Maybe we can just mention that one can derive some results under PL and other convex-like assumptions. We can say that the analysis is very similar to what we discussed in this part.

3.2 Catalyst: a universal framework for acceleration of randomized optimization methods

From given starting point $x^0 \in \mathbb{R}^n$ Catalyst generates the points $\{x^k\}_{k\geq 0}$ solving the inner problem

$$x^k \approx \arg\min_{x\in\mathbb{R}^n} \left\{ f(x) + \frac{\kappa}{2}\left\|x - y^{k-1}\right\|^2 \right\} \tag{3.116}$$

with method $\mathcal{M}$ and compute auxiliary variable y^k using Nesterov's extrapolation step

$$y^k = x^k + \beta_k (x^k - x^{k-1}), \tag{3.117}$$

with $\beta_k = \frac{\alpha_{k-1}(1-\alpha_{k-1})}{\alpha_{k-1}^2+\alpha_k}$, where $\alpha_k \in (0,1)$ and updated from equation $\alpha_k^2 = (1-\alpha_k)\alpha_{k-1}^2 + q\alpha_k$. That is, at iteration k method compute approximately solution of the problem

3.2.1 Analysis of standard Catalyst algorithm

Before...

The Moreau envelope results from adding the quadratic regularization to f:

$$F(x) := \min_{z\in\mathbb{R}^n} \left\{ f(z) + \frac{\kappa}{2}\|z-x\|^2 \right\}$$

where κ is the positive parameter. The proximal operator is the unique minimizer of , that is,

$$d(x) := \text{prox}_{f/\kappa}(x) = \arg\min_{z\in\mathbb{R}^n} \left\{ f(z) + \frac{\kappa}{2}\|z-x\|^2 \right\}.$$

The next proposition characterizes the smoothing effect of the Moreau Envelope.

Proposition 3.1 (Regularization properties of the Moreau Envelope) *Given a convex continuous function f and a regularization parameter $\kappa > 0$, consider the Moreau Envelope F. Then,*

1. *F is convex and minimizing f and F are equivalent in the sense that*

$$\min_{x\in\mathbb{R}^n} F(x) = \min_{x\in\mathbb{R}^n} f(x).$$

Moreover the solution set of the two above problems coincide with each other.

2. *F is continuously differentiable even f is not and:*

$$\nabla F = \kappa(x - d(x)).$$

Moreover F is L_F-smooth with constant $L_F = \kappa$.

3. *If f is μ-strongly convex, then F is μ_F-strongly convex with constant $\mu_F = \frac{\mu\kappa}{\mu+\kappa}$.*

$$y^{k+1} = y^k - \frac{1}{L_F}\nabla F(y^k) \quad \text{and} \quad y^{k+1} = x^{k+1} + \beta_{k+1}(x^{k+1} - x^k),$$

where β_{k+1} is Nesterov's extrapolation parameter. By noticing that $\nabla F(y) = \kappa(y - d(y))$ and $L_F = \kappa$, we obtain in fact

$$x^{k+1} = d(y^k) \quad \text{and} \quad y^{k+1} = x^{k+1} + \beta_{k+1}(x^{k+1} - x^k),$$

which is known as the accelerated proximal point algorithm.

Given a proximal center x, a smoothing parameter κ, and an accuracy $\varepsilon > 0$, we denote the set of ε-approximations of the proximal operator $d(x)$ by

$$d^{\varepsilon}(x) := \{z \in \mathbb{R}^n \text{ s.t. } h(z) - h(z^*) \le \varepsilon\}, \text{ where } h(z) = f(z) + \frac{\kappa}{2}\|x - z\|^2$$

and $h(z^*)$ is the minimum function value of $h(z)$.

Theorem 3.17 *Consider the sequences $\{x^k\}_{k\ge 0}$ and $\{y^k\}_{k\ge 0}$ produced by* Catalyst*, assuming that x^k is in $d^{\varepsilon_k}(y^{k-1})$ for all $k \ge 1$. Then,*

$$f(x^k) - f(x^*) \le A_{k-1}\left(\sqrt{(1-\alpha_0)(f(x^0) - f(x^*)) + \frac{\gamma_0}{2}\|x^* - x^0\|^2} + 3\sum_{j=1}^{k}\sqrt{\frac{\varepsilon_j}{A_{j-1}}}\right)^2,$$

where

$$\gamma_0 = (\kappa + \mu)\alpha_0(\alpha_0 - q) \text{ and } A_k = \prod_{j=1}^{k}(1 - \alpha_j) \text{ with } A_0 = 1.$$

Proposition 3.2 *In* Catalyst *algorithm, choose $\alpha_0 = \sqrt{q}$ and*

$$\varepsilon_k = \frac{2}{9}\left(f(x^0) - f(x^*)\right)(1-\rho)^k, \text{ with } \rho < \sqrt{q}.$$

Then, the sequence of iterates $\{x^k\}_{k\ge 0}$ satisfies

$$f(x^k) - f(x^*) \le \frac{8}{(\sqrt{q} - \rho)^2}(1-\rho)^{k+1}(f(x^0) - f(x^*)).$$

3.2.2 Modifications and generalizations of Catalyst

TODO for Eduard: polish the section once the previous section is finished.

Note that in the both versions of Catalyst presented above we faced the need to use $\tilde{O}(\cdot)$ notation to write out their convergence rates. The weak point of this analysis is that it inevitably burdens the estimates with the factor of $O(\log \frac{1}{\varepsilon})$, due to which the estimates of Catalyst-accelerated methods are slightly more complex than those of individual accelerated versions of these methods. However, there is a way to overcome this complication. It turns out that one can select another stopping condition in acceleration envelope so that the convergence rate of Catalyst with this condition is free of logarithmic factors.

TODO for Dmitry P.: once Section 3.2.1 is finished, one needs to check the consitence of the notation.

Let us consider the following Monteiro–Svaiter stopping condition:

$$\|\nabla h_k(x)\|_2 \leq \frac{\kappa}{2}\|x - y^{k-1}\|_2. \tag{3.118}$$

If we use this condition in a slightly modified Catalyst algorithm, the convergence rate is still asymptotically the same as in Section 3.2.1 (we take it without proof). But we can transform the condition (3.118) and get a new one:

$$\|x - x^*\|_2 \leq \frac{\kappa}{3\kappa + 2L}\|y^{k-1} - x^*\|_2, \tag{3.119}$$

where x^* is an exact solution of the auxiliary minimization problem. How does it connected with (3.118)?

Proposition 3.3 *Condition* (3.119) *is sufficient for condition* (3.118).

Proof Using the triangle inequality,

$$\|y^{k-1} - x^*\|_2 - \|x - x^*\|_2 \leq \|x - y^{k-1}\|_2.$$

Therefore, it is sufficient to fulfill the following condition:

$$\|\nabla h_k(x)\|_2 \leq \frac{\kappa}{2}\left(\|y^{k-1} - x^*\|_2 - \|x - x^*\|_2\right) \quad (\Rightarrow (3.118))$$

On the other hand, due to the $(L + \kappa)$-strong convexity of h_k, it holds that

$$\|\nabla h_k(x)\|_2 \leq (L + \kappa)\|x - x^*\|_2.$$

Together with the previous statement, it implies that

$$\begin{aligned}\frac{\kappa}{2}\left(\|y^{k-1} - x^*\|_2 - \|x - x^*\|_2\right) &\leq (L + \kappa)\|x - x^*\|_2 \\ \Leftrightarrow \|x - x^*\|_2 &\leq \frac{\kappa}{3\kappa + 2L}\|y^{k-1} - x^*\|_2\end{aligned}$$

is sufficient for (3.118). □

The next theorem shows that the condition (3.119) is exactly the desired stopping condition with which the convergence rate of Catalyst is free of redundant logarithmic term.

Theorem 3.18 *Let us use algorithm $\mathcal{M}$ to solve the auxiliary minimization problem in* Catalyst. *Assume that $\mathcal{M}$ produces a sequence of points x^t such that*

$$h_k(x^t) - h_k(x^*) \leq C \cdot L \exp\left(-\frac{\kappa}{\kappa + L}t\right), \tag{3.120}$$

for some constant C. Then, the total complexity of Catalyst *with $\mathcal{M}$ and stopping condition* (3.119) *is*

$$N = O\left(\sqrt{\frac{LR^2}{\varepsilon}}\right). \tag{3.121}$$

Proof Due to the κ-strong convexity of h_k,

$$\frac{\kappa}{2}\|x^t - x^*\|_2^2 \leq h_k(x^t) - h_k(x^*),$$

then using (3.120) we get

$$\|x^t - x^*\|_2 \leq C \cdot \sqrt{\frac{2L}{\kappa}} \exp\left(-\frac{\kappa}{2(\kappa + L)}t\right).$$

Considering the condition (3.119), number of sufficient iterations T for $\mathcal{M}$ is determined from

$$C \cdot \sqrt{\frac{2L}{\kappa}} \exp\left(-\frac{\kappa}{2(\kappa + L)}T\right) = \frac{\kappa}{3\kappa + 2L}\|x^0 - x^*\|_2,, \tag{3.122}$$

where x^0 is set to y^{k-1}, and hence

$$T = O\left(\frac{\kappa + L}{\kappa} \log\left(\frac{3\kappa + 2L}{\kappa}\sqrt{\frac{2L}{\kappa}}\right)\right).$$

Since we choose $\kappa = L$, number of sufficient iterations is $T = O(1)$. Using the reasoning from Theorem's ? proof, we obtain (3.121). TODO: theorem from Alexander's part □

We see that convergence of Catalyst with stopping condition (3.119), as well as complexity of auxiliary minimization problem, is devoided of the logarithmic factor. Moreover, complexity of auxiliary minimization problem is independent on desired accuracy ε and equal to some small constant. It is very practical property of this Catalyst modification, because it allows to choose the number of iterations for internal method manually, without using any stop condition nor additional computations.

Eduard: I think it is better to move this and the next paragraphs to end of this subsection. Taking into account the subsubsection'ed Catalyst CDM example, it is more readable to leave the following content here

The simplest Catalyst algorithm described there is only applicable to the first order optimization methods and for smooth convex minimization problems, but the scope of Catalyst-like universal acceleration frameworks is not limited to this. The concept of Catalyst is evolved in Universal Meta-algorithm framework which, in turn, is applicable to tensor optimization methods (of a second- or more order), as well as for saddle point and composite optimization problems. Besides, this method provides the means to implement such an advanced technique as oracle complexity separation, so called gradient sliding. We see that the simplicity of Catalyst basement allows one to successfully generalize it on many different settings.

Another universal acceleration framework, that is more general than Catalyst, is Strongly convex accelerated hybrid proximal extragradient method. It is notable, that inexactness of proximal operator in it is taken into account with a framework of duality gap, that allows to mix the additive and multiplicative error terms. But the main advantage of this framework is that it exploits the strong convexity of the problem. The complexity of L-smooth μ-strongly convex problems is

$$O\left(\sqrt{\frac{\mu}{L}}\log\frac{1}{\varepsilon}\right), \tag{3.123}$$

and if the problem is composite and one of its terms is μ-strongly convex, this framework guarantee the convergence rate similar to corresponding to (3.123) for the outer algorithm (certainly not such a simple as (3.123) and with a lot of details we omit in our exposition).

3.2.2.1 Catalyst accelerated Coordinate Descent

Catalyst algorithm is very useful when the individual accelerated version of some method is computationally inefficient. In some cases, Catalyst is only an option to accelerate method preserving its computational advantages.

Let's consider CDM algorithm as the internal for Catalyst. The only difference with the standard CDM will be that components to step along are chosen not uniformly, but in proportional to coordinate-wise Lipschitz smoothness constants of h_k (equal to $\kappa + L_i$). For the sequence of points x^t generated by CDM we have:

$$\mathbb{E}[h_k(x^t)] - h_k(x^*) \leq (h_k(x^0) - h_k(x^*)) \cdot \exp\left(-\frac{\kappa}{\sum_{i=1}^n(\kappa + L_i)}t\right), \tag{3.124}$$

that is similar to condition in Theorem 3.18. But to estimate complexity of Catalyst with CDM in expectation, we should modify the reasoning of Theorem 3.18. Namely, we extend Theorem 3.18 with the following

Lemma 3.12 *Let $\tilde{\varepsilon}$ denote the r.h.s. of* (3.122) *and $T(\tilde{\varepsilon})$ be expressed from* (3.122) *as a function of $\tilde{\varepsilon}$. Assume that internal method stops immediately after fulfillment of* (3.119)*. Then, the expected total complexity of* Catalyst *with $\mathcal{M}$ and stopping condition* (3.119) *is*

$$\mathbb{E}[N] = K \cdot (T(\tilde{\varepsilon}) + 1) = O\left(\sqrt{\frac{LR^2}{\varepsilon}}\right)$$

Proof Let T be the number of iterations to fulfill (3.119) by $\mathcal{M}$ in expectation. Further, we prove that the expected number $\mathbb{E}[T]$ of iterations of $\mathcal{M}$ is bounded by $T(\varepsilon) + 1$, that implies the statement of lemma. So,

$$\begin{aligned}
\mathbb{P}(T \geq t) &\leq \mathbb{P}\left(\|x^t - x^*\|_2 \geq \tilde{\varepsilon}\right) \\
&\leq \min\left\{1,\ \mathbb{E}[\|x^t - x^*\|_2^2] \ / \ \tilde{\varepsilon}^2\right\} \quad \text{(Markov inequality)} \\
&\leq \min\left\{1,\ C_1 \cdot \exp\left(-\frac{\kappa}{\sum_{i=1}^n (\kappa + L_i)} t\right)\right\},
\end{aligned}$$

where the latter follows from (3.124) and strong-convexity of h_k, and C_1 depends only on L and κ. Finally,

$$\begin{aligned}
\mathbb{E}[T] = \sum_{t=1}^{\infty} \mathbb{P}(T \geq t) &\leq \int_0^{T(\tilde{\varepsilon})} dt + C_1 \cdot \exp\left(\frac{\kappa}{\sum_{i=1}^n (\kappa + L_i)}\right) \cdot \int_{T(\tilde{\varepsilon})}^{\infty} \exp(-t) \\
&= T(\tilde{\varepsilon}) + C_1 \cdot \exp\left(-\frac{\kappa}{\sum_{i=1}^n (\kappa + L_i)} T(\tilde{\varepsilon})\right) \leq T(\tilde{\varepsilon}) + 1,
\end{aligned}$$

where the latter follows from (3.122) (we exactly set $T(\tilde{\varepsilon})$ such that it holds). □

It is also important that since (3.124) we used coordinate Lipschitz smoothness ($\sum_{i=1}^n (\kappa + L_i)$ instead of $\kappa + L$), so the final complexity will be the $\mathbb{E}[N] = O\left(n\sqrt{\frac{\overline{L}R^2}{\varepsilon}}\right)$, where $\overline{L} = \frac{1}{n}\sum_{i=1}^n L_i$ (the proof is quite similar to that for Theorem 3.18).

Now, let us estimate the algorithmic complexity of Catalyst with CDM (i.e., in terms of number of arithmetic operations needed). Assume that one step of CDM takes $O(s)$ a.o. (it is convenient for the sparse optimization problems). So, very step of CDM within Catalyst also takes $O(s)$. But it is also necessary to perform one gradient step once an outer iteration of Catalyst, and it takes $O(ns)$ a.o. On the other hand, we must perform $T \sim n$ steps of internal method between the gradient steps, so the amortized complexity remains $O(s)$ a.o. per CDM step. The resulting algorithmic complexity of Catalyst with CDM is $O\left(sn \cdot \sqrt{\frac{\overline{L}R^2}{\varepsilon}}\right)$.

There are another ways to accelerate CDM: in particular, by modifying the method itself — its oracle complexity is $O\left(n\sqrt{\frac{\tilde{L}R^2}{\varepsilon}}\right)$, where $\tilde{L} = \left(\frac{1}{n}\sum_{i=1}^n \sqrt{L_i}\right)^2$. Nevertheless, such an acceleration leads to worsening of algorithmic complexity of CDM

iteration from $O(s)$ a.o. to $O(n)$ a.o., so the resulting algorithmic complexity is $O\left(n^2 \cdot \sqrt{\frac{\tilde{L}R^2}{\varepsilon}}\right)$. Thus, the direct acceleration makes us to forget about additional efficiency of CDM iteration (possibly, $s \ll n$), whilst the Catalyst with CDM gains very convenient and expectable accelerated and efficient complexity.

3.3 Obtaining estimates of the rate of convergence on average based on inexact gradients and batching

Estimates of the rate of convergence of optimal methods for smooth (strongly) convex stochastic optimization can be obtained from the results on the convergence of optimal methods under conditions of inexact gradients with small non-random noise.

Consider the stochastic optimization problem

$$\min_{x \in Q} \mathbb{E} f(x, \xi), \tag{3.125}$$

where $Q \subset \mathbb{R}^n$ is a closed and convex set, ξ is a random variable, expectation $\mathbb{E} f(x, \xi)$ is defined and finite for any $x \in Q$, f is μ-strongly convex ($\mu \geq 0$) and has L-Lipschitz gradient, i.e.

$$f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|_2^2 \leqslant f(y) \leqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2.$$

Suppose that we have access to the stochastic gradient $\nabla f(x, \xi)$ that satisfies

$$\mathbb{E}[\nabla f(x, \xi)] \equiv \nabla f(x),$$

for all $x \in Q$. The goal of this section is to show that the convergence rate

$$\tilde{O}\left(\min\left\{\frac{LR^2}{N^p} + \delta_1 + N^{p-1}\delta_2, LR^2 \exp\left(-\left(\frac{\mu}{L}\right)^{\frac{1}{p}} \frac{N}{2}\right) + \delta_1 + \left(\frac{L}{\mu}\right)^{\frac{p-1}{2}} \delta_2\right\}\right),$$

can be obtained based on results of convergence in conditions of inexact gradients with low noise of non-random nature, where $R = \|x_* - x^0\|$, x_* is the solution of 3.125, $p = 1$ corresponds to stochastic gradient descent, and $p = 2$ to accelerated stochastic descent.

We introduce the concept of inexact model. Let us define the function $\psi_\delta(y, x)$ as the (δ, L) model of the objective $f(x)$ if for all $x, y \in Q$ the function ψ_δ is convex w.r.t y, $\psi_\delta(y, y) = 0$, and

$$f(x) + \psi_\delta(y, x) + \frac{\mu}{2}\|y - x\|_2^2 - \delta_1 \leqslant f(y) \leqslant f(x) + \psi_\delta(y, x) + \frac{L}{2}\|y - x\|_2^2 + \delta_2. \tag{3.126}$$

Should we put the graphic explaining the concept of the model and example (composite function)?

3.3.0.1 Gradient Method for the Inexact Model of Objective

Algorithm 1 Gradient Method

1: **Input:** Starting point x_0, strong convexity constant $\mu \geq 0$, Lipschitz gradient constant $L > 0$.
2: **for** $k \geq 0$ **do**
3:
$$\phi_{k+1}(x) := \psi_{\delta^k}(x, x^k) + \frac{L}{2}\|x - x^k\|_2^2,$$
$$x^{k+1} := \arg\min_{x \in Q} \phi_{k+1}(x). \tag{3.127}$$
4: **end for**
5: **Output:** $y^N = \frac{1}{\sum_{i=1}^N q^{N-i}} \sum_{i=1}^N q^{N-i} x^i$

Lemma 3.13 *Let $\psi(x)$ be a convex function and*

$$y = \arg\min_{x \in Q} \left\{ \psi(x) + \frac{\beta}{2}\|x - z\|_2^2 + \frac{\gamma}{2}\|x - u\|_2^2 \right\},$$

where $\beta \geq 0$ and $\gamma \geq 0$. Then,

$$\begin{aligned} &\psi(x) + \tfrac{\beta}{2}\|x - z\|_2^2 + \tfrac{\gamma}{2}\|x - u\|_2^2 \\ &\quad \geq \psi(y) + \tfrac{\beta}{2}\|y - z\|_2^2 + \tfrac{\gamma}{2}\|y - u\|_2^2 + \tfrac{\beta+\gamma}{2}\|x - y\|_2^2, \quad \forall x \in Q. \end{aligned}$$

Proof Using the optimality criteria

$$\exists g \in \partial\psi(y), \quad \langle g + \tfrac{\beta}{2}\nabla_y \|y - z\|_2^2 + \tfrac{\gamma}{2}\nabla_y \|y - u\|_2^2, x - y\rangle \geq 0, \quad \forall x \in Q.$$

From $\beta + \gamma$–strong convexity $\psi(x) + \frac{\beta}{2}\|x - z\|_2^2 + \frac{\gamma}{2}\|x - u\|_2^2$

$$\begin{aligned} \psi(x) + \tfrac{\beta}{2}\|x - z\|_2^2 + \tfrac{\gamma}{2}\|x - u\|_2^2 \geq \psi(y) + \tfrac{\beta}{2}\|y - z\|_2^2 + \tfrac{\gamma}{2}\|y - u\|_2^2 \\ + \langle g + \tfrac{\beta}{2}\nabla_y \|y - u\|_2^2 + \tfrac{\gamma}{2}\nabla_y \|y - z\|_2^2, x - y\rangle + \tfrac{\beta+\gamma}{2}\|x - y\|_2^2 \end{aligned}$$

The last two inequalities prove the lemma. □

Let us denote $q = 1 - \frac{\mu}{L}$.

Theorem 3.19 *After N steps of Algorithm 1 the following inequality holds:*

$$f(y^N) - f(x_*) \leq \min\left\{ \frac{LR^2}{2N}, \frac{LR^2}{2} \exp\left(-\frac{\mu}{L}N\right) \right\} + \frac{1}{\sum_{i=1}^N q^{N-i}} \sum_{i=1}^N q^{N-i}(\delta_1^{i-1} + \delta_2^{i-1}).$$

Proof From (3.126) we obtain

$$f(x^N) \leq f(x^{N-1}) + \psi_{\delta^{N-1}}(x^N, x^{N-1}) + \frac{L}{2}\left\|x^N - x^{N-1}\right\|_2^2 + \delta_2^{N-1}.$$

Using Lemma 3.13 for (3.127):

$$f(x^N) \leq f(x^{N-1}) + \psi_{\delta^{N-1}}(x, x^{N-1}) + \frac{L}{2}\left\|x - x^{N-1}\right\|_2^2 - \frac{L}{2}\left\|x - x^N\right\|_2^2 + \delta_2^{N-1}.$$

Using the left inequality in (3.126):

$$f(x^N) \leq f(x) + \frac{L-\mu}{2}\left\|x - x^{N-1}\right\|_2^2 - \frac{L}{2}\left\|x - x^N\right\|_2^2 + \delta_1^{N-1} + \delta_2^{N-1}. \quad (3.128)$$

Setting $x = x_*$:

$$\frac{1}{2}\left\|x_* - x^N\right\|_2^2 \leq \frac{1}{L}\left(f(x_*) - f(x^N) + \delta_1^{N-1} + \delta_2^{N-1}\right) + \frac{q}{2}\left\|x_* - x^{N-1}\right\|_2^2.$$

Recursively we get that

$$\frac{1}{2}\left\|x_* - x^N\right\|_2^2 \leq \sum_{i=1}^{N}\left(\frac{q^{N-i}}{L}(f(x_*) - f(x^i) + \delta_1^{i-1} + \delta_2^{i-1})\right) + \frac{q^N}{2}\left\|x_* - x^0\right\|_2^2.$$

Considering that $\frac{1}{2}\left\|x_* - x^N\right\|_2^2 \geq 0$ and the definition of y^N, we get:

$$\begin{aligned}\frac{q^N}{2}\left\|x_* - x^0\right\|_2^2 &\geq \sum_{i=1}^{N}\left(\frac{q^{N-i}}{L}(f(x^i) - f(x_*) - \delta_1^{i-1} - \delta_2^{i-1})\right) \\ &\geq (f(y^N) - f(x_*))\sum_{i=1}^{N}\frac{q^{N-i}}{L} - \frac{1}{L}\sum_{i=1}^{N} q^{N-i}(\delta_1^{i-1} + \delta_2^{i-1}).\end{aligned}$$

Dividing the both sides of the last inequality by $\sum_{i=1}^{N}\frac{q^{N-i}}{L}$:

$$f(y^N) - f(x_*) \leq \frac{\frac{q^N}{2}}{\sum_{i=1}^{N}\frac{q^{N-i}}{L}}\left\|x_* - x^0\right\|_2^2 + \frac{1}{\sum_{i=1}^{N} q^{N-i}}\sum_{i=1}^{N} q^{N-i}(\delta_1^{i-1} + \delta_2^{i-1}).$$

Using that $\sum_{i=1}^{N}\frac{q^{N-i}}{L} \geq \frac{1}{L}$ and $q^{N-i} \geq q^N$ for all $i \geq 0$, we have the following inequality:

$$f(y^N) - f(x_*) \leq \frac{L}{2}\min\left\{q^N, \frac{1}{N}\right\}\left\|x_* - x^0\right\|_2^2 + \frac{1}{\sum_{i=1}^{N} q^{N-i}}\sum_{i=1}^{N} q^{N-i}(\delta_1^{i-1} + \delta_2^{i-1}).$$

That inequality and $q^N \leq \exp(-\frac{\mu}{L}N)$ complete the proof of the theorem. □

Assumption 3.6 Let sequences $\delta_1^k(y,x)$ and δ_2^k $(k \geq 0)$ be given. Assume that exists constant $\tilde{\delta_1}$ such that

$$\mathbb{E}\left[\delta_1^k(y,x)|\delta_{1,2}^{k-1}, \delta_{1,2}^{k-2}, ...\right] \leq \tilde{\delta_1},$$

and

$$\mathbb{E}\left[\delta_2^k|\delta_{1,2}^{k-1}, \delta_{1,2}^{k-2}, ...\right] \leq \hat{\delta}_2.$$

Theorem 3.20 *Let δ_1^k and δ_2^k $(k \geq 0)$ satisfy Assumption 3.6. Then after N steps of Algorithm 1 the following inequality holds:*

$$\mathbb{E}[f(y_N)] - f(x_*) \leq \min\left\{\frac{LR^2}{2N}, \frac{LR^2}{2}\exp\left(-\frac{\mu}{L}N\right)\right\} + \tilde{\delta}_1 + O(\hat{\delta}_2). \tag{3.129}$$

Proof Taking a expectation from inequality in 3.19

$$\mathbb{E}f(y_N) - f(x_*) \leq \min\left\{\frac{LR^2}{2N}, \frac{LR^2}{2}\exp\left(-\frac{\mu}{L}N\right)\right\} + \frac{1}{\sum_{i=1}^N q^{N-i}}\sum_{i=1}^N q^{N-i}\mathbb{E}(\delta_1^{i-1} + \delta_2^{i-1}).$$

By Assumption 3.6 we obtain

$$\mathbb{E}f(y_N) - f(x_*) \leq \min\left\{\frac{LR^2}{2N}, \frac{LR^2}{2}\exp\left(-\frac{\mu}{L}N\right)\right\} + \tilde{\delta}_1 + O(\hat{\delta}_2), \tag{3.130}$$

since the second moment of sub-Gaussian sequence $\sqrt{\delta_2^{i-1}}$ has the bound $\mathbb{E}\delta_2^{i-1} \leq 4\hat{\delta_2}$. □

3.3.0.2 Fast Gradient Method for the Inexact Model of Objective

For the case of the Fast Gradient Method we have to change the definition of the inexact model of the objective. Let us denote the function $\psi_\delta(y,x)$ as the (δ, L) model of the objective $f(x)$ if for all $x, y \in Q$ the function ψ_δ is convex w.r.t y, $\psi_\delta(y,y) = 0$, and

$$f(x) + \psi_\delta(y,x) + \frac{\mu}{2}\|y-x\|_2^2 - \delta_1(y,x) \leq f(y) \leq f(x) + \psi_\delta(y,x) + \frac{L}{2}\|y-x\|_2^2 + \delta_2. \tag{3.131}$$

Note, that in this definition we assume that $\delta_1(y,x)$ is a function of two arguments $y, x \in Q$.

The Fast Gradient Method is listed as Algorithm 2

Lemma 3.14 *For all $x \in Q$ the following inequality holds*

Algorithm 2 Fast Gradient Method

1: **Input:** Starting point x_0, strong convexity constant $\mu \geq 0$, Lipschitz gradient constant $L > 0$.
2: Set $y_0 := x_0$, $u_0 := x_0$, $\alpha_0 := 0$, $A_0 := \alpha_0$
3: **for** $k \geq 0$ **do**
4: Constant α_{k+1} is the largest solution of equation

$$A_{k+1}(1 + A_k\mu) = L\alpha_{k+1}^2, \quad A_{k+1} := A_k + \alpha_{k+1}. \tag{3.132}$$

5:

$$y^{k+1} := \frac{\alpha_{k+1}u^k + A_k x^k}{A_{k+1}}. \tag{3.133}$$

6:

$$\phi^{k+1}(x) = \alpha_{k+1}\psi_{\delta^k}(x, y^{k+1}) + \frac{1 + A_k\mu}{2}\|x - u^k\|_2^2 + \frac{\alpha_{k+1}\mu}{2}\|x - y^{k+1}\|_2^2,$$
$$u^{k+1} := \arg\min_{x \in Q} \phi^{k+1}(x). \tag{3.134}$$

7:

$$x^{k+1} := \frac{\alpha_{k+1}u^{k+1} + A_k x^k}{A_{k+1}}. \tag{3.135}$$

8: **end for**
9: **Output:** x_N,

$$A_{k+1}f(x^{k+1}) - A_k f(x^k) + \frac{1 + A_{k+1}\mu}{2}\left\|x - u^{k+1}\right\|_2^2 - \frac{1 + A_k\mu}{2}\left\|x - u^k\right\|_2^2$$
$$\leq \alpha_{k+1}f(x) + A_k\delta_1^k(x^k, y^{k+1}) + \alpha_{k+1}\delta_1^k(x, y^{k+1}) + A_{k+1}\delta_2^k.$$

Proof By (3.131)

$$f(x^{k+1}) \leq f(y^{k+1}) + \psi_{\delta^k}(x^{k+1}, y^{k+1}) + \frac{L}{2}\left\|x^{k+1} - y^{k+1}\right\|_2^2 + \delta_2^k.$$

From (3.135) and (3.133) for sequences x^{k+1} and y^{k+1} we have, that

$$f(x^{k+1}) \leq f(y^{k+1}) + \psi_{\delta^k}\left(\frac{\alpha_{k+1}u^{k+1} + A_k x^k}{A_{k+1}}, y^{k+1}\right)$$
$$+ \frac{L}{2}\left\|\frac{\alpha_{k+1}u^{k+1} + A_k x^k}{A_{k+1}} - y^{k+1}\right\|_2^2 + \delta_1^k$$
$$= f(y^{k+1}) + \psi_{\delta^k}\left(\frac{\alpha_{k+1}u^{k+1} + A_k x^k}{A_{k+1}}, y^{k+1}\right) + \frac{L\alpha_{k+1}^2}{2A_{k+1}^2}\left\|u^{k+1} - u^k\right\|_2^2 + \delta_2^k.$$

Since the model $\psi_{\delta^k}(\cdot, y^{k+1})$ is convex, we obtain

$$f(x^{k+1}) \leq \frac{A_k}{A_{k+1}}\left(f(y^{k+1}) + \psi_{\delta^k}(x^k, y^{k+1})\right) + \frac{\alpha_{k+1}}{A_{k+1}}\left(f(y^{k+1}) + \psi_{\delta^k}(u^{k+1}, y^{k+1})\right) + \frac{L\alpha_{k+1}^2}{2A_{k+1}^2}\left\|u^{k+1} - u^k\right\|_2^2 + \delta_2^k.$$

From (3.132) for sequence α_{k+1} we have

$$f(x^{k+1}) \leq \frac{A_k}{A_{k+1}}\left(f(y^{k+1}) + \psi_{\delta^k}(x^k, y^{k+1})\right) + \frac{\alpha_{k+1}}{A_{k+1}}\Big(f(y^{k+1}) + \psi_{\delta^k}(u^{k+1}, y^{k+1}) + \frac{1 + A_k\mu}{2\alpha_{k+1}}\left\|u^{k+1} - u^k\right\|_2^2\Big) + \delta_2^k. \tag{3.136}$$

Using Lemma 3.13 for optimization problem (3.134) we obtain

$$\begin{aligned}&\alpha_{k+1}\psi_{\delta^k}(u^{k+1}, y^{k+1}) + \frac{1 + A_k\mu}{2}\left\|u^{k+1} - u^k\right\|_2^2 + \frac{\alpha_{k+1}\mu}{2}\left\|u^{k+1} - y^{k+1}\right\|_2^2 \\ &\quad + \frac{1 + A_{k+1}\mu}{2}\left\|x - u^{k+1}\right\|_2^2 \\ &\leq \alpha_{k+1}\psi_{\delta^k}(x, y^{k+1}) + \frac{1 + A_k\mu}{2}\left\|x - u^k\right\|_2^2 + \frac{\alpha_{k+1}\mu}{2}\left\|x - y^{k+1}\right\|_2^2.\end{aligned}$$

Since $\frac{1}{2}\left\|u^{k+1} - y^{k+1}\right\|_2^2 \geq 0$

$$\begin{aligned}&\alpha_{k+1}\psi_{\delta^k}(u^{k+1}, y^{k+1}) + \frac{1 + A_k\mu}{2}\left\|u^{k+1} - u^k\right\|_2^2 \\ &\leq \alpha_{k+1}\psi_{\delta^k}(x, y^{k+1}) + \frac{1 + A_k\mu}{2}\left\|x - u^k\right\|_2^2 \\ &\quad - \frac{1 + A_{k+1}\mu}{2}\left\|x - u^{k+1}\right\|_2^2 + \frac{\alpha_{k+1}\mu}{2}\left\|x - y^{k+1}\right\|_2^2.\end{aligned} \tag{3.137}$$

By combining inequalities (3.136) and (3.137)

$$\begin{aligned}f(x^{k+1}) \leq\ & \frac{A_k}{A_{k+1}}\left(f(y^{k+1}) + \psi_{\delta^k}(x^k, y^{k+1})\right) \\ &+ \frac{\alpha_{k+1}}{A_{k+1}}\Big(f(y^{k+1}) + \psi_{\delta^k}(x, y^{k+1}) + \frac{\mu}{2}\left\|x - y^{k+1}\right\|_2^2 \\ &+ \frac{1 + A_k\mu}{2\alpha_{k+1}}\left\|x - u^k\right\|_2^2 - \frac{1 + A_{k+1}\mu}{2\alpha_{k+1}}\left\|x - u^{k+1}\right\|_2^2\Big) + \delta_2^k.\end{aligned}$$

By the left inequality (3.131):

$$f(x^{k+1}) \leq \frac{A_k}{A_{k+1}} f(x^k) + \frac{\alpha_{k+1}}{A_{k+1}} f(x) + \frac{1 + A_k\mu}{2A^{k+1}} \left\|x - u^k\right\|_2^2 - \frac{1 + A_{k+1}\mu}{2A^{k+1}} \left\|x - u^{k+1}\right\|_2^2 + \frac{A_k}{A_{k+1}} \delta_1^k(x^k, y^{k+1}) + \frac{\alpha_{k+1}}{A_{k+1}} \delta_1^k(x, y^{k+1}) + \delta_2^k.$$

The last inequality proves Lemma 3.14. □

Theorem 3.21 *After N steps of the Algorithm 2 the following inequality holds:*

$$f(x^N) - f(x_*) \leq \frac{R^2}{2A_N} + \frac{1}{A_N} \sum_{k=0}^{N-1} A_k \delta_1^k(x^k, y^{k+1}) + \frac{1}{A_N} \sum_{k=0}^{N-1} \alpha_{k+1} \delta_1^k(x_*, y^{k+1}) + \frac{1}{A_N} \sum_{k=0}^{N-1} A_{k+1} \delta_2^k.$$

Proof By summing inequalities from Lemma 3.14 for k from 0 to $N-1$ and, by setting $x = x_*$, we have

$$A_N f(x^N) \leq A_N f(x_*) + \frac{1}{2} \|x_* - u_0\|_2^2 - \frac{1 + A_N\mu}{2} \|x_* - u_N\|_2^2 + \sum_{k=0}^{N-1} A_k \delta_1^k(x^k, y^{k+1}) + \sum_{k=0}^{N-1} \alpha_{k+1} \delta_1^k(x_*, y^{k+1}) + \sum_{k=0}^{N-1} A_{k+1} \delta_2^k.$$

Since $\frac{1+A_N\mu}{2} \|x_* - u_N\|_2^2 \geq 0$

$$A_N f(x^N) - A_N f(x_*) \leq \frac{1}{2} \|x_* - u_0\|_2^2 + \sum_{k=0}^{N-1} A_k \delta_1^k(x^k, y^{k+1}) + \sum_{k=0}^{N-1} \alpha_{k+1} \delta_1^k(x_*, y^{k+1}) + \sum_{k=0}^{N-1} A_{k+1} \delta_2^k.$$

The last inequality proves the theorem. □

Lemma 3.15 *For all $N \geq 1$,*

$$\frac{1}{A_N} \leq \min\left\{\frac{4L}{N^2}, L\exp\left(-(N-1)\sqrt{\frac{\mu}{L}}\right)\right\}.$$

Proof At first, let us consider the case when $\mu = 0$. By (3.132)

$$A_{k+1} = L\alpha_{k+1}^2 = L(A_{k+1} - A_k)^2 = L\left(A_{k+1}^{1/2} - A_k^{1/2}\right)^2 \left(A_{k+1}^{1/2} + A_k^{1/2}\right)^2 \leq 4LA_{k+1}\left(A_{k+1}^{1/2} - A_k^{1/2}\right)^2.$$

Therefore $\frac{1}{2\sqrt{L}} \leq A_{k+1}^{1/2} - A_k^{1/2}$, which implies $\frac{1}{A_N} \leq \frac{4L}{N^2}$.

Now, assume that $\mu > 0$. By (3.132)

$$A_{k+1}A_k\mu \leq L\alpha_{k+1} = L(A_{k+1} - A_k)^2$$
$$= L\left(A_{k+1}^{1/2} - A_k^{1/2}\right)^2 \left(A_{k+1}^{1/2} + A_k^{1/2}\right)^2 \leq 4LA_{k+1}\left(A_{k+1}^{1/2} - A_k^{1/2}\right)^2.$$

Therefore $\mu A_k \leq 4L\left(A_{k+1}^{1/2} - A_k^{1/2}\right)^2$. Then $\left(1 + \frac{1}{2}\sqrt{\frac{\mu}{L}}\right) A_k^{1/2} \leq A_{k+1}^{1/2}$.

By the fact that $A_0 = \alpha_0 = 0$ and $A_1 = \alpha_1 = L\alpha_1^2$ we have

$$A_N \geq \left(1 + \frac{1}{2}\sqrt{\frac{\mu}{L}}\right)^2 A_{N-1} \geq \left(1 + \frac{1}{2}\sqrt{\frac{\mu}{L}}\right)^{2(N-1)} A_{N-1}$$
$$\geq \left(1 + \frac{1}{2}\sqrt{\frac{\mu}{L}}\right)^{2(N-1)} A_1 \geq \left(1 + \frac{1}{2}\sqrt{\frac{\mu}{L}}\right)^{2(N-1)} \frac{1}{L}.$$

Therefore, since $1 + x \leq e^x$, we have

$$\frac{1}{A_N} \leq L\left(1 + \frac{1}{2}\sqrt{\frac{\mu}{L}}\right)^{-2(N-1)} \leq L\exp\left(-(N-1)\sqrt{\frac{\mu}{L}}\right).$$

Lemma 3.16 *For all $N \geq 1$,*

$$\frac{1}{A_N}\sum_{k=0}^{N-1} A_{k+1} \leq O\left(\min\left\{N, \sqrt{\frac{L}{\mu}}\right\}\right). \tag{3.138}$$

Proof First, let $\mu = 0$. Note, that $\alpha_k \geq 0$ for $k = 0, \ldots, N$. Indeed, by (3.132), we have $L\alpha_{k+1}^2 = A_k + \alpha_{k+1}$. By taking the largest solution of this quadratic equation we have that

$$\alpha_{k+1} = \frac{\sqrt{\frac{1}{L^2} + \frac{4}{L}A_k} + \frac{1}{L}}{2} > 0, \quad k = 0, \ldots, N-1.$$

Therefore $A_k = A_{k+1} - \alpha_{k+1} \leq A_{k+1}$ for all $k \geq 0$. Finally,

$$\frac{1}{A_N}\sum_{k=0}^{N-1} A_{k+1} \leq N.$$

Now, let $\mu > 0$. By (3.132) $1 + \mu A_k = \frac{L(A_{k+1} - A_k)^2}{A_{k+1}}$. Then, A_k satisfies the following recurrence equation

$$LA_{k+1}^2 - 2LA_{k+1}A_k + LA_k^2 - A_{k+1} - \mu A_k A_{k+1} = 0.$$

Equivalently,

$$A_{k+1} = \frac{\sqrt{\left(\frac{1}{L} + 2A_k + \frac{\mu}{L}A_k\right)^2 - 4A_k^2} + \left(\frac{1}{L} + 2A^k + \frac{\mu}{L}A_k\right)}{2}.$$

Let us define the recurrence $\tilde{A}^{k+1} = \tilde{A}^k \left(\frac{\frac{\mu}{L}+\sqrt{\frac{\mu}{L}}\sqrt{\frac{\mu}{L}+4}}{2} + \frac{1}{L}\right)$. It is easy to verify that $\frac{A_{k+1}}{A_k} \geq \frac{\tilde{A}^{k+1}}{\tilde{A}^k}$ and therefore $\frac{A_k}{A_i} \geq \frac{\tilde{A}^k}{\tilde{A}^i}$ $\forall i < k$, $\forall k \geq 1$. Thus,

$$\frac{\sum\limits_{i=0}^{k} A_i}{A_k} \leq \frac{\sum\limits_{i=0}^{k} \tilde{A}^k}{\tilde{A}^i}.$$

By (3.132) $A_0 = 0$ and $\alpha_1 = A_1 = \frac{1}{L}$. Let $\tilde{A}_0 = 0$ and $\tilde{A}_1 = \frac{1}{L}$. Then for all $k \geq 1$ $\tilde{A}^k = \tilde{A}_1 C^{k-1}$, where $C = \left(\frac{\frac{\mu}{L}+\sqrt{\frac{\mu}{L}}\sqrt{\frac{\mu}{L}+4}}{2} + \frac{1}{L}\right)$. Therefore,

$$\begin{aligned}\frac{\sum\limits_{i=0}^{N} \tilde{A}^i}{\tilde{A}^k} &= \frac{\tilde{A}_1(C^k - 1)}{C - 1}\frac{1}{\tilde{A}_1 C^{k-1}} \leq \frac{C}{C-1} \leq \frac{C}{C - \frac{1}{L}} \\ &= \frac{\frac{\mu}{L} + \sqrt{\frac{\mu}{L}}\sqrt{\frac{\mu}{L}+4} + \frac{2}{L}}{\frac{\mu}{L} + \sqrt{\frac{\mu}{L}}\sqrt{\frac{\mu}{L}+4}} \leq 1 + \sqrt{\frac{L}{\mu}}.\end{aligned}$$

Assumption 3.7 Let sequence $\delta_1^k(y,x)$ and δ_2^k $(k \geq 0)$ be given. Let the random variable $\delta_1^k(x,y)$ has such a condition expectation that

- $\mathbb{E}\left[\delta_1^k(x,y) | \delta_1^{k-1}(x,y), \delta_2^{k-1}, \delta_1^{k-2}(x,y) \dots\right] \leq \tilde{\delta}_1^k(x-y)$ $\forall x, y \in Q$, where $\tilde{\delta}_1^k(\cdot)$ is a non-random function of one argument.
- $\tilde{\delta}_1^k(\alpha z) \leq \alpha \tilde{\delta}_1^k(z)$ for all $\alpha \geq 0$ and $z \in B(0,R)$.
- $\tilde{\delta}_1 < +\infty$, where $\tilde{\delta}_1 \geq \sup^{z \in B(0,R)} \tilde{\delta}_1^k(z)$,

and

$$\mathbb{E}\left[\delta_2^k | \delta_{1,2}^{k-1}, \delta_{1,2}^{k-2}, \dots\right] \leq \hat{\delta}_2.$$

Theorem 3.22 *Let sequences $\delta_1^k(x,y)$ and δ_2^k $(k \geq 0)$ satisfy 3.7 for all $x, y \in Q$. Then after N steps of Algorithm 2 the following inequality holds:*

$$\begin{aligned}\mathbb{E}f(x_N) - f(x_*) \leq \min\left\{\frac{4LR^2}{N^2}, 2LR^2\exp\left(-\frac{N-1}{2}\sqrt{\frac{\mu}{L}}\right)\right\} + \tilde{\delta}_1 \\ + O\left(\min\left\{N, \sqrt{\frac{L}{\mu}}\right\}\hat{\delta}_2\right).\end{aligned}$$

Proof By 3.16

$$\frac{1}{A_N}\sum_{k=0}^{N-1} A_{k+1} \leq O\left(\min\left\{N, \sqrt{\frac{L}{\mu}}\right\}\right).$$

By Assumption 3.7 we have

$$A_k\tilde{\delta}_1^k(x^k - y^{k+1}) = A_k\tilde{\delta}_1^k\left(\frac{\alpha_{k+1}}{A_k}(y^{k+1} - u^k)\right) \leq \alpha_{k+1}\tilde{\delta}_1^k(y^{k+1} - u^k) \leq \alpha_{k+1}\tilde{\delta}_1.$$

Finally,

$$\mathbb{E}f(x_N) - f(x_*) \leq \min\left\{\frac{4LR^2}{N^2}, 2LR^2\exp\left(-\frac{N-1}{2}\sqrt{\frac{\mu}{L}}\right)\right\} + 2\tilde{\delta}_1 + O\left(\min\left\{N, \sqrt{\frac{L}{\mu}}\right\}\hat{\delta}_2\right).$$

3.4 High-Probability Bounds for Stochastic Methods

In this part, we focus on the following problem

$$\min_{x\in\mathbb{R}^n} f(x), \quad f(x) = \mathbb{E}_\xi\left[f(x,\xi)\right], \tag{3.139}$$

where $f(x)$ is a μ-strongly convex or convex but possibly non-smooth function.

Definition 3.1 Differentiable function $f : Q \subseteq \mathbb{R}^n \to \mathbb{R}$ is called μ-strongly convex for some $\mu \geq 0$ if for all $x, y \in Q$

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|_2^2.$$

When $\mu = 0$ function f is called convex.

Next, at each point $x \in \mathbb{R}^n$ there is an access to the unbiased estimator $\nabla f(x,\xi)$ of $\nabla f(x)$ such that $\mathbb{E}_\xi\left[\|\nabla f(x,\xi) - \nabla f(x)\|_2^2\right] < \infty$ and if additionally $x \in Q \subseteq \mathbb{R}^n$, then

$$\mathbb{E}_\xi[\nabla f(x,\xi)] = \nabla f(x), \quad \mathbb{E}_\xi\left[\|\nabla f(x,\xi) - \nabla f(x)\|_2^2\right] \leq \sigma^2, \quad \sigma > 0. \tag{3.140}$$

Q is the ball centered at the solution x^* of (3.139) with radius $\sim R_0 \geq \|x^0 - x^*\|_2$, where x^0 is a starting point of the method. Function f has (ν, M_ν)-Hölder continuous gradients on a compact set $Q \subseteq \mathbb{R}^n$ for some $\nu \in [0,1]$, $M_\nu > 0$ meaning that

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq M_\nu\|x - y\|_2^\nu \quad \forall x, y \in Q. \tag{3.141}$$

When $\nu = 1$ inequality (3.141) implies M_1-smoothness of f, and when $\nu = 0$ it means $\nabla f(x)$ has bounded variation which is equivalent to being uniformly bounded. Moreover, when $\nu = 0$ differentiability of f is not needed: one can assume uniform boundedness of the subgradients of f throughout the proofs. When (3.141) holds for

$\nu = 0$ and $\nu = 1$ simultaneously then it holds for all $\nu \in [0, 1]$ with $M_\nu \leq M_0^{1-\nu} M_1^\nu$. It is sufficient to assume that the set Q is the ball centered at the solution x^* of (3.139) with radius $\sim R_0 \geq \|x^0 - x^*\|_2$, where x^0 is a starting point of the method, i.e., analysis in this part does not require (3.141) to hold on $\mathbb{R}^n$.

In this chapter "gradient" is used, though the entire analysis below works for non-differentiable convex functions as well (when $\nu = 0$): one just needs to replace gradients by subgradients. This remark is valid for Definition 3.1 as well.

High-Probability Bounds

All analysis in this part of the book are presented in terms of high-probability bounds. For a given accuracy $\varepsilon > 0$ and confidence level $\beta \in (0, 1)$ the goal is to find ε-solutions of problem (3.139) with probability at least $1 - \beta$, i.e., such $\widehat{x}$ that $\mathbb{P}\{f(\widehat{x}) - f(x^*) \leq \varepsilon\} \geq 1 - \beta$. For brevity, such (in general, random) points $\widehat{x}$ are called as (ε, β)-solution of (3.139). Moreover, by high-probability iteration/oracle complexity of a stochastic method $\mathcal{M}$ it means a sufficient number of iterations/oracle calls (number of $\nabla f(x, \xi)$ computations) needed to guarantee that $\mathcal{M}$ returns an (ε, β)-solution of (3.139).

Gradient clipping

In this chapter, we will need the concept of gradient clipping. The methods based on gradient clipping and normalization are popular in different machine learning and deep learning tasks due to their robustness in practice to the noise in the stochastic gradients and rapid changes of the objective function. We consider methods such as clipped-SGD and clipped-SSTM. These methods are based on the clipping of the stochastic gradients:

$$\text{clip}(\nabla f(x, \boldsymbol{\xi}), \lambda) = \min\left\{1, \frac{\lambda}{\|\nabla f(x, \boldsymbol{\xi})\|_2}\right\} \nabla f(x, \boldsymbol{\xi}) \tag{3.142}$$

where $\nabla f(x, \boldsymbol{\xi}) = \frac{1}{m} \sum_{i=1}^m \nabla f(x, \xi_i)$ is a mini-batched stochastic gradient. Gradient clipping ensures that the resulting vector has a norm bounded by the clipping level λ. Since the clipped stochastic gradient cannot have arbitrary large norm, the clipping helps to avoid unstable behavior of the method when the noise is heavy-tailed and the clipping level λ is properly adjusted.

3.4.1 Clipped Stochastic Gradient Descent

The first method we are considering is Clipped Stochastic Gradient Descent (clipped-SGD).

Algorithm 3 Clipped Stochastic Gradient Descent (clipped-SGD): case $\nu \in [0,1]$

Input: starting point x^0, number of iterations N, batchsize m, stepsize γ, clipping parameter $B > 0$.

1: **for** $k = 0, \ldots, N-1$ **do**
2: Draw mini-batch of m fresh i.i.d. samples $\xi_1^k, \ldots, \xi_m^k$ and compute $\nabla f(x^{k+1}, \xi^k) = \frac{1}{m}\sum_{i=1}^m \nabla f(x^{k+1}, \xi_i^k)$
3: Compute $\widetilde{\nabla} f(x^k, \xi^k) = \text{clip}(\nabla f(x^k, \xi^k), \lambda)$ using (3.142) with $\lambda = B/\gamma$
4: $x^{k+1} = x^k - \gamma \widetilde{\nabla} f(x^k, \xi^k)$
5: **end for**

Output: $\bar{x}^N = \frac{1}{N}\sum_{k=0}^{N-1} x^k$

Convex Case

Theorem 3.23 *Assume that function f is convex, achieves its minimum at a point x^*, and its stochastic gradient and its gradient satisfy* (3.140) *and* (3.141) *respectively with $\sigma > 0$, $\nu \in [0,1]$, $M_\nu > 0$ on $Q = B_{7R_0}(x^*)$, where $R_0 \geq \|x^0 - x^*\|_2$. Then, for all $\beta \in (0,1)$ and N such that*

$$\ln \frac{4N}{\beta} \geq 2, \tag{3.143}$$

we have that after N iterations of clipped-SGD *with*

$$\lambda = \frac{R_0}{\gamma \ln \frac{4N}{\beta}}, \quad m \geq \max\left\{1, \frac{81N\sigma^2}{\lambda^2 \ln \frac{4N}{\beta}}\right\} \tag{3.144}$$

and stepsize

$$\gamma \leq \min\left\{\frac{\varepsilon^{\frac{1-\nu}{1+\nu}}}{8M_\nu^{\frac{2}{1+\nu}}}, \frac{R_0}{\sqrt{2N}\varepsilon^{\frac{\nu}{1+\nu}} M_\nu^{\frac{1}{1+\nu}}}, \frac{R_0^{1-\nu}}{2C^\nu M_\nu \ln \frac{4N}{\beta}}\right\}, \tag{3.145}$$

with probability at least $1-\beta$ it holds that

$$f(\bar{x}^N) - f(x^*) \leq \frac{C^2 R_0^2}{\gamma N}, \tag{3.146}$$

where $\bar{x}^N = \frac{1}{N}\sum_{k=0}^{N-1} x^k$ and

$$C = 7. \tag{3.147}$$

In other words, clipped-SGD *with* $\gamma = \min\left\{\frac{\varepsilon^{\frac{1-\nu}{1+\nu}}}{8M_\nu^{\frac{2}{1+\nu}}}, \frac{R_0}{\sqrt{2N}\varepsilon^{\frac{\nu}{1+\nu}} M_\nu^{\frac{1}{1+\nu}}}, \frac{R_0^{1-\nu}}{2C^\nu M_\nu \ln \frac{4N}{\beta}}\right\}$ *achieves $f(\bar{x}^N) - f(x^*) \leq \varepsilon$ with probability at least $1-\beta$ after $O\left(\max\left\{\frac{M_\nu^{\frac{2}{1+\nu}} R_0^2}{\varepsilon^{\frac{2}{1+\nu}}}, \frac{M_\nu R_0^{1+\nu}}{\varepsilon} \ln \frac{M_\nu R_0^{1+\nu}}{\varepsilon\beta}\right\}\right)$ iterations and requires*

$$O\left(\max\left\{\frac{M_\nu^{\frac{2}{1+\nu}}R_0^2}{\varepsilon^{\frac{2}{1+\nu}}}, \max\left\{\frac{M_\nu R_0^{1+\nu}}{\varepsilon}, \frac{\sigma^2 R_0^2}{\varepsilon^2}\right\}\ln\frac{M_\nu R_0^{1+\nu}}{\varepsilon\beta}\right\}\right) \tag{3.148}$$

oracle calls.

Proof Since $f(x)$ is convex and its gradients satisfy (3.141), we get the following inequality under assumption that $x^k \in B_{7R_0}(x^*)$:

$$\begin{aligned}
\|x^{k+1}-x^*\|_2^2 &= \|x^k - \gamma\widetilde{\nabla}f(x^k,\boldsymbol{\xi}^k) - x^*\|_2^2 \\
&= \|x^k-x^*\|_2^2 + \gamma^2\|\widetilde{\nabla}f(x^k,\boldsymbol{\xi}^k)\|_2^2 - 2\gamma\left\langle x^k - x^*, \widetilde{\nabla}f(x^k,\boldsymbol{\xi}^k)\right\rangle \\
&= \|x^k-x^*\|_2^2 + \gamma^2\|\nabla f(x^k)+\theta_k\|_2^2 - 2\gamma\left\langle x^k-x^*, \nabla f(x^k)+\theta_k\right\rangle \\
&\overset{(3.165)}{\leq} \|x^k-x^*\|_2^2 + 2\gamma^2\|\nabla f(x^k)\|_2^2 + 2\gamma^2\|\theta_k\|_2^2 - 2\gamma\left\langle x^k-x^*, \nabla f(x^k)+\theta_k\right\rangle \\
&\overset{(3.174)}{\leq} \|x^k-x^*\|_2^2 - 2\gamma\left(1-2\gamma\left(\frac{1}{\varepsilon}\right)^{\frac{1-\nu}{1+\nu}}M_\nu^{\frac{2}{1+\nu}}\right)\left(f(x^k)-f(x^*)\right) + 2\gamma^2\|\theta_k\|_2^2 \\
&\qquad -2\gamma\left\langle x^k-x^*,\theta_k\right\rangle + 2\gamma^2\varepsilon^{\frac{2\nu}{1+\nu}}M_\nu^{\frac{2}{1+\nu}},
\end{aligned}$$

where $\theta_k = \widetilde{\nabla}f(x^k,\boldsymbol{\xi}^k) - \nabla f(x^k)$ and the last inequality follows from the convexity of f. Using notation $R_k \overset{\text{def}}{=} \|x^k - x^*\|_2$, $k > 0$ we derive that for all $k \geq 0$

$$\begin{aligned}
R_{k+1}^2 \leq R_k^2 &- 2\gamma\left(1-2\gamma\left(\frac{1}{\varepsilon}\right)^{\frac{1-\nu}{1+\nu}}M_\nu^{\frac{2}{1+\nu}}\right)\left(f(x^k)-f(x^*)\right) + 2\gamma^2\|\theta_k\|_2^2 \\
&-2\gamma\left\langle x^k-x^*,\theta_k\right\rangle + 2\gamma^2\varepsilon^{\frac{2\nu}{1+\nu}}M_\nu^{\frac{2}{1+\nu}}
\end{aligned}$$

under assumption that $x^k \in B_{7R_0}(x^*)$. Let us define $A = 2\gamma\left(1-2\gamma\left(\frac{1}{\varepsilon}\right)^{\frac{1-\nu}{1+\nu}}M_\nu^{\frac{2}{1+\nu}}\right) \overset{(3.145)}{\geq}$ $\gamma > 0$, then

$$A\left(f(x^k)-f(x^*)\right) \leq R_k^2 - R_{k+1}^2 + 2\gamma^2\|\theta_k\|_2^2 - 2\gamma\left\langle x^k-x^*,\theta_k\right\rangle + 2\gamma^2\varepsilon^{\frac{2\nu}{1+\nu}}M_\nu^{\frac{2}{1+\nu}}$$

under assumption that $x^k \in B_{7R_0}(x^*)$. Summing up these inequalities for $k = 0,\ldots,N-1$, we obtain

$$\frac{A}{N}\sum_{k=0}^{N-1}\left[f(x^k)-f(x^*)\right] \le \frac{1}{N}\sum_{k=0}^{N-1}\left(R_k^2-R_{k+1}^2\right)+2\gamma^2\varepsilon^{\frac{2\nu}{1+\nu}}M_\nu^{\frac{2}{1+\nu}}+\frac{2\gamma^2}{N}\sum_{k=0}^{N-1}\|\theta_k\|_2^2$$
$$-\frac{2\gamma}{N}\sum_{k=0}^{N-1}\left\langle x^k-x^*,\theta_k\right\rangle$$
$$=\frac{1}{N}\left(R_0^2-R_N^2\right)+2\gamma^2\varepsilon^{\frac{2\nu}{1+\nu}}M_\nu^{\frac{2}{1+\nu}}+\frac{2\gamma^2}{N}\sum_{k=0}^{N-1}\|\theta_k\|_2^2$$
$$-\frac{2\gamma}{N}\sum_{k=0}^{N-1}\left\langle x^k-x^*,\theta_k\right\rangle$$

under assumption that $x^k \in B_{7R_0}(x^*)$. Noticing that for $\bar{x}^N = \frac{1}{N}\sum\limits_{k=0}^{N-1} x^k$ Jensen's inequality gives $f(\bar{x}^N) = f\left(\frac{1}{N}\sum\limits_{k=0}^{N-1} x^k\right) \le \frac{1}{N}\sum\limits_{k=0}^{N-1} f(x^k)$, we have

$$AN\left(f(\bar{x}^N)-f(x^*)\right) \le R_0^2-R_N^2+2\gamma^2N\varepsilon^{\frac{2\nu}{1+\nu}}M_\nu^{\frac{2}{1+\nu}}+2\gamma^2\sum_{k=0}^{N-1}\|\theta_k\|_2^2$$
$$-2\gamma\sum_{k=0}^{N-1}\left\langle x^k-x^*,\theta_k\right\rangle \tag{3.149}$$

under assumption that $x^k \in B_{7R_0}(x^*)$ for $k = 0, 1, \ldots, N-1$. Taking into account that $f(\bar{x}^N) - f(x^*) \ge 0$ and changing the indices we get that for all $k = 0, 1, \ldots, N$

$$R_k^2 \le R_0^2+2\gamma^2k\varepsilon^{\frac{2\nu}{1+\nu}}M_\nu^{\frac{2}{1+\nu}}+2\gamma^2\sum_{l=0}^{k-1}\|\theta_l\|_2^2-2\gamma\sum_{l=0}^{k-1}\left\langle x^l-x^*,\theta_k\right\rangle. \tag{3.150}$$

under assumption that $x^l \in B_{7R_0}(x^*)$ for $l = 0, 1, \ldots, k-1$. The remaining part of the proof is based on the analysis of inequality (3.150). In particular, via induction we prove that for all $k = 0, 1, \ldots, N$ with probability at least $1 - \frac{k\beta}{N}$ the following statement holds: inequalities

$$R_t^2 \overset{(3.150)}{\le} R_0^2+2\gamma^2t\varepsilon^{\frac{2\nu}{1+\nu}}M_\nu^{\frac{2}{1+\nu}}+2\gamma^2\sum_{l=0}^{t-1}\|\theta_k\|_2^2-2\gamma\sum_{l=0}^{t-1}\left\langle x^k-x^*,\theta_k\right\rangle$$
$$\le C^2R_0^2 \tag{3.151}$$

hold for $t = 0, 1, \ldots, k$ simultaneously where C is defined in (3.147). Let us define the probability event when this statement holds as E_k. Then, our goal is to show that $\mathbb{P}\{E_k\} \ge 1 - \frac{k\beta}{N}$ for all $k = 0, 1, \ldots, N$. For $t = 0$ inequality (3.151) holds with probability 1 since $C \ge 1$. Next, assume that for some $k = T-1 \le N-1$ we have $\mathbb{P}\{E_k\} = \mathbb{P}\{E_{T-1}\} \ge 1 - \frac{(T-1)\beta}{N}$. Let us prove that $\mathbb{P}\{E_T\} \ge 1 - \frac{T\beta}{N}$. First of all,

probability event E_{T-1} implies that $x^t \in B_{7R_0}(x^*)$ for $t = 0, 1, \ldots, T-1$, and, as a consequence, (3.150) holds for $k = T$. Since $\nabla f(x)$ is (ν, M_ν)-Hölder continuous on $B_{7R_0}(x^*)$, we have that probability event E_{T-1} implies

$$\left\|\nabla f(x^t)\right\|_2 \overset{(??)}{\leq} M_\nu \|x^t - x^0\|^\nu \leq M_\nu C^\nu R_0^\nu \overset{(3.145)}{\leq} \frac{\lambda}{2} \tag{3.152}$$

for $t = 0, \ldots, T-1$. Next, we introduce new random variables:

$$\eta_l = \begin{cases} x^* - x^l, & \text{if } \|x^* - x^l\|_2 \leq CR_0, \\ 0, & \text{otherwise,} \end{cases} \tag{3.153}$$

for $l = 0, 1, \ldots T-1$. Note that these random variables are bounded with probability 1, i.e. with probability 1 we have

$$\|\eta_l\|_2 \leq CR_0. \tag{3.154}$$

Using the introduced notation, we obtain that E_{T-1} implies

$$R_T^2 \overset{(3.144),(3.145),(3.150),(3.151),(3.153)}{\leq} 2R_0^2 + 2\gamma \sum_{l=0}^{T-1} \langle \theta_l, \eta_l \rangle + 2\gamma^2 \sum_{l=0}^{T-1} \|\theta_l\|_2^2.$$

Finally, we do some preliminaries in order to apply Bernstein's inequality (see Lemma 3.19) and obtain that E_{T-1} implies

$$\begin{aligned} R_T^2 \overset{(3.165)}{\leq} 2R_0^2 &+ \underbrace{2\gamma \sum_{l=0}^{T-1} \left\langle \theta_l^u, \eta_l \right\rangle}_{①} + \underbrace{2\gamma \sum_{l=0}^{T-1} \left\langle \theta_l^b, \eta_l \right\rangle}_{②} + \underbrace{4\gamma^2 \sum_{l=0}^{T-1} \left(\|\theta_l^u\|_2^2 - \mathbb{E}_{\xi^l}\left[\|\theta_l^u\|_2^2\right] \right)}_{③} \\ &\underbrace{+ 4\gamma^2 \sum_{l=0}^{T-1} \mathbb{E}_{\xi^l}\left[\|\theta_l^u\|_2^2\right]}_{④} + \underbrace{4\gamma^2 \sum_{l=0}^{T-1} \|\theta_l^b\|_2^2}_{⑤}, \end{aligned} \tag{3.155}$$

where we introduce new notations:

$$\theta_l^u \overset{\text{def}}{=} \widetilde{\nabla} f(x^l, \xi^l) - \mathbb{E}_{\xi^l}\left[\widetilde{\nabla} f(x^l, \xi^l)\right], \quad \theta_l^b \overset{\text{def}}{=} \mathbb{E}_{\xi^l}\left[\widetilde{\nabla} f(x^l, \xi^l)\right] - \nabla f(x^l), \tag{3.156}$$

$$\theta_l = \theta_l^u + \theta_l^b.$$

It remains to provide tight upper bounds for ①, ②, ③, ④ and ⑤, i.e. in the remaining part of the proof we show that ① + ② + ③ + ④ + ⑤ $\leq \delta C^2 R_0^2$ for some $\delta < 1$.

Upper bound for ①. First of all, since $\mathbb{E}_{\xi^l}[\theta_l^u] = 0$ summands in ① are conditionally unbiased:

$$\mathbb{E}_{\xi^l}\left[2\gamma \left\langle \theta_l^u, \eta_l \right\rangle\right] = 0.$$

Secondly, these summands are bounded with probability 1:

$$\left|2\gamma\left\langle\theta_l^u,\eta_l\right\rangle\right| \le 2\gamma\|\theta_l^u\|_2\,\|\eta_l\|_2 \overset{(3.169),(3.154)}{\le} 4\gamma\lambda CR_0.$$

Finally, one can bound conditional variances $\sigma_l^2 \overset{\text{def}}{=} \mathbb{E}_{\xi^l}\left[4\gamma^2\left\langle\theta_l^u,\eta_l\right\rangle^2\right]$ in the following way:

$$\sigma_l^2 \le \mathbb{E}_{\xi^l}\left[4\gamma^2\left\|\theta_l^u\right\|_2^2\|\eta_l\|_2^2\right] \overset{(3.154)}{\le} 4\gamma^2(CR_0)^2\mathbb{E}_{\xi^l}\left[\left\|\theta_l^u\right\|_2^2\right],$$

i.e., σ_l^2 is finite due to finiteness of $\|\theta_{l+1}^u\|_2$ (see Lemma 3.18). In other words, sequence $\left\{2\gamma\left\langle\theta_l^u,\eta_l\right\rangle\right\}_{l\ge0}$ is a bounded martingale difference sequence with bounded conditional variances $\{\sigma_l^2\}_{l\ge0}$. Therefore, we can apply Bernstein's inequality, i.e., we apply Lemma 3.19 with $X_l = 2\gamma\left\langle\theta_l^u,\eta_l\right\rangle$, $c = 4\gamma\lambda CR_0$ and $F = \frac{c^2\ln\frac{4N}{\beta}}{6}$ and get that for all $b > 0$

$$\mathbb{P}\left\{\left|\sum_{l=0}^{T-1}X_l\right| > b \text{ and } \sum_{l=0}^{T-1}\sigma_l^2 \le F\right\} \le 2\exp\left(-\frac{b^2}{2F + {}^{2cb}/_3}\right)$$

or, equivalently, with probability at least $1 - 2\exp\left(-\frac{b^2}{2F+{}^{2cb}/_3}\right)$

$$\text{either } \sum_{l=0}^{T-1}\sigma_l^2 > F \quad \text{or} \quad \underbrace{\left|\sum_{l=0}^{T-1}X_l\right|}_{|①|} \le b.$$

The choice of F will be clarified further, let us now choose b in such a way that $2\exp\left(-\frac{b^2}{2F+{}^{2cb}/_3}\right) = \frac{\beta}{2N}$. This implies that b is the positive root of the quadratic equation

$$b^2 - \frac{2c\ln\frac{4N}{\beta}}{3}b - 2F\ln\frac{4N}{\beta} = 0,$$

hence

$$\begin{aligned} b &= \frac{c\ln\frac{4N}{\beta}}{3} + \sqrt{\frac{c^2\ln^2\frac{4N}{\beta}}{9} + 2F\ln\frac{4N}{\beta}} = \frac{c\ln\frac{4N}{\beta}}{3} + \sqrt{\frac{4c^2\ln^2\frac{4N}{\beta}}{9}} \\ &= c\ln\frac{4N}{\beta} = 4\gamma\lambda CR_0\ln\frac{4N}{\beta}. \end{aligned}$$

That is, with probability at least $1 - \frac{\beta}{2N}$

$$\underbrace{\text{either } \sum_{l=0}^{T-1}\sigma_l^2 > F \quad \text{or} \quad |①| \le 4\gamma\lambda CR_0\ln\frac{4N}{\beta}}_{\text{probability event } E_①}.$$

Here and below, we notice that the conditions of Lemma 3.18 hold when E_{T-1} holds, since event E_{T-1} implies that $x^0, x^1, \ldots, x^T$ lie in $B_{7R_0}(x^*)$. Therefore, probability event E_{T-1} implies that

$$\sum_{l=0}^{T-1} \sigma_l^2 \leq 4\gamma^2 (CR_0)^2 \sum_{l=0}^{T-1} \mathbb{E}_{\xi^l}\left[\|\theta_l^u\|_2^2\right] \overset{(3.172)}{\leq} 72\gamma^2 (CR_0)^2 \sigma^2 \frac{T}{m}$$
$$\overset{T \leq N}{\leq} 72\gamma^2 (CR_0)^2 \sigma^2 \frac{N}{m} \leq \frac{c^2 \ln \frac{4N}{\beta}}{6} = F,$$

where the last inequality follows from $c = 4\gamma\lambda CR_0$ and simple arithmetic.

Upper bound for ②. First of all, we notice that probability event E_{T-1} implies

$$2\gamma \left\langle \theta_l^b, \eta_l \right\rangle \leq 2\gamma \left\|\theta_l^b\right\|_2 \|\eta_l\|_2 \overset{(3.170),(3.154)}{\leq} 2\gamma \frac{4\sigma^2}{m\lambda} CR_0 = \frac{8\gamma\sigma^2 CR_0}{m\lambda}.$$

This implies that

$$② = 2\gamma \sum_{l=0}^{T-1} \left\langle \theta_l^b, \eta_l \right\rangle \overset{T \leq N}{\leq} \frac{8\gamma\sigma^2 CR_0 N}{m\lambda} \overset{(3.144)}{\leq} \frac{8}{81}\lambda\gamma CR_0 \ln \frac{4N}{\beta}.$$

Upper bound for ③. We derive the upper bound for ③ using the same technique as for ①. First of all, we notice that the summands in ③ are conditionally unbiased:

$$\mathbb{E}_{\xi^l}\left[4\gamma^2 \left(\|\theta_l^u\|_2^2 - \mathbb{E}_{\xi^l}\left[\|\theta_l^u\|_2^2\right]\right)\right] = 0.$$

Secondly, the summands are bounded with probability 1:

$$\left|4\gamma^2 \left(\|\theta_l^u\|_2^2 - \mathbb{E}_{\xi^l}\left[\|\theta_l^u\|_2^2\right]\right)\right| \leq 4\gamma^2 \left(\|\theta_l^u\|_2^2 + \mathbb{E}_{\xi^l}\left[\|\theta_l^u\|_2^2\right]\right) \overset{(3.169)}{\leq} 4\gamma^2 \left(4\lambda^2 + 4\lambda^2\right)$$
$$= 32\gamma^2\lambda^2 \overset{\text{def}}{=} c_1. \tag{3.157}$$

Finally, one can bound conditional variances $\hat{\sigma}_l^2 \overset{\text{def}}{=} \mathbb{E}_{\xi^l}\left[\left|4\gamma^2 \left(\|\theta_l^u\|_2^2 - \mathbb{E}_{\xi^l}\left[\|\theta_l^u\|_2^2\right]\right)\right|^2\right]$ in the following way:

$$\hat{\sigma}_l^2 \overset{(3.157)}{\leq} c_1 \mathbb{E}_{\xi^l}\left[\left|4\gamma^2 \left(\|\theta_l^u\|_2^2 - \mathbb{E}_{\xi^l}\left[\|\theta_l^u\|_2^2\right]\right)\right|\right]$$
$$\leq 4\gamma^2 c_1 \mathbb{E}_{\xi^l}\left[\|\theta_l^u\|_2^2 + \mathbb{E}_{\xi^l}\left[\|\theta_l^u\|_2^2\right]\right] = 8\gamma^2 c_1 \mathbb{E}_{\xi^l}\left[\|\theta_l^u\|_2^2\right], \tag{3.158}$$

i.e., $\hat{\sigma}_l^2$ is finite due to finiteness of $\|\theta_{l+1}^u\|_2$ (see Lemma 3.18) In other words, sequence $\left\{4\gamma^2 \left(\|\theta_l^u\|_2^2 - \mathbb{E}_{\xi^l}\left[\|\theta_l^u\|_2^2\right]\right)\right\}_{l \geq 0}$ is a bounded martingale difference sequence with bounded conditional variances $\{\hat{\sigma}_l^2\}_{l \geq 0}$. Therefore, we can apply Bernstein's inequality, i.e. we apply Lemma 3.19 with $X_l = \hat{X}_l = 4\gamma^2 \left(\|\theta_l^u\|_2^2 - \mathbb{E}_{\xi^l}\left[\|\theta_l^u\|_2^2\right]\right)$,

$c = c_1 = 32\gamma^2\lambda^2$ and $F = F_1 = \frac{c_1^2 \ln \frac{4N}{\beta}}{18}$ and get that for all $b > 0$

$$\mathbb{P}\left\{\left|\sum_{l=0}^{T-1} \hat{X}_l\right| > b \text{ and } \sum_{l=0}^{T-1} \hat{\sigma}_l^2 \le F_1\right\} \le 2\exp\left(-\frac{b^2}{2F_1 + 2c_1 b/3}\right)$$

or, equivalently, with probability at least $1 - 2\exp\left(-\frac{b^2}{2F_1+2c_1 b/3}\right)$

$$\text{either } \sum_{l=0}^{T-1} \hat{\sigma}_l^2 > F_1 \quad \text{or} \quad \underbrace{\left|\sum_{l=0}^{T-1} \hat{X}_l\right|}_{|③|} \le b.$$

As in our derivations of the upper bound for ① we choose such b that $2\exp\left(-\frac{b^2}{2F_1+2c_1 b/3}\right) = \frac{\beta}{2N}$, i.e.,

$$b = \frac{c_1 \ln \frac{4N}{\beta}}{3} + \sqrt{\frac{c_1^2 \ln^2 \frac{4N}{\beta}}{9} + 2F_1 \ln \frac{4N}{\beta}} \le c_1 \ln \frac{4N}{\beta} = 32\gamma^2\lambda^2 \ln \frac{4N}{\beta}.$$

That is, with probability at least $1 - \frac{\beta}{2N}$

$$\underbrace{\text{either } \sum_{l=0}^{T-1} \hat{\sigma}_l^2 > F_1 \quad \text{or} \quad |③| \le 32\gamma^2\lambda^2 \ln \frac{4N}{\beta}}_{\text{probability event } E_③}.$$

Next, we notice that probability event E_{T-1} implies that

$$\sum_{l=0}^{T-1} \hat{\sigma}_l^2 \overset{(3.158)}{\le} 8\gamma^2 c_1 \sum_{l=0}^{T-1} \mathbb{E}_{\xi^l}\left[\left\|\theta_l^u\right\|_2^2\right] \overset{(3.172)}{\le} 144\gamma^2 c_1 \sigma^2 \frac{T}{m}$$
$$\overset{T \le N}{\le} 144\gamma^2 c_1 \sigma^2 \frac{N}{m} = \frac{c_1^2 \ln \frac{4N}{\beta}}{18} \le F_1.$$

Upper bound for ④. The probability event E_{T-1} implies

$$④ = 4\gamma^2 \sum_{l=0}^{T-1} \mathbb{E}_{\xi^l}\left[\|\theta_l^u\|_2^2\right] \overset{(3.172)}{\le} 72\gamma^2\sigma^2 \sum_{l=0}^{T-1} \frac{1}{m} \overset{T \le N}{\le} \frac{72\gamma^2 N\sigma^2}{m} \overset{(3.144)}{\le} \frac{8}{9}\lambda^2\gamma^2 \ln \frac{4N}{\beta}.$$

Upper bound for ⑤. Again, we use corollaries of probability event E_{T-1}:

$$⑤ = 4\gamma^2 \sum_{l=0}^{T-1} \|\theta_l^b\|_2^2 \overset{(3.170)}{\le} 64\gamma^2\sigma^4 \frac{T}{m^2\lambda^2} \overset{T \le N}{\le} 64\gamma^2\sigma^4 \frac{N}{m^2\lambda^2} \overset{(3.144)}{\le} \frac{64}{6561} \frac{\lambda^2\gamma^2 \ln^2 \frac{4N}{\beta}}{N}.$$

Now we summarize all bound that we have: probability event E_{T-1} implies

$$\begin{aligned} R_T^2 &\overset{(3.150)}{\leq} 2R_0^2 + 2\gamma^2 \sum_{l=0}^{T-1} \|\theta_l\|_2^2 - 2\gamma \sum_{l=0}^{T-1} \left\langle x^l - x^*, \theta_l \right\rangle \\ &\overset{(3.155)}{\leq} 2R_0^2 + ① + ② + ③ + ④ + ⑤, \\ ② &\leq \frac{8}{81}\lambda\gamma C R_0 \ln\frac{4N}{\beta}, \quad ④ \leq \frac{8}{9}\lambda^2\gamma^2 \ln\frac{4N}{\beta}, \quad ⑤ \leq \frac{64}{6561}\frac{\lambda^2\gamma^2 \ln^2 \frac{4N}{\beta}}{N}, \\ \sum_{l=0}^{T-1} \sigma_l^2 &\leq F, \quad \sum_{l=0}^{T-1} \hat{\sigma}_l^2 \leq F_1 \end{aligned}$$

and

$$\mathbb{P}\{E_{T-1}\} \geq 1 - \frac{(T-1)\beta}{N}, \quad \mathbb{P}\{E_①\} \geq 1 - \frac{\beta}{2N}, \quad \mathbb{P}\{E_③\} \geq 1 - \frac{\beta}{2N},$$

where

$$\begin{aligned} E_① &= \left\{ \text{either } \sum_{l=0}^{T-1} \sigma_l^2 > F \quad \text{or} \quad |①| \leq 4\gamma\lambda C R_0 \ln\frac{4N}{\beta} \right\}, \\ E_③ &= \left\{ \text{either } \sum_{l=0}^{T-1} \hat{\sigma}_l^2 > F_1 \quad \text{or} \quad |③| \leq 32\gamma^2\lambda^2 \ln\frac{4N}{\beta} \right\}. \end{aligned}$$

Taking into account these inequalities and our assumptions on λ and γ (see (3.144) and (3.145)) we get that probability event $E_{T-1} \cap E_① \cap E_③$ implies

$$\begin{aligned} R_T^2 &\overset{(3.150)}{\leq} 2R_0^2 + 2\gamma^2 \sum_{l=0}^{T-1} \|\theta_l\|_2^2 - 2\gamma \sum_{l=0}^{T-1} \left\langle x^l - x^*, \theta_l \right\rangle \\ &\leq 2R_0^2 + \left(\frac{4}{7} + \frac{8}{567} + \frac{16}{49} + \frac{4}{441} + \frac{64}{321489}\right) C^2 R_0^2 \overset{(3.147)}{\leq} C^2 R_0^2 \end{aligned} \tag{3.159}$$

Moreover, using union bound we derive

$$\mathbb{P}\left\{E_{T-1} \cap E_① \cap E_③\right\} = 1 - \mathbb{P}\left\{\overline{E}_{T-1} \cup \overline{E}_① \cup \overline{E}_③\right\} \geq 1 - \frac{T\beta}{N}. \tag{3.160}$$

That is, by definition of E_T and E_{T-1} we have proved that

$$\mathbb{P}\{E_T\} \overset{(3.159)}{\geq} \mathbb{P}\left\{E_{T-1} \cap E_① \cap E_③\right\} \overset{(3.160)}{\geq} 1 - \frac{T\beta}{N},$$

which implies that for all $k = 0, 1, \ldots, N$ we have $\mathbb{P}\{E_k\} \geq 1 - \frac{k\beta}{N}$. Then, for $k = N$ we have that with probability at least $1 - \beta$

$$AN\left(f(\bar{x}^N) - f(x^*)\right) \overset{(3.149)}{\leq} 2R_0^2 + 2\gamma^2 \sum_{k=0}^{N-1} \|\theta_k\|_2^2 - 2\gamma \sum_{k=0}^{N-1} \left\langle x^k - x^*, \theta_k \right\rangle \overset{(3.151)}{\leq} C^2 R_0^2.$$

Since $A = 2\gamma\left(1 - 2\gamma\left(\frac{1}{\varepsilon}\right)^{\frac{1-\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}\right) \overset{(3.145)}{\geq} \gamma$ we get that with probability at least $1-\beta$

$$f(\bar{x}^N) - f(x^*) \leq \frac{C^2 R_0^2}{AN} = \frac{C^2 R_0^2}{\gamma N}.$$

When

$$\gamma = \min\left\{\frac{\varepsilon^{\frac{1-\nu}{1+\nu}}}{8M_\nu^{\frac{2}{1+\nu}}}, \frac{R_0}{\sqrt{2N}\varepsilon^{\frac{\nu}{1+\nu}} M_\nu^{\frac{1}{1+\nu}}}, \frac{R_0^{1-\nu}}{2C^\nu M_\nu \ln\frac{4N}{\beta}}\right\}$$

we have that with probability at least $1-\beta$

$$f(\bar{x}^N) - f(x^*) \leq \max\left\{\frac{8C^2 M_\nu^{\frac{2}{1+\nu}} R_0^2}{\varepsilon^{\frac{1-\nu}{1+\nu}} N}, \frac{\sqrt{2}C^2 M_\nu^{\frac{1}{1+\nu}} R_0 \varepsilon^{\frac{\nu}{1+\nu}}}{\sqrt{N}}, \frac{2C^{2+\nu} M_\nu R_0^{1+\nu} \ln\frac{4N}{\beta}}{N}\right\}.$$

Next, we estimate the iteration and oracle complexities of the method and consider 3 possible situations.

1. If $\gamma = \frac{\varepsilon^{\frac{1-\nu}{1+\nu}}}{8M_\nu^{\frac{2}{1+\nu}}}$, then with probability at least $1-\beta$

$$f(\bar{x}^N) - f(x^*) \leq \frac{8C^2 M_\nu^{\frac{2}{1+\nu}} R_0^2}{\varepsilon^{\frac{1-\nu}{1+\nu}} N}.$$

In other words, clipped-SGD achieves $f(\bar{x}^N) - f(x^*) \leq \varepsilon$ with probability at least $1-\beta$ after

$$O\left(\frac{M_\nu^{\frac{2}{1+\nu}} R_0^2}{\varepsilon^{\frac{2}{1+\nu}}}\right)$$

iterations and requires

$$\begin{aligned} Nm &\overset{(3.144)}{=} O\left(\max\left\{N, \frac{N^2\sigma^2\gamma^2 \ln\frac{N}{\beta}}{R_0^2}\right\}\right) = O\left(\max\left\{N, \frac{N^2\varepsilon^{\frac{2(1-\nu)}{1+\nu}}\sigma^2 \ln\frac{N}{\beta}}{M_\nu^{\frac{4}{1+\nu}} R_0^2}\right\}\right) \\ &= O\left(\max\left\{\frac{M_\nu^{\frac{2}{1+\nu}} R_0^2}{\varepsilon^{\frac{2}{1+\nu}}}, \frac{\sigma^2 R_0^2}{\varepsilon^2} \ln\frac{M_\nu^{\frac{2}{1+\nu}} R_0^2}{\varepsilon^{\frac{2}{1+\nu}}\beta}\right\}\right) \end{aligned}$$

oracle calls.

2. If $\gamma = \frac{R_0}{\sqrt{2N}\,\varepsilon^{\frac{\nu}{1+\nu}} M_\nu^{\frac{1}{1+\nu}}}$, then with probability at least $1-\beta$

$$f(\bar{x}^N) - f(x^*) \leq \frac{\sqrt{2}C^2 M_\nu^{\frac{1}{1+\nu}} R_0 \varepsilon^{\frac{\nu}{1+\nu}}}{\sqrt{N}}.$$

In other words, clipped-SGD achieves $f(\bar{x}^N) - f(x^*) \leq \varepsilon$ with probability at least $1 - \beta$ after

$$O\left(\frac{M_\nu^{\frac{2}{1+\nu}} R_0^2}{\varepsilon^{\frac{2}{1+\nu}}}\right)$$

iterations and requires

$$\begin{aligned} Nm &\overset{(3.144)}{=} O\left(\max\left\{N, \frac{N^2\sigma^2\gamma^2 \ln\frac{N}{\beta}}{R_0^2}\right\}\right) = O\left(\max\left\{N, \frac{N\sigma^2 \ln\frac{N}{\beta}}{\varepsilon^{\frac{2\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}}\right\}\right) \\ &= O\left(\max\left\{\frac{M_\nu^{\frac{2}{1+\nu}} R_0^2}{\varepsilon^{\frac{2}{1+\nu}}}, \frac{\sigma^2 R_0^2}{\varepsilon^2} \ln\frac{M_\nu^{\frac{2}{1+\nu}} R_0^2}{\varepsilon^{\frac{2}{1+\nu}}\beta}\right\}\right) \end{aligned}$$

oracle calls.

3. If $\gamma = \frac{R_0^{1-\nu}}{2C^\nu M_\nu \ln\frac{4N}{\beta}}$, then with probability at least $1 - \beta$

$$f(\bar{x}^N) - f(x^*) \leq \frac{2C^{2+\nu} M_\nu R_0^{1+\nu} \ln\frac{4N}{\beta}}{N}.$$

In other words, clipped-SGD achieves $f(\bar{x}^N) - f(x^*) \leq \varepsilon$ with probability at least $1 - \beta$ after

$$O\left(\frac{M_\nu R_0^{1+\nu} \ln\frac{M_\nu R_0^{1+\nu}}{\varepsilon\beta}}{\varepsilon}\right)$$

iterations and requires

$$\begin{aligned} Nm &\overset{(3.144)}{=} O\left(\max\left\{N, \frac{N^2\sigma^2\gamma^2 \ln\frac{N}{\beta}}{R_0^2}\right\}\right) = O\left(\max\left\{N, \frac{N^2\sigma^2}{M_\nu^2 R_0^{2\nu} \ln\frac{N}{\beta}}\right\}\right) \\ &= O\left(\max\left\{\frac{M_\nu R_0^{1+\nu}}{\varepsilon}, \frac{\sigma^2 R_0^2}{\varepsilon^2}\right\} \ln\frac{M_\nu R_0^{1+\nu}}{\varepsilon\beta}\right) \end{aligned}$$

oracle calls.

Putting all together and noticing that $\ln\frac{M_\nu^{\frac{2}{1+\nu}} R_0^2}{\varepsilon^{\frac{2}{1+\nu}}\beta} = O\left(\ln\frac{M_\nu R_0^{1+\nu}}{\varepsilon\beta}\right)$ we get the desired result. □

It is possible to get rid of using large batchsizes without sacrificing the oracle complexity via a proper choice of γ.

Corollary 3.7 *Let the assumptions of Theorem 3.23 hold and*

$$\gamma = \min\left\{\frac{\varepsilon^{\frac{1-\nu}{1+\nu}}}{8M_\nu^{\frac{2}{1+\nu}}}, \frac{R_0}{\sqrt{2N}\varepsilon^{\frac{\nu}{1+\nu}}M_\nu^{\frac{1}{1+\nu}}}, \frac{R_0^{1-\nu}}{2C^\nu M_\nu \ln\frac{4N}{\beta}}, \frac{R_0}{9\sigma\sqrt{N\ln\frac{4N}{\beta}}}\right\}. \tag{3.161}$$

Then for all $k = 0, 1, \ldots, N-1$ *one can use* $m = 1$ *and to achieve* $f(\bar{x}^N) - f(x^*) \le \varepsilon$ *with probability at least* $1-\beta$ clipped-SGD *requires*

$$O\left(\max\left\{\frac{M_\nu^{\frac{2}{1+\nu}}R_0^2}{\varepsilon^{\frac{2}{1+\nu}}}, \max\left\{\frac{M_\nu R_0^{1+\nu}}{\varepsilon}, \frac{\sigma^2R_0^2}{\varepsilon^2}\right\}\ln\left(\frac{M_\nu R_0^{1+\nu}}{\varepsilon\beta} + \frac{\sigma^2R_0^2}{\varepsilon^2\beta}\right)\right\}\right) \tag{3.162}$$

iterations/oracle calls.

Proof First of all, we verify that $m = 1$ is a valid choice. The only assumption on m is given in (3.144):

$$m \overset{(3.144)}{\ge} \max\left\{1, \frac{81N\sigma^2}{\lambda^2\ln\frac{4N}{\beta}}\right\}$$

Since $\gamma \le \frac{R_0}{9\sigma\sqrt{N\ln\frac{4N}{\beta}}}$, we have

$$\begin{aligned}\max\left\{1, \frac{81N\sigma^2}{\lambda^2\ln\frac{4N}{\beta}}\right\} &\overset{(3.144)}{=} \max\left\{1, \frac{81\gamma^2\sigma^2N\ln\frac{4N}{\beta}}{R_0^2}\right\}\\ &\le \max\left\{1, \frac{R_0^2}{81\sigma^2N\ln\frac{4N}{\beta}}\cdot\frac{81\sigma^2N\ln\frac{4N}{\beta}}{R_0^2}\right\} = 1.\end{aligned}$$

Therefore, for γ given in (3.161) one can use $m = 1$.

Next, if the minimum in (3.161) is attained on any of the first three terms, then applying the derivations from the end of the proof of Theorem 3.23, we get that the method requires

$$O\left(\max\left\{\frac{M_\nu^{\frac{2}{1+\nu}}R_0^2}{\varepsilon^{\frac{2}{1+\nu}}}, \max\left\{\frac{M_\nu R_0^{1+\nu}}{\varepsilon}, \frac{\sigma^2R_0^2}{\varepsilon^2}\right\}\ln\frac{M_\nu R_0^{1+\nu}}{\varepsilon\beta}\right\}\right)$$

iterations/oracle calls to achieve $f(\bar{x}^N) - f(x^*) \le \varepsilon$ with probability at least $1-\beta$. If $\gamma = \frac{R_0}{9\sigma\sqrt{N\ln\frac{4N}{\beta}}}$, then with probability at least $1-\beta$

$$f(\bar{x}^N) - f(x^*) \overset{(3.146)}{\le} \frac{9C^2R_0\sigma\sqrt{\ln\frac{4N}{\beta}}}{\sqrt{N}}.$$

In other words, clipped-SGD achieves $f(\bar{x}^N) - f(x^*) \leq \varepsilon$ with probability at least $1 - \beta$ after

$$O\left(\frac{\sigma^2 R_0^2 \ln \frac{\sigma^2 R_0^2}{\varepsilon^2 \beta}}{\varepsilon^2}\right)$$

iterations/oracle calls. Putting all together we get the desired result. □

Strongly Convex Case

For the strongly convex problems, we consider restarted version of Algorithm 3 (R-clipped-SGD, see Algorithm 4) and derive high-probability complexity result for this version.

Algorithm 4 Restarted clipped-SGD (R-clipped-SGD): case $\nu \in [0, 1]$

Input: starting point x^0, number of restarts τ, number of steps of clipped-SGD in restarts $\{N_t\}_{t=1}^{\tau}$, batchsizes $\{m_t\}_{k=1}^{\tau}$, stepsizes $\{\gamma_t\}_{t=1}^{\tau}$, clipping parameters $\{B_t\}_{t=1}^{\tau}$

1: $\hat{x}^0 = x^0$
2: **for** $t = 1, \ldots, \tau$ **do**
3: Run clipped-SGD (Algorithm 3) for N_t iterations with batchsize m_t, stepsize γ_t, clipping parameter B_t, and starting point $\hat{x}^{t-1}$. Define the output of clipped-SGD by $\hat{x}^t$.
4: **end for**

Output: $\hat{x}^{\tau}$

Note that due to strong convexity the solution x^* is unique.

Theorem 3.24 *Assume that function f is μ-strongly convex, its stochastic gradient and its gradient satisfy* (3.140) *and* (3.141) *respectively with $\sigma > 0$, $\nu \in [0, 1]$, $M_\nu > 0$ on $Q = B_{7R_0}(x^*)$, where $R_0 \geq \|x^0 - x^*\|_2$. Let $\varepsilon > 0$, $\beta \in (0, 1)$, and for all $t = 1, \ldots, \tau$*

$$N_t = \max\left\{\frac{2C^4 M_\nu^{\frac{2}{1+\nu}} R_0^2}{2^t \varepsilon_t^{\frac{2}{1+\nu}}}, \frac{4C^{2+\nu} M_\nu R_0^{1+\nu} \ln \frac{16C^{2+\nu} M_\nu R_0^{1+\nu}}{2^{\frac{(1+\nu)t}{2}} \varepsilon_t \beta}}{2^{\frac{(1+\nu)t}{2}} \varepsilon_t}\right\}, \quad \varepsilon_t = \frac{\mu R_0^2}{2^{t+1}},$$

$$\lambda_t = \frac{R_0}{2^{\frac{t}{2}} \gamma_t \ln \frac{4N_t\tau}{\beta}}, \quad m_t \geq \max\left\{1, \frac{81 N_t \sigma^2}{\lambda_t^2 \ln \frac{4N_t\tau}{\beta}}\right\}, \quad \ln \frac{4N_t\tau}{\beta} \geq 2,$$

$$\gamma_t = \min\left\{\frac{\varepsilon_t^{\frac{1-\nu}{1+\nu}}}{8M_\nu^{\frac{2}{1+\nu}}}, \frac{R_0}{2^{\frac{t}{2}}\sqrt{2N_t}\varepsilon_t^{\frac{\nu}{1+\nu}} M_\nu^{\frac{1}{1+\nu}}}, \frac{R_0^{1-\nu}}{2^{1+\frac{(1-\nu)t}{2}} C^\nu M_\nu \ln \frac{4N_t\tau}{\beta}}\right\}.$$

Then R-clipped-SGD *achieves $f(\bar{x}^\tau) - f(x^*) \leq \varepsilon$ with probability at least $1 - \beta$ after*

$$O\left(\max\left\{D_1^{\frac{2}{1+\nu}}\ln\frac{\mu R_0^2}{\varepsilon}, D_2^{\frac{2}{1+\nu}}, \max\left\{D_1\ln\frac{\mu R_0^2}{\varepsilon}, D_2\right\}\ln\frac{D}{\beta}\right\}\right)$$

iterations of Algorithm 3 in total and requires

$$O\left(\max\left\{D_1^{\frac{2}{1+\nu}}\ln\frac{\mu R_0^2}{\varepsilon}, D_2^{\frac{2}{1+\nu}}, \max\left\{D_1\ln\frac{\mu R_0^2}{\varepsilon}, D_2, \frac{\sigma^2}{\mu\varepsilon}\right\}\ln\frac{D}{\beta}\right\}\right) \tag{3.163}$$

oracle calls, where

$$D_1 = \frac{M_\nu}{\mu R_0^{1-\nu}}, \quad D_2 = \frac{M_\nu}{\mu^{\frac{1+\nu}{2}}\varepsilon^{\frac{1-\nu}{2}}}, \quad D = D_2\ln\frac{\mu R_0^2}{\varepsilon}.$$

Proof Applying Theorem 3.23, we obtain that with probability at least $1-\frac{\beta}{\tau}$

$$f(\hat{x}^1) - f(x^*) \le \frac{\mu R_0^2}{4}.$$

Since f is μ-strongly convex we have

$$\frac{\mu\|\hat{x}^1 - x^*\|_2^2}{2} \le f(\hat{x}^1) - f(x^*).$$

Therefore, with probability at least $1-\frac{\beta}{\tau}$

$$f(\hat{x}^1) - f(x^*) \le \frac{\mu R_0^2}{4}, \quad \|\hat{x}^1 - x^*\|_2^2 \le \frac{R_0^2}{2}.$$

From mathematical induction and the union bound for probability events it follows that inequalities

$$f(\hat{x}^t) - f(x^*) \le \frac{\mu R_0^2}{2^{t+1}}, \quad \|\hat{x}^t - x^*\|_2^2 \le \frac{R_0^2}{2^t}$$

hold simultaneously for $t = 1, \ldots, \tau$ with probability at least $1-\beta$. In particular, it means that after $\tau = \left\lceil\log_2\frac{\mu R_0^2}{\varepsilon}\right\rceil - 1$ restarts R-clipped-SGD finds an ε-solution with probability at least $1-\beta$. The total number of iterations $\hat{N}$ is

$$\begin{aligned}
\sum_{t=1}^{\tau} N_t &= O\left(\sum_{t=1}^{\tau} \max\left\{\frac{M_\nu^{\frac{2}{1+\nu}} R_0^2}{2^t \varepsilon_t^{\frac{2}{1+\nu}}}, \frac{M_\nu R_0^{1+\nu}}{2^{\frac{(1+\nu)t}{2}} \varepsilon_t} \ln \frac{M_\nu R_0^{1+\nu} \tau}{2^{\frac{(1+\nu)t}{2}} \varepsilon_t \beta}\right\}\right) \\
&= O\left(\sum_{t=1}^{\tau} \max\left\{\frac{M_\nu^{\frac{2}{1+\nu}} \cdot 2^{\frac{(1-\nu)t}{1+\nu}}}{\mu^{\frac{2}{1+\nu}} R_0^{\frac{2(1-\nu)}{1+\nu}}}, \frac{M_\nu \cdot 2^{\frac{(1-\nu)t}{2}}}{\mu R_0^{1-\nu}} \ln \frac{M_\nu \cdot 2^{\frac{(1-\nu)\tau}{2}} \tau}{\mu R_0^{1-\nu} \beta}\right\}\right) \\
&= O\left(\max\left\{\frac{M_\nu^{\frac{2}{1+\nu}}}{\mu^{\frac{2}{1+\nu}} R_0^{\frac{2(1-\nu)}{1+\nu}}}, \frac{M_\nu}{\mu R_0^{1-\nu}} \ln \frac{M_\nu \ln \frac{\mu R_0^2}{\varepsilon}}{\mu^{\frac{1+\nu}{2}} \varepsilon^{\frac{1-\nu}{2}} \beta}\right\} \cdot \max\left\{\ln \frac{\mu R_0^2}{\varepsilon}, \left(\frac{\mu R_0^2}{\varepsilon}\right)^{\frac{1-\nu}{2}}\right\}\right) \\
&= O\left(\max\left\{D_1^{\frac{2}{1+\nu}} \ln \frac{\mu R_0^2}{\varepsilon}, D_2^{\frac{2}{1+\nu}}, \max\left\{D_1 \ln \frac{\mu R_0^2}{\varepsilon}, D_2\right\} \ln \frac{D}{\beta}\right\}\right),
\end{aligned}$$

where

$$D_1 = \frac{M_\nu}{\mu R_0^{1-\nu}}, \quad D_2 = \frac{M_\nu}{\mu^{\frac{1+\nu}{2}} \varepsilon^{\frac{1-\nu}{2}}}, \quad D = D_2 \ln \frac{\mu R_0^2}{\varepsilon}.$$

Finally, the total number of oracle calls equals

$$\begin{aligned}
\sum_{t=1}^{\tau} \sum_{k=0}^{N_t - 1} m_k^t &= O\left(\max\left\{\sum_{t=1}^{\tau} N_t, \sum_{t=1}^{\tau} \frac{\sigma^2 R_0^2}{2^t \varepsilon_t^2} \ln \frac{M_\nu R_0^{1+\nu} \tau}{2^{\frac{(1+\nu)t}{2}} \varepsilon_t \beta}\right\}\right) \\
&= O\left(\max\left\{\hat{N}, \sum_{t=1}^{\tau} \frac{\sigma^2 \cdot 2^t}{\mu^2 R_0^2} \ln \frac{D}{\beta}\right\}\right) = O\left(\max\left\{\hat{N}, \frac{\sigma^2}{\mu \varepsilon} \ln \frac{D}{\beta}\right\}\right).
\end{aligned}$$

Eduard: we need the analysis of clipped-SGD and clipped-SSTM (for Hölder continuous gradients).

Useful Inequalities

For all $a, b \in \mathbb{R}^n$ and $\lambda > 0$

$$|\langle a, b \rangle| \leq \frac{\|a\|_2^2}{2\lambda} + \frac{\lambda \|b\|_2^2}{2}, \tag{3.164}$$

$$\|a + b\|_2^2 \leq 2\|a\|_2^2 + 2\|b\|_2^2, \tag{3.165}$$

$$\langle a, b \rangle = \frac{1}{2}\left(\|a + b\|_2^2 - \|a\|_2^2 - \|b\|_2^2\right). \tag{3.166}$$

Auxiliary Lemmas

Lemma 3.17 *Let f has (ν, M_ν)-Hölder continuous gradient on $Q \subseteq \mathbb{R}^n$. Then for all $x, y \in Q$ and for all $\delta > 0$*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{M_\nu}{1+\nu} \|x - y\|_2^{1+\nu}, \tag{3.167}$$

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L(\delta, \nu)}{2} \|x - y\|_2^2 + \frac{\delta}{2}, \quad L(\delta, \nu) = \left(\frac{1}{\delta}\right)^{\frac{1-\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}. \tag{3.168}$$

Lemma 3.18 *For all $k \geq 0$ the following inequality holds:*

$$\left\| \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k) - \mathbb{E}_{\boldsymbol{\xi}^k} \left[\widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k) \right] \right\|_2 \leq 2\lambda_{k+1}. \tag{3.169}$$

Moreover, if the stochastic gradient satisfies (3.140) *on $Q = B_{3R_0}(x^*)$ and $\|\nabla f(x^{k+1})\|_2 \leq \frac{\lambda_{k+1}}{2}$ for some $k \geq 0$, then for this k we have:*

$$\left\| \mathbb{E}_{\boldsymbol{\xi}^k} \left[\widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k) \right] - \nabla f(x^{k+1}) \right\|_2 \leq \frac{4\sigma^2}{m_k \lambda_{k+1}}, \tag{3.170}$$

$$\mathbb{E}_{\boldsymbol{\xi}^k} \left[\left\| \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k) - \nabla f(x^{k+1}) \right\|_2^2 \right] \leq \frac{18\sigma^2}{m_k}, \tag{3.171}$$

$$\mathbb{E}_{\boldsymbol{\xi}^k} \left[\left\| \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k) - \mathbb{E}_{\boldsymbol{\xi}^k} \left[\widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k) \right] \right\|_2^2 \right] \leq \frac{18\sigma^2}{m_k}. \tag{3.172}$$

Lemma 3.19 *Let the sequence of random variables $\{X_i\}_{i \geq 1}$ form a martingale difference sequence, i.e. $\mathbb{E}[X_i \mid X_{i-1}, \ldots, X_1] = 0$ for all $i \geq 1$. Assume that conditional variances $\sigma_i^2 \overset{def}{=} \mathbb{E}\left[X_i^2 \mid X_{i-1}, \ldots, X_1\right]$ exist and are bounded and assume also that there exists deterministic constant $c > 0$ such that $\|X_i\|_2 \leq c$ almost surely for all $i \geq 1$. Then for all $b > 0$, $F > 0$ and $n \geq 1$*

$$\mathbb{P}\left\{ \left| \sum_{i=1}^n X_i \right| > b \text{ and } \sum_{i=1}^n \sigma_i^2 \leq F \right\} \leq 2 \exp\left(-\frac{b^2}{2F + 2cb/3} \right). \tag{3.173}$$

Technical Lemmas

Lemma 3.20 *Let f have Hölder continuous gradients on $Q \subseteq \mathbb{R}^n$ for some $\nu \in [0, 1]$ with constant $M_\nu > 0$, be convex and $x^* \in Q$ be some minimum of $f(x)$ on $\mathbb{R}^n$. Then, for all $x \in \mathbb{R}^n$ and all $\delta > 0$,*

$$\|\nabla f(x)\|_2^2 \leq 2 \left(\frac{1}{\delta}\right)^{\frac{1-\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}} (f(x) - f(x^*)) + \delta^{\frac{2\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}. \tag{3.174}$$

Proof For a given $\delta > 0$ we consider an arbitrary point $x \in Q$ and $y = x - \frac{1}{L(\delta,\nu)} \nabla f(x)$, where $L(\delta, \nu) = \left(\frac{1}{\delta}\right)^{\frac{1-\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}$. Since $x^* \in Q$ and f is convex one can easily show that $y \in Q$. For the pair of points x, y we apply (3.168) and get

$$
\begin{aligned}
f(y) &\leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L(\delta, \nu)}{2} \|x - y\|_2^2 + \frac{\delta}{2} \\
&= f(x) - \frac{1}{2L(\delta, \nu)} \|x - y\|_2^2 + \frac{\delta}{2}
\end{aligned}
$$

implying

$$
\begin{aligned}
\|\nabla f(x)\|_2^2 &\leq 2L(\delta, \nu) \left(f(x) - f(y)\right) + \delta L(\delta, \nu) \\
&\leq 2 \left(\frac{1}{\delta}\right)^{\frac{1-\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}} \left(f(x) - f(x^*)\right) + \delta^{\frac{2\nu}{1+\nu}} M_\nu^{\frac{2}{1+\nu}}.
\end{aligned}
$$

3.5 Historical Notes

SGD was proposed in the seminal work [103]. The analysis of SGD under Assumption 3.1 was developed in [**?**].

I suggest to write here just short sentences to acknowledge who proposed the method/analysis/idea that we mention in the chapter.

Eduard: don't forget to mention recent advances for over-parameterized modeld (line searches, Polyak stepsizes)

Chapter 4
Adaptive Methods for Stochastic optimization Problems and Stochastic Variational Inequalities

Abstract TODO
TODO: write a short introduction

4.1 AdaGrad non-smooth case

Step size adaptivity is a most interesting type of adaptivity since stochastic gradient descent for stochastic approximation approach to stochastic optimization is very sensitive to the choice of step size schedule. For example, asymptotically optimal step size schedule $\gamma_k \sim k^{-1}$ is known to be impractical, while step sizes $\gamma_k \sim k^{-1/2}$ proposed by Polyak and Juditsky, together with averaging, allows one to achieve much better performance. In the non-smooth convex case, these algorithms match unimprovable $O(k^{-1/2})$ convergence rate. Further improvements will address more advanced trajectory-dependent step size schedules.

Note, that mentioned asymptotic optimal step size schedule, more precisely, is $\gamma_k = \nabla^2 f(x^*) \cdot k^{-1/2}$ and is of the matrix form, unlike standard ones. AdaGrad (and its versions) we consider in this section is based on the idea of variable matrix step size or, more generally, variable metric.

Let's move on to the description of gradient method in variable metric setting. Its iteration is

$$x^{k+1} = \min_{x \in Q} \left\{ \langle g^k, x \rangle + \frac{1}{2} \langle x, H_k x \rangle \right\} \tag{4.1}$$

Theorem 4.1 *For the algorithm with iteration* (4.1)*, it holds that*

$$\mathbb{E}\left[\sum_{k=0}^{N} (f(x^k) - f(x^*))\right] \leq \frac{1}{2}\|x^0 - x^*\|_{H_0}^2 + \frac{1}{2}\mathbb{E}\left[\sum_{k=0}^{N} \|g^k\|_{H_k^{-1}}^2\right] + \frac{1}{2}\mathbb{E}\left[\sum_{k=0}^{N-1} (\|x^{k+1} - x^*\|_{H_{k+1}}^2 - \|x^{k+1} - x^*\|_{H_k}^2)\right]. \tag{4.2}$$

Proof Projection is non-expansive, so

$$\begin{aligned}\frac{1}{2}\|x^{k+1}-x^*\|_{H_k}^2 &\le \frac{1}{2}\|x^k - H_k^{-1}g^k - x^*\|_{H_k}^2 \\ &= \frac{1}{2}\|x^k-x^*\|_{H_k}^2 - \langle g^k, x^k - x^*\rangle + \frac{1}{2}\|g^k\|_{H_k^{-1}}^2.\end{aligned} \tag{4.3}$$

By convexity,

$$\frac{1}{2}\mathbb{E}\left[\|x^{k+1}-x^*\|_{H_k}^2\right] = \mathbb{E}\left[\frac{1}{2}\|x^k-x^*\|_{H_k}^2 - (f(x^k)-f(x^*)) + \frac{1}{2}\|g^k\|_{H_k^{-1}}^2\right]. \tag{4.4}$$

Rearranging,

$$\mathbb{E}\left[f(x^k)-f(x^*)\right] = \frac{1}{2}\mathbb{E}\left[\|x^k-x^*\|_{H_k}^2 - \|x^{k+1}-x^*\|_{H_k}^2 + \|g^k\|_{H_k^{-1}}^2\right]. \tag{4.5}$$

Summing up from $k=1$ to N finishes the proof. □

Motivation of AdaGrad is to lower the second term in (4.2) by a proper choice of H_k. Let's consider such an optimization problem:

$$\min_S \sum_{k=0}^N \langle g^k, S^{-1}g^k\rangle, \tag{4.6}$$

such that $S \in \mathbb{R}^{n\times n}$ is positive-definite and $\mathrm{Tr}(S) \le c$. This problem can be solved using Lagrange multipliers method, and the solution is $S = cG_N^{-1/2}/\mathrm{Tr}(G_N^{-1/2})$ with $G_N = \sum_{k=0}^N g^k g^{k\top}$. Iteration of AdaGrad is, therefore, (4.1) with $H_k = G_k^{1/2}$. This leads to the following convergence guarantees.

Theorem 4.2 *For the* AdaGrad*, it holds that*

$$\begin{aligned}\mathbb{E}\left[\sum_{k=0}^N (f(x^k)-f(x^*))\right] &\le \frac{L}{2}\|x^0-x^*\|_2^2 + \mathbb{E}\left[\mathrm{Tr}(G_N^{1/2})\right] \\ &\quad + \frac{1}{2}\mathbb{E}\left[\max_{k=1,\dots,N}\|x^k-x^*\|_2^2 \cdot \mathrm{Tr}(G_N^{1/2})\right],\end{aligned} \tag{4.7}$$

where $L \ge 0$ *is such that* $\max_{k=1,\dots,N}\|g^k\|_2 \le L$.

Proof $\mathrm{Tr}(A) \ge \lambda_{\max}(A)$ for positive-definite $A \in \mathbb{R}^{n\times n}$, so

$$\begin{aligned}&\|x^{k+1}-x^*\|_{G_{k+1}^{1/2}}^2 - \|x^{k+1}-x^*\|_{G_k^{1/2}}^2 \\ &\quad = \langle x^{k+1}-x^*, (G_{k+1}^{1/2}-G_k^{1/2})(x^{k+1}-x^*)\rangle \\ &\quad \le \|x^{k+1}-x^*\|_2^2 \cdot \mathrm{Tr}(G_{k+1}^{1/2}-G_k^{1/2}).\end{aligned} \tag{4.8}$$

On the other hand,

$$\sum_{k=0}^{N-1} \|x^{k+1} - x^*\|_2^2 \cdot \mathrm{Tr}(G_{k+1}^{1/2} - G_k^{1/2}) \tag{4.9}$$
$$\leq \max_{k=1,\dots,N} \|x^k - x^*\|_2^2 \cdot \mathrm{Tr}(G_N^{1/2}) - \|x^0 - x^*\|_2^2 \cdot \mathrm{Tr}(G_0^{1/2}).$$

Further, let's demonstrate that

$$\sum_{k=0}^{N} \langle g^k, G_k^{-1/2} g^k \rangle \leq 2 \sum_{k=0}^{N} \langle g^k, G_N^{-1/2} g^k \rangle = 2\mathrm{Tr}(G_N^{-1/2}), \tag{4.10}$$

by induction. The base case is

$$\langle g^0, G_0^{-1/2} g^0 \rangle = \frac{\langle g^0, g^0 \rangle}{\|g^0\|_2} \leq 2\|g^0\|_2. \tag{4.11}$$

Induction assumption for $N-1$ implies

$$\sum_{k=0}^{N} \langle g^k, G_k^{-1/2} g^k \rangle \leq 2 \sum_{k=0}^{N-1} \langle g^k, G_{N-1}^{-1/2} g^k \rangle + \langle g^N, G_N^{-1/2} g^N \rangle \tag{4.12}$$
$$= 2\mathrm{Tr}\left(G_{N-1}^{-1/2} \cdot \sum_{k=0}^{N-1} g^k g^{k\top} \right) + \langle g^N, G_N^{-1/2} g^N \rangle$$
$$= 2\mathrm{Tr}(G_{N-1}^{-1/2} G_{N-1}) + \langle g^N, G_N^{-1/2} g^N \rangle.$$

Using (without proof) the relation

$$2\mathrm{Tr}((G - gg^\top)^{1/2}) \leq 2\mathrm{Tr}(G^{-1/2}) - \mathrm{Tr}(G^{-1/2} g g^\top), \tag{4.13}$$

G is positive-definite, for $G = G_N$, $g = g^N$, we proof the statement for N and for all $N \geq 1$.

Finally, $\|x^0 - x^*\|_{H_0}^2 \leq \|x^0 - x^*\|_2^2 \cdot \mathrm{Tr}(G_0^{1/2})$. Combining it with (4.9), (4.10) and Theorem 4.2 finishes the proof. □

This version of AdaGrad is full-matrix which may be computationally expensive. Very practical and wide-spread version of AdaGrad uses diagonal matrices H_k. More precisely, instead of optimization problem (4.6) defining H_k, we consider such one:

$$\min_s \sum_{k=0}^{N} \sum_{i=1}^{n} \frac{[g^k]_i^2}{s_i}, \tag{4.14}$$

such that $\langle s, 1 \rangle \leq c$, $s_i \geq 0$ for all $i = 1, ..., n$. The solution for this problem is $s_i = \sqrt{\sum_{k=0}^{N} [g^k]_i^2}$. So, we can use $H_k = (s)$. This leads to the following

Corollary 4.1 *For the* AdaGrad *with diagonal adaptation, it holds that*

$$\mathbb{E}\left[\sum_{k=0}^{N}(f(x^k)-f(x^*))\right] \leq \frac{L}{2}\|x^0-x^*\|_2^2+\mathbb{E}\left[\sum_{i=1}^{n}\sqrt{\sum_{k=0}^{N}[g^k]_i^2}\right] \tag{4.15}$$
$$+\frac{1}{2}\mathbb{E}\left[\max_{k=1,\ldots,N}\|x^k-x^*\|_2^2\cdot\sum_{i=1}^{n}\sqrt{\sum_{k=0}^{N}[g^k]_i^2}\right],$$

where $L \geq 0$ *is such that* $\max_{k=1,\ldots,N}\|g^k\|_2 \leq L$.

To understand this convergence rate, let's bound gradient term:

$$\sum_{i=1}^{n}\sqrt{\sum_{k=0}^{N}[g^k]_i^2} \leq \sqrt{n\sum_{k=0}^{N}\|g^k\|^2} \leq L\sqrt{nN}. \tag{4.16}$$

Therefore, asymptotic convergence rate of AdaGrad is $O(1/\sqrt{N})$, the same as for SGD. On the other hand, former gradient term depending on $\|g^k\|_{H_k^{-1}}$ in AdaGrad reaches its minimum. More precisely,

4.2 Adam smooth case

TODO

4.3 Universal stochastic Mirror-Prox for variational inequalities

Let Q be convex subset of $\mathbb{R}^n$, $F: Q \longrightarrow \mathbb{R}^n$ be a monotone operator, i.e.

$$\langle F(x)-F(y), x-y\rangle \geq 0, \quad \forall x, y \in Q.$$

We focus on the stochastic setting of the problem of finding a vector $x^* \in Q$ (called a strong solution), such that

$$\langle F(x^*), x^*-x\rangle \leq 0, \quad \forall x \in Q. \tag{4.17}$$

This means that we have an access to an unbiased noisy estimate of the exact monotone mapping $F(x)$. That is, we have an access to an oracle $\tilde{F}: Q \longrightarrow \mathbb{R}^n$, such that for any $x \in Q$ we have

$$\mathbb{E}[\tilde{F}(x)|x] = F(x).$$

Since F is a monotone operator, a strong solution is also a weak solution, that is $\langle F(x), x^*-x\rangle \leq 0,\ \forall x \in Q$ [**?**].

We assume that there exists a bound G on all of unbiased estimates of the operator F, i.e. $\|\tilde{F}(x)\|_* \leq G,\ \forall x \in Q$, where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$.

Definition 4.1 [9] Let $Q \subseteq \mathbb{R}^n$ be a convex set, and $\Delta : Q \times Q \longrightarrow \mathbb{R}$ is convex with respect to its first argument. We say that the function Δ is a *gap function (or merit function)* compatible with the monotone operator $F : Q \longrightarrow \mathbb{R}^n$, if

$$\Delta(x, y) \leq \langle F(x), x - y\rangle; \quad \forall x, y \in Q.$$

The duality gap of Δ is defined as $\text{DualGap}(x) := \max\limits_{y \in Q} \Delta(x, y)$. We say that x^* is a solution of (4.17) if and only if $\text{DualGap}(x^*) = 0$

For the stochastic setting of the problem (4.17), in [9], it was proposed, in the general set-up of arbitrary norms and compatible Bregman divergences, a universal algorithm, listed as Algorithm 5 below. This algorithm simultaneously achieves the optimal rates for the smooth/non-smooth, and noisy/noiseless settings.

Algorithm 5 Universal Stochastic Mirror-Prox [9].

Input: Number of iterations $N > 0$, a prox-function d and the connected Bregmann divergence $V(\cdot, \cdot)$, $y_0 = \text{argmin}_{x \in Q}\, d(x)$, $D > 0$, s.t. $D^2 = \max_{x \in Q} d(x) - \min_{x \in Q} d(x)$, $G_0 > 0$.

1: **for** $k = 1, \ldots, N$ **do**

2: Set $\tilde{M}_k := \tilde{F}(y_{k-1})$, the noisy estimate of $M_k := F(y_{k-1})$.

3: Calculate

$$\gamma_k = \frac{D}{\sqrt{G_0^2 + \sum_{i=1}^{k-1} w_i^2}}, \quad \text{where } w_i^2 := \frac{\|x_i - y_i\|^2 + \|x_i - y_{i-1}\|^2}{5\eta_i^2}. \tag{4.18}$$

4: Calculate $x_k = \underset{x \in Q}{\text{argmin}} \left\{\langle \tilde{M}_k, x\rangle + \frac{1}{\gamma_k} V(x, y_{k-1})\right\}$.

5: Set $\tilde{g}_k := \tilde{F}(x_k)$, the noisy estimate of $g_k := F(x_k)$.

6: Calculate $y_k = \underset{x \in Q}{\text{argmin}} \left\{\langle \tilde{g}_k, x\rangle + \frac{1}{\gamma_k} V(x, y_{k-1})\right\}$.

7: **end for**

Output: $\overline{x}_N = \frac{1}{N}\sum_{k=1}^{N} x_k$.

For Algorithm 5, we mention the first results regards the non-smooth noisy case, in the following theorem

Theorem 4.3 *For the output point $\overline{x}_N \in \mathbb{R}^n$, after N iterations of Algorithm 5, it holds the following bound*

$$\mathbb{E}\left[\text{DualGap}\left(\overline{x}_N\right)\right] := \mathbb{E}\left[\max_{x \in Q} \Delta\left(\overline{x}_N, x\right)\right] \leq O\left(\frac{\alpha G D \sqrt{\log N}}{\sqrt{N}}\right), \tag{4.19}$$

where

$$G := \max_{1 \leq k \leq N} \left\{\max\{\|g_k\|_*, \|M_k\|_*\}\right\}, \quad \textit{and} \quad \alpha := \max\left\{\frac{G}{G_0}, \frac{G_0}{G}\right\}$$

Proof (Proof Sketch) Assuming that the learning rate sequence $\{\gamma_k\}_{k\geq 1}$ is monotonically non-increasing, then for any $x \in Q$ we have

$$\begin{aligned}\sum_{k=1}^{N} \tilde{g}_k\,(x_k - x) \leq \frac{D^2}{\gamma_1} + \frac{D^2}{\eta_N} + \sum_{k=1}^{N} \left\|\tilde{g}_k - \tilde{M}_k\right\|_* \cdot \|x_k - y_k\| \\ - \frac{1}{2}\sum_{k=1}^{N} \gamma_k^{-1}\left(\|x_k - y_k\|^2 + \|x_k - y_{k-1}\|^2\right),\end{aligned} \tag{4.20}$$

and (see [9])

$$\sum_{k=1}^{N} \tilde{g}_k\,(x_k - x) \leq O(\alpha G D\sqrt{N\log N}). \tag{4.21}$$

Let we set $\zeta_k := \tilde{g}_k - g_k$. Then we have $\mathbb{E}[\zeta_k | x_k] = 0$ and $\{\zeta_k\}_k$ is a martingale difference sequence.

From the definition of the output point $\bar{x}_N$, and using Jensen's inequality, for any $x \in Q$, we obtain

$$\begin{aligned} N \cdot \Delta\,(\bar{x}_N, x) &\leq \sum_{k=1}^{N} \Delta\,(x_k, x) \\ &\leq \sum_{k=1}^{N} g_k\,(x_k - x) \\ &= \sum_{k=1}^{N} \tilde{g}_k\,(x_k - x) - \sum_{k=1}^{N} \zeta_k\,(x_k - x) \\ &\leq O\left(\alpha G D\sqrt{N\log N}\right) - \sum_{k=1}^{N} \zeta_k\,(x_k - x)\,, \end{aligned} \tag{4.22}$$

Let $x^* = \operatorname{argmax}_{x\in Q} \Delta(\bar{x}_N, x)$, by substituting $x = x^*$ in (4.22) and taking the expectation over (4.22), we find

$$\begin{aligned} N \cdot \mathbb{E}\left[\Delta\,(\bar{x}_N, x^*)\right] &\leq O\left(\alpha G D\sqrt{N\log N}\right) - \mathbb{E}\left[\sum_{k=1}^{N} \zeta_k\,(x_k - x^*)\right] \\ &= O\left(\alpha G D\sqrt{N\log N}\right) + \mathbb{E}\left[\sum_{k=1}^{N} \zeta_k x^*\right]. \end{aligned} \tag{4.23}$$

But (see [9])

$$\mathbb{E}\left[\sum_{k=1}^{N} \zeta_k x^*\right] \leq O\left(G D\sqrt{N\log N}\right) \tag{4.24}$$

By combining (4.23) and (4.24) we find the desired estimate (4.19). □

The next theorem regards the smooth noisy case.

Theorem 4.4 *Let F be an L-smooth operator on Q, and $\mathbb{E}\left[\left\|\tilde{F}(x) - F(x)\right\|_*^2 | x\right] \leq \sigma^2$; $\forall x \in Q$. Then for the output point $\overline{x}_N \in \mathbb{R}^n$, after N iterations of Algorithm 5, it holds the following bound*

$$\mathbb{E}\left[\text{DualGap}\left(\overline{x}_N\right)\right] \leq O\left(\frac{\alpha GD + \alpha^2 LD^2 + LD^2 \log\left(LD/G_0\right)}{N} + \frac{\alpha\sigma D\sqrt{\log N}}{\sqrt{N}}\right). \tag{4.25}$$

Proof (Proof Sketch) For $\tilde{g}$ and $\tilde{M}$, from [**?**], we have

$$\left\|\tilde{g}_k - \tilde{M}_k\right\|_* \cdot \|x_k - y_k\| = \min_{\rho>0}\left\{\frac{\rho}{2}\left\|\tilde{g}_k - \tilde{M}_k\right\|_*^2 + \frac{1}{2\rho}\|x_k - y_k\|^2\right\}.$$

Since F is an L-smooth operator and by taking $\rho = \frac{1}{L}$, we find

$$\left\|\tilde{g}_k - \tilde{M}_k\right\|_* \cdot \|x_k - y_k\| \leq \frac{L}{2}\|x_k - y_{k-1}\|^2 + \frac{L}{2}\|x_k - y_k\|^2. \tag{4.26}$$

Let $\xi_k := \tilde{g}_k - g_k - (\tilde{M}_k - M_k)$. By using (4.26) with triangle inequality we get,

$$\left\|\tilde{g}_k - \tilde{M}_k\right\|_* \cdot \|x_k - y_k\| \leq \|\xi_k\|_* \cdot \|x_k - y_k\| + \frac{L}{2}\|x_k - y_{k-1}\|^2 + \frac{L}{2}\|x_k - y_k\|^2,$$

Now, for any $x \in Q$, we have

$$\sum_{k=1}^{N} \tilde{g}_k\left(x_k - x\right) \leq O\left(\alpha DG + \alpha^2 LD^2 + LD^2 \log\left(\frac{LD}{G_0}\right)\right) + \sum_{k=1}^{N}\|\xi_k\|_* \cdot \|x_k - y_k\|. \tag{4.27}$$

Let we set $\zeta_k := \tilde{g}_k - g_k$. From the definition of the output point $\overline{x}_N$, and using Jensen's inequality, for $x^* = \text{argmax}_{x\in Q}\,\Delta(\overline{x}_N, x)$, we obtain

$$\begin{aligned} N \cdot \mathbb{E}\left[\Delta\left(\bar{x}_N, x^*\right)\right] &\leq \mathbb{E}\sum_{k=1}^{N}\tilde{g}_k\left(x_k - x\right) - \mathbb{E}\left[\sum_{k=1}^{N}\zeta_k\left(x_k - x\right)\right] \\ &\leq O\left(\alpha DG + \alpha^2 LD^2 + LD^2\log\left(\frac{LD}{G_0}\right)\right) \\ &\quad + \mathbb{E}\left[\sum_{k=1}^{N}\|\xi_k\|_* \cdot \|x_k - y_k\|\right] + \mathbb{E}\left[\sum_{k=1}^{N}\zeta_k x^*\right]. \end{aligned} \tag{4.28}$$

But

$$\mathbb{E}\left[\sum_{k=1}^{N}\|\xi_k\|_* \cdot \|x_t - y_k\|\right] \leq 12\alpha D\sigma\sqrt{T(1+\log T)}, \tag{4.29}$$

and

$$\mathbb{E}\left[\sum_{k=1}^{N}\zeta_k x^*\right] \leq \frac{D}{2}\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\|\zeta_k\|_*^2\right]} \leq D\sigma\sqrt{N}. \tag{4.30}$$

Thus, as a result, we have the following estimate

$$N\cdot\mathbb{E}\left[\Delta\left(\bar{x}_N, x^*\right)\right] \leq O\left(\alpha DG + \alpha^2 LD^2 + LD^2\log\left(\frac{LD}{G_0}\right)\right) + O(\alpha\sigma D\sqrt{N\log N}), \tag{4.31}$$

from which we find the desired estimate (4.25). □

4.3.1 Geometry-aware universal Mirror-Prox

In Algorithm 5, the proposed step-size depends on the norm of the past updates at each iteration. In [**?**] it was relaxed the step-size dependence from norm to Bregman divergence. This relaxation makes the step-size more geometry-aware and allows to extends the analysis of universal Mirror Prox to the settings where the operator is not smooth or bounded, which occur in many applied problems, such as support vector machine, GAN with Kullback-Leibler losses or resource allocation problem.

Let we introduce the notions of the *Bregman smoothness* and *Bregman Boundedness*, based on the local norm on the convex set Q, which were introduced in [**?**], in order to overcame the problem of the failure of the Lipschitz continuity to solve many practical problems that appear in many applications.

Definition 4.2 Let $\mathcal{V}$ be a finite dimensional vector space and $\mathcal{V}^*$ its dual. Let $\mathcal{Z} = span(Q-Q)$, the subspace of $\mathcal{V}$ spanned by all vectors $\{x-y|\, x,y\in Q\}$. A *local norm* on Q is a continuous assignment of a norm $\|\cdot\|_x$ on $\mathcal{Z}$ at each $x\in Q$. The induced *dual local norm*, for all $v\in\mathcal{V}^*$, is defined as follows

$$\|v\|_{x,*} = \max_{z\in\mathcal{Z}}\left\{|\langle v,z\rangle| : \|z\|_x \leq 1\right\}.$$

Here we introduce two examples for clarity of the definition 4.2 (see [**?**])

Euclidean norm

Let $Q=\mathcal{Z}=\mathbb{R}^n$. The *Euclidean norm* on Q is given by the standard expression $x=\sum_{i=1}^n x_i^2$, for any $x=(x_1,\ldots,x_n)\in Q$. The associated dual norm is the same.

Shahshahani p-norm

Let $Q=\mathbb{R}^n_{++}$ and $\mathcal{Z}=\mathbb{R}^n$. For all $p>1$, the Shahshahani p-norm on Q is defined as

$$\|z\|_x = \left(\frac{|z_1|^p}{x_1}+\cdots+\frac{|z_n|^p}{x_n}\right)^{1/p}, \quad \forall x\in Q, z\in\mathcal{Z}.$$

By Hölder's inequality, the corresponding dual norm is given by

$$\|v\|_{x,*} = \left(x_1^{q-1}\,|v_1|^q + \cdots + x_d^{q-1}\,|v_n|^q\right)^{1/q},$$

with $p^{-1} + q^{-1} = 1$.

Definition 4.3 Let $\|\cdot\|_x$ be a local norm on Q and d be a prox-function on Q. An operator $F : Q \longrightarrow \mathcal{V}^*$ is L-Bregman smooth on Q if

$$\|F(y) - F(x)\|_{x,*} \leq L\sqrt{2V(y,x)}, \quad \forall x, y \in Q.$$

Note that, if we take $d(x) = \frac{1}{2}\|x\|^2$, we recover the standard Lipschitz continuity condition

$$\|F(y) - F(x)\|_{x,*} \leq L\|y - x\|, \quad \forall x, y \in Q.$$

Definition 4.4 An operator $F : Q \longrightarrow \mathcal{V}^*$ is M-Bregman bounded on Q , if there exist $M > 0$, such that

$$\|F(x)\|_* \leq M\frac{\sqrt{V(y,x)}}{\|y - x\|}, \quad \forall x, y \in Q.$$

Note that, if we take $d(x) = \frac{1}{2}\|x\|^2$, we recover the standard recover the bounded setting $\|F(x)\|_* \leq M$.

Let d be a prox-function on a convex set Q and $V(\cdot,\cdot)$ the Bregmann divergence connected with d. For the stocahstic setting of the problem (4.17), we assume the following assumpsions

A1: The unbiased estimator of the operator F, i.e.

$$\mathbb{E}[\tilde{F}(x)|x] = F(x), \ \forall x \in Q.$$

A1: The bounded variance, i.e.

$$\mathbb{E}\left[\left\|\tilde{F}(x) - F(x)\right\|_*^2 |x\right] \leq \sigma^2; \ \forall x \in Q.$$

A3: The almost sure boundedness, i.e. there exist $G' > 0$, such that $\|\tilde{F}(x)\|_* \leq G'$, for all $x \in Q$.

A4: $V(x^*, x) \leq D^2$, where $D^2 = \max_{x\in Q} d(x) - \min_{x\in Q} d(x)$ and x^* is a solution of the problem (4.17).

In [**?**], it was proposed the following adaptive step-size with different constant c:

$$\gamma_k = \frac{D}{\sqrt{G_0^2 + \sum_{i=1}^{k-1} w_k^2}}, \quad w_k^2 := \frac{V(x_k, y_{k-1}) + V(y_k, x_k)}{c^2\gamma_i^2}, \tag{4.32}$$

where G_0 is a constant.

For the Bregmann smooth setting of the problem (4.17), the following theorem is proved in [**?**].

Theorem 4.5 *Let F be a Bregman smooth on Q. Under Assumptions **A1–A4**, after N iterations of Algorithm 5, with γ_k as in* (4.32) *and $c = 5$, we have*

$$\mathbb{E}\left[\text{DualGap}\left(\overline{x}_N\right)\right] \leq O\left(\frac{\sqrt{\log N}}{\sqrt{N}}\right).$$

Also, For the Bregmann bounded setting of the problem (4.17), the following theorem is proved in [**?**].

Theorem 4.6 *Let F be a Bregman bounded on Q. Under Assumption **A1**, after N iterations of Algorithm 5, with η_k as in* (4.32) *and $c = 1$, we have*

$$\mathbb{E}\left[\text{DualGap}\left(\overline{x}_N\right)\right] \leq O\left(\frac{\sqrt{\log N}}{\sqrt{N}}\right).$$

4.3.2 Generalized extragradient framework

It is known that when the operator F is monotone and in order to solve (4.17) if a method is run with a stochastic first-order oracle, then its convergence rate is $O\left(1/\sqrt{N}\right)$ after N oracle calls [56]. This rate is, in general, not improvable [89] without additional assumptions. The rate $O\left(1/\sqrt{N}\right)$ can be improved when the operator F is strongly monotone, i.e. there exist $\mu > 0$, such that

$$\langle F(x) - F(y), x - y\rangle \geq \mu\|x - y\|^2, \quad \forall x, y \in Q. \tag{4.33}$$

In this case, we can obtain a fast rate $O\left(1/N\right)$ with a rapidly decreasing step-size [**?**]. This acceleration requires knowledge of the strong monotonicity modulus μ, and there is no known way to adapt to it.

In [**?**], for solving monotone variational inequalities in the presence of randomness and uncertainty, it was proposed an algorithmic framework which interpolates between $O\left(1/\sqrt{N}\right)$ and $O\left(1/N\right)$ depending on the setting of the problem. It includes various popular stochastic first-order methods, such as dual averaging [88], dual extrapolation [**?**] and optimistic gradient algorithms [**?**, **?**], which allow to provide a unified framework for their analysis.

To clarify the settings of the problem under consideration, we will denote the assumptions **A1, A2, A3** as *absolutely random* [9]. On the other hand, the assumptions **A1, A2** and the following additional assumption

A5: Relative variance, i.e. there exist $c > 0$, such that

$$\mathbb{E}\left[\left\|\tilde{F}(x) - F(x)\right\|_*^2 |x\right] \leq c\|F(x)\|_*^2; \ \forall x \in Q.$$

denoted as *relatively random* [95].

It is well known that the absolutely random assumptions are typical in order to get the rate $O\left(1/\sqrt{N}\right)$ for various optimization methods (see [56] and references therein). On the other hand, the relatively random assumptions allow us to recover the well known order-optimal bound $O\left(1/N\right)$ for deterministic settings [**?**].

For problem (4.17), in [**?**], it was proposed a generalized extragradient (GEG) framework. Where by the given two sequences of dual vectors $v_k, v_{k+1/2}$, and the sequence of step-sizes $\{\gamma_k\}_{k\geq 0}$, GEG has the following form

$$\begin{aligned} x_{k+1/2} &= x_k - \gamma_k v_k, \\ y_{k+1} &= y_k - v_{k+1/2}, \\ x_{k+1} &= \gamma_{k+1} y_{k+1}. \end{aligned} \tag{4.34}$$

Under different choices of v_k and $v_{k+1/2}$, it can be write the Dual Averaging, Dual Extrapolation and Optimistic Dual Averaging algorithms in the form of GEG, as the follows.

Stochastic Dual Averaging: Let us choose

$$v_k = 0 \quad \text{and} \quad v_{k+1/2} = \tilde{F}\left(x_{k+1/2}\right).$$

Then the GEG scheme reduces to the dual averaging scheme (DA):

$$\begin{aligned} y_{k+1} &= y_k - \tilde{F}(x_k), \\ x_{k+1} &= \gamma_{k+1} y_{k+1}. \end{aligned} \tag{4.35}$$

Stochastic Dual Extrapolation: Let us choose

$$v_k = \tilde{F}(x_k) \quad \text{and} \quad v_{k+1/2} = \tilde{F}\left(x_{k+1/2}\right).$$

Then GEG yields to Nesterov's dual extrapolation method (DE):

$$\begin{aligned} x_{k+1/2} &= x_k - \gamma_k \tilde{F}\left(x_k\right), \\ y_{k+1} &= y_k - \tilde{F}\left(x_{k+1/2}\right), \\ x_{k+1} &= \gamma_{k+1} y_{k+1}. \end{aligned} \tag{4.36}$$

Stochastic Optimistic Dual Averaging: Let us choose

$$v_k = \tilde{F}\left(x_{k-1/2}\right) \quad \text{and} \quad v_{k+1/2} = \tilde{F}\left(x_{k+1/2}\right).$$

Then GEG yields to the optimistic dual averaging method (ODA):

$$\begin{aligned} x_{k+1/2} &= x_k - \gamma_k \tilde{F}\left(x_{k-1/2}\right), \\ y_{k+1} &= y_k - \tilde{F}\left(x_{k+1/2}\right), \\ x_{k+1} &= \gamma_{k+1} y_{k+1}. \end{aligned} \tag{4.37}$$

In [**?**], it was proposed and analyzed an adaptive and non-adaptive scenarios of GEG, for the absolute and relative random noise,

4.3.2.1 Non-adaptive generalized extragradient

In this subsection we mention to a series of tight convergence rates for GEG with a non-adaptive step-size sequence.

Theorem 4.7 *Let $x_k, x_{k+1/2}$ be generated by* GEG *with a decreasing step-size $\gamma_k = O\left(1/\sqrt{k}\right)$. Then under the absolutely random assumptions (i.e. **A1, A2, A3**), for $\overline{x}_N = \frac{1}{N}\sum_{i=1}^{N} x_{i+1/2}$, after N calls of oracles we have*

$$\mathbb{E}\left[\text{DualGap}\left(\overline{x}_N\right)\right] \leq O\left(\frac{1}{\sqrt{N}}\right).$$

Under the additional assumption of relative variance (assumption **A5**), it is possible to achieve an accelerated rate of $O\left(1/N\right)$. To this end, in [**?**], it was proved the following results.

Theorem 4.8 *Let $x_k, x_{k+1/2}$ be generated by* GEG *with a constant step-size $\gamma_k := \gamma,\ \forall k \geq 0$, which satisfies*

$$\min\left\{(2L)^{-1}, \left(4L^2\gamma\right)^{-1}\right\} - 2\gamma c > 0.$$

*Then under the relatively random assumptions (i.e. **A1, A2, A5**), for $\overline{x}_N = \frac{1}{N}\sum_{i=1}^{N} x_{i+1/2}$,after N calls of oracles we have*

$$\mathbb{E}\left[\text{DualGap}\left(\overline{x}_N\right)\right] \leq O\left(\frac{1}{N}\right).$$

4.3.2.2 Adaptive generalized extragradient

In this subsection we mention to a series of tight convergence rates for GEG with the following adaptive step-size sequence

$$\gamma_k = \left(1 + \sum_{i=1}^{k-1} \left\|v_i - v_{i+1/2}\right\|_*^2\right)^{-1/2}. \tag{4.38}$$

The next result gives the convergence rate of GEG with absolutely random assumptions

Theorem 4.9 *Let $x_k, x_{k+1/2}$ be generated by* GEG *with the step-size* (4.38)*. Then under the absolutely random assumptions (i.e. **A1, A2, A3**), for $\overline{x}_N = \frac{1}{N}\sum_{i=1}^{N} x_{i+1/2}$, after N calls of oracles we have*

$$\mathbb{E}\left[\text{DualGap}\left(\overline{x}_N\right)\right] \leq O\left(\frac{1}{\sqrt{N}}\right).$$

Under the additional assumption of relative variance (assumption **A5**), it is possible to achieve an accelerated rate of $O\left(1/N\right)$. To this end, in [**?**], it was proved the following results.

Theorem 4.10 *Let $x_k, x_{k+1/2}$ be generated by* GEG *with the step-size* (4.38)*. Then under the relatively random assumptions (i.e. **A1, A2, A5**), for $\overline{x}_N = \frac{1}{N}\sum_{i=1}^{N} x_{i+1/2}$, after N calls of oracles we have*

$$\mathbb{E}\left[\text{DualGap}\left(\overline{x}_N\right)\right] \leq O\left(\frac{1}{N}\right).$$

4.4 Extragradient method with line search

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\xi : \Omega \longrightarrow \Xi$ be a random variable with distribution $\mathbf{P}$. Let $Q \subset \mathbb{R}^n$ be a given closed and convex set and $F : \Xi \times Q \longrightarrow \mathbb{R}^n$ be a measurable random operator. The expected operator defined as

$$T(x) = \mathbb{E}[F(\xi, x)] = \int_{\Xi} F(\xi, x)\, \mathrm{d}\mathbf{P}(\xi), \quad (x \in X). \tag{4.39}$$

We assume that there is an access to the stochastic oracle of F, i.e. to the random operator F via samples drawn from the distribution $\mathbf{P}$.

Under stochastic oracle, a famous approach to solving stochastic variational inequalities (SVIs) is the stochastic approximation (SA) method, which was firstly proposed by Robbins and Monro [103] for the stochastic optimization problems and recently analyzed for SVIs [56, **?**, **?**, **?**, **?**]. In stochastic approximation methods, the samples are accessed in an interior and online manner along the progress of a chosen algorithm [83]. In order to construct an efficient SA method in terms of sample complexity, a rapidly growing line of search proposes SA methods with variance reduction using more than one oracle per iteration to relax the role of the stepsize in reducing variances, such as gradient aggregation methods and dynamic sampling methods [15]. These variance reduction methods use a constant stepsize policy $O\left(1/L\right)$ with assuming knowledge of the Lipschitz constant L, and they improve the convergence of stochastic approximation methods, but they are non-practical because the constant L is rarely known or it is challenging estimated [79, 83]. In [**?**],

it was focused on the problem of the construction of a robust and efficient adaptive variance reduction method, in which the stepsizes bounded away from zero, with a faster rate of convergence and near optimal oracle complexity. It was proposed a dynamic sampled stochastic approximated (DS-SA) extragradient method to solve SVIs in the large sample setting, using the stochastic approximation scheme with variance reduction and a line search scheme without requiring knowledge of the Lipschitz constant and it is near optimal oracle complexity $O\left(\varepsilon^{-2}\right)$, up to logarithmic factors on ε and L.

For any $x \in \mathbb{R}^n$, we will use the notations: $d(x, Q) = \inf_{y \in Q}\{\|x-y\|_2\}$, $\text{Proj}_Q(x) = \text{argmin}_{y \in Q} \|y - x\|_2^2$. Let $X^* \neq \emptyset$ be the set of solutions of the problem (4.17) with operator T in (4.39). For the proposed DS-SA extragradient method with a DS-SA line search scheme in [**?**], about its asymptotic convergence, it was proved that under assumptions: the operator T in (4.39) is L-Lipschitz continuous, i.e.

$$\|T(x) - T(y)\|_2 \leq L\|x - y\|_2, \quad \forall x, y \in Q,$$

and is pseudomonotone, i.e.

$$\langle T(x), y - x\rangle \geq 0 \Longrightarrow \langle T(y), y - x\rangle \geq 0, \quad \forall x, y \in Q,$$

(note that the monotonicity is a special case from pseudomonotonicity), and i.i.d. sampling of some generated samples form the distribution **P**. The proposed algorithm generates an almost surly (a.s.) bounded sequence $\{x_k\}_{k\geq 0}$, such that $\lim_{k\to\infty} d(x_k, X^*) = 0$, and

$$\lim_{k\to\infty} \left\|x_k - \text{Proj}_Q\left[x_k - T(x_k)\right]\right\|_2 = 0 \tag{4.40}$$

Also, for the rate of convergence of this proposed algorithm, it was proved the following bound for all $k \geq 1$:

$$\min_{i\in\{0,1,\ldots,k-1\}} \mathbb{E}\left[\left\|x_i - \text{Proj}_Q\left[x_i - T(x_i)\right]\right\|_2^2\right] \leq O\left(\frac{1}{k}\right) \tag{4.41}$$

Let we set $N := O(n)$, as a result from the previous mentioned results, for the DS-SA extragradient method with a DS-SA line search, after $K = c^{-1}O\left(\varepsilon^{-1}\right)$ iterations, for any $c > 0$ and a given $\varepsilon > 0$, we have

$$\min_{0\leq k\leq K} \mathbb{E}\left[\left\|x_k - \text{Proj}_Q\left[x_k - T(x_k)\right]\right\|_2^2\right] \leq \varepsilon.$$

4.5 Permutation-based stochastic gradient methods

In this section we consider only finite-sum problems

$$\min_{x\in\mathbb{R}^n}\left\{f(x)=\frac{1}{m}\sum_{i=1}^{m}f_i(x)\right\} \tag{4.42}$$

where each individual function $f_i:\mathbb{R}^n\to\mathbb{R}$ is L_i- smooth for each $i\in[m]$. In this section we explain popular and extremely efficient approach, which was used as heuristics, and provide sharp guarantees for family of methods based on *data permutation/shuffling*. Alternatively, this class of methods is called as *sampling without replacement*.

Let us consider the definition of permutations, which we will use in this section. The bijective function $\pi:\{1,\dots,m\}\to\{\pi_0,\dots,\pi_{m-1}\}$ is called *permutation or shuffling*. Note that permuted indices π_i start from π_0. It simplifies some derivations below.

We consider the class of permutation-based methods, which rely on random or deterministic shuffling of data points and perform sequential gradient steps following the permuted order. We focus our attention on three most popular methods belong to this class.

The most popular algorithm from the class of permutation-based methods is the **Random Reshuffling (RR)**. Let us describe how this method works. In the beginning of each epoch we sample set of indices $\{\pi_0,\pi_1,\dots,\pi_{m-1}\}$ without replacement from the set $\{1,\dots,m\}$ uniformly at random. We use this random permutation to perform m sequential gradient steps of the form:

$$x_{i+1}^k=x_i^k-\gamma\nabla f_{\pi_i}\left(x_i^k\right). \tag{4.43}$$

where $\gamma>0$ is a stepsize of the method. The starting iteration x^{k+1} in next epoch is updated as the last iteration of previous epoch: $x^{k+1}=x_0^{k+1}=x_n^k$. After updating we repeat m steps again for next epoch and in total we have K epochs. Note that we generate a new permutation at the beginning of each epoch, this explains why we use the term *reshuffling*.

We also consider another method, which is called **Shuffle-Once (SO)**. This method is identical to RR with only one difference that this method makes permutation only once in the beginning of optimization process and it utilizes this permutation in all epochs.

Finally, we consider deterministic algorithm **Incremental Gradient (IG)**, which is equivalent to SO with one exception that initial permutation is deterministic and not random as it is used in SO. In our analysis we provide bounds for IG in worst case scenario of all possible permutations.

4.5.1 Strong convexity assumption for individual functions

In this section we provide convergence guarantees for Random Reshuffling and Shuffle-Once methods in case of strong convexity of individual functions f_i. The standard approach of analyzing SGD method is based on the fact that all iterates

converge to neighborhood of the optimum x^* and the radius of neighborhood is characterized by level of noise. However, the proof technique for Random Reshuffling and Shuffle-Once in case of strong convexity of individual functions is significantly different and relies on the observation that each inner iterate converges to its own point. For fixed permutation π these points have the following form:

Definition 4.5 Given a permutation π, we define x_i^* as i gradient steps from point x^*:

$$x_i^* \stackrel{\text{def}}{=} x^* - \gamma \sum_{j=0}^{i-1} \nabla f_{\pi_j}(x^*), \quad i = 1, \ldots, m-1. \tag{4.44}$$

Using this new sequence allows us to introduce a better notion of variance for Random Reshuffling and Shuffle-Once algorithms:

Definition 4.6 For fixed stepsize $\gamma > 0$ and a random permutation π of $\{1, 2, \ldots, n\}$ and using definition 4.5 for the point x_i^* we can define shuffling variance as

$$\sigma_{\text{Shuffle}}^2 \stackrel{\text{def}}{=} \max_{i=1,\ldots,n-1} \left[\frac{1}{\gamma^2} \mathbb{E}\left[D_{f_{\pi_i}}\left(x_i^*, x^*\right) \right] \right] \tag{4.45}$$

Chapter 5
Stochastic linear coupling under strong growth condition

Abstract In this chapter, we ...

In this chapter, we consider

$$\min_{x \in \mathbb{R}^n} \left\{ f(x) = \mathbb{E}_{\xi \sim \mathcal{D}}[f(x, \xi)] \right\} \tag{5.1}$$

with f being L-smooth meaning that f is differentiable and for all $x, y \in \mathbb{R}^n$ its gradient is L-Lipschitz:

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2. \tag{5.2}$$

and satisfies the following assumption

Assumption 5.1 (Strong growth condition (SGC)) We say, that function f satisfies the strong growth condition, if for a stochastic subgradient $g(x, \xi)$ the following holds for any $x \in \mathbb{R}^n$ with constant η and additive error σ^2:

$$\mathbb{E}\left[\|g(x, \xi)\|_2^2\right] \leq \eta\|\nabla f(x)\|_2^2 + \sigma^2$$

5.1 Stochastic Linear coupling

Theorem 5.1 (Convergence rate of Algorithm 6) *Let $f(x)$ be convex and L-Lipschitz smooth. Let for stochastic gradient $g(x)$, the SGC holds. Then for all $N > 0$, the output of Algorithm 6 satisfies the following:*

$$\mathbb{E}[f(y^N) - f(x^*)] \leq \frac{2L\eta^2\|x^0 - x^*\|_2^2}{(N+1)^2} + \frac{\sigma^2(N+1)}{L\eta^2}. \tag{5.6}$$

Proof **Step 1** From L-smoothness of $f(x)$ and (5.4) we have:

Algorithm 6 Stochastic Linear Coupling (SLC).

Input: starting point $x^0 \in \mathbb{R}^d$, number of iterations N.

1: $y^0 = z^0 = x^0$.

2: **for** $k = 0, 1, \ldots, N-1$ **do**

3: $\alpha_{k+1} = \frac{k+2}{2L\eta^2}$ and $\tau_k = \frac{1}{\alpha_{k+1}L\eta^2} = \frac{2}{k+2}$

4:
$$x^{k+1} = \tau_k z^k + (1-\tau_k)y^k \tag{5.3}$$

5:
$$y^{k+1} = x^{k+1} - \frac{1}{L\eta}g^{k+1}(x^{k+1}) \tag{5.4}$$

6:
$$z^{k+1} = z^k - \alpha_{k+1}g^{k+1}(x^{k+1}) \tag{5.5}$$

7: **end for**

Output: y^N

$$\begin{aligned} f(y^{k+1}) &\leq f(x^{k+1}) + \langle \nabla f(x^{k+1}), y^{k+1} - x^{k+1}\rangle + \frac{L}{2}\|y^{k+1} - x^{k+1}\|_2^2 \\ &= f(x^{k+1}) - \frac{1}{L\eta}\langle \nabla f(x^{k+1}), g(x^{k+1})\rangle + \frac{1}{2L\eta^2}\|g(x^{k+1})\|_2^2. \end{aligned}$$

Let $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot \mid z^k, y^k]$. Then $\mathbb{E}_k[g(x^{k+1})] = \nabla f(x^{k+1})$. Using the SGC we obtain the following

$$\begin{aligned} \mathbb{E}_k[f(y^{k+1})] &\leq f(x^{k+1}) - \frac{1}{L\eta}\langle \nabla f(x^{k+1}), \mathbb{E}_k[g^{k+1}]\rangle + \frac{1}{2L\eta^2}\mathbb{E}_k\left[\|g(x^{k+1})\|_2^2\right] \\ &\leq f(x^{k+1}) - \frac{1}{2L\eta}\|\nabla f(x^{k+1})\|_2^2 + \frac{\sigma^2}{2L\eta^2}. \end{aligned}$$

Thus, rearranging the terms we get the following

$$\|\nabla f(x^{k+1})\|_2^2 \leq 2L\eta\left(f(x^{k+1}) - \mathbb{E}_k[f(y^{k+1})]\right) + \frac{\sigma^2}{\eta} \tag{5.7}$$

Step 2

From (5.5) it holds

$$\begin{aligned} \alpha_{k+1}\langle g(x^{k+1}), z^k - x^*\rangle = -\langle z^{k+1} - z^k, z^k - x^*\rangle &= \frac{1}{2}\|z^{k+1} - z^k\|_2^2 + \frac{1}{2}\|z^k - x^*\|_2^2 \\ &\quad - \frac{1}{2}\|z^{k+1} - z^*\|_2^2, \end{aligned}$$

Next we use (5.5) and $-\langle a, b\rangle = \frac{1}{2}\|a\|_2^2 + \frac{1}{2}\|b\|_2^2 - \frac{1}{2}\|a+b\|_2^2$ for all $a, b \in \mathbb{R}^n$ and get

$$\alpha_{k+1}\langle g(x^{k+1}), z^k - x^*\rangle = \frac{\alpha_{k+1}^2}{2}\|g(x^{k+1})\|_2^2 + \frac{1}{2}\|z^k - x^*\|_2^2 - \frac{1}{2}\|z^{k+1} - x^*\|_2^2.$$

Taking the conditional expectation $\mathbb{E}_k[\cdot]$ we get

$$\alpha_{k+1}\langle \nabla f(x^{k+1}), z^k - x^*\rangle \leq \frac{\alpha_{k+1}^2\eta}{2}\|\nabla f(x^{k+1})\|_2^2 + \frac{\alpha_{k+1}^2\sigma^2}{2} + \frac{1}{2}\|z^k - x^*\|_2^2 - \frac{1}{2}\mathbb{E}_k\left[\|z^{k+1} - x^*\|_2^2\right].$$

Using (5.7) we obtain

$$\alpha_{k+1}\langle \nabla f(x^{k+1}), z^k - x^*\rangle \leq L\eta^2\alpha_{k+1}^2\left(f(x^{k+1}) - \mathbb{E}_k[f(y^{k+1})]\right) + \alpha_{k+1}^2\sigma^2 + \frac{1}{2}\|z^k - x^*\|_2^2 - \frac{1}{2}\mathbb{E}_k\left[\|z^{k+1} - x^*\|_2^2\right]. \tag{5.8}$$

Step 3
Now we consider

$$\begin{aligned}\alpha_{k+1}\left(f(x^{k+1}) - f(x^*)\right) &\leq \alpha_{k+1}\langle \nabla f(x^{k+1}), x^{k+1} - x^*\rangle \\ &= \alpha_{k+1}\langle \nabla f(x^{k+1}), x^{k+1} - z^k\rangle + \alpha_{k+1}\langle \nabla f(x^{k+1}), z^k - x^*\rangle \\ &\overset{(5.8)}{\leq} \frac{(1-\tau_k)\alpha_{k+1}}{\tau_k}\langle \nabla f(x^{k+1}), y^k - x^{k+1}\rangle \\ &+ L\eta^2\alpha_{k+1}^2\left(f(x^{k+1}) - \mathbb{E}_k[f(y^{k+1})]\right) + \alpha_{k+1}^2\sigma^2 \\ &+ \frac{1}{2}\|z^k - x^*\|_2^2 - \frac{1}{2}\mathbb{E}_k\left[\|z^{k+1} - x^*\|_2^2\right],\end{aligned}$$

where we used (5.3). Using $\tau_k = \frac{1}{\alpha_{k+1}L\eta^2}$ and the convexity of f we get

$$\begin{aligned}\alpha_{k+1}\left(f(x^{k+1}) - f(x^*)\right) &\leq (\alpha_{k+1}^2 L\eta^2 - \alpha_{k+1})(f(y^k) - f(x^{k+1})) \\ &+ L\eta^2\alpha_{k+1}^2\left(f(x^{k+1}) - \mathbb{E}_k[f(y^{k+1})]\right) + \alpha_{k+1}^2\sigma^2 \\ &+ \frac{1}{2}\|z^k - x^*\|_2^2 - \frac{1}{2}\mathbb{E}_k\left[\|z^{k+1} - x^*\|_2^2\right].\end{aligned}$$

We rewrite this as follows

$$\begin{aligned}\alpha_{k+1}f(x^*) + \alpha_{k+1}^2\sigma^2 &\geq \alpha_{k+1}^2 L\eta^2\mathbb{E}_k[f(y^{k+1})] - (\alpha_{k+1}^2 L\eta^2 - \alpha_{k+1})f(y^k) \\ &+ \frac{1}{2}\mathbb{E}_k\left[\|z^{k+1} - x^*\|_2^2\right] - \frac{1}{2}\|z^k - x^*\|_2^2.\end{aligned}$$

Step 4
Taking the expectation from the inequality from Step 3 and summing for $k = 0, 1, \ldots, N-1$, we get

$$\sum_{k=0}^{N-1}\left(f(x^*)\alpha_{k+1}+\sigma^2\alpha_{k+1}^2\right) \geq \sum_{k=0}^{N-1}\alpha_{k+1}^2 L\eta^2\mathbb{E}[f(y^{k+1})] - \sum_{k=0}^{N-1}(\alpha_{k+1}^2 L\eta^2-\alpha_{k+1})\mathbb{E}[f(y^k)] + \underbrace{\sum_{k=0}^{N-1}\left(\frac{1}{2}\mathbb{E}\left[\|z^{k+1}-x^*\|_2^2\right]-\frac{1}{2}\mathbb{E}\left[\|z^k-x^*\|_2^2\right]\right)}_{\frac{1}{2}\mathbb{E}\left[\|z^N-x^*\|_2^2\right]-\frac{1}{2}\|x^0-x^*\|_2^2}$$

Using $\alpha_{k+1}=\frac{k+2}{2L\eta^2}$, we have

$$\alpha_{k+1}^2 L\eta^2-\alpha_{k+1}-\alpha_k^2 L\eta^2 = \frac{k^2+4k+4}{4L\eta^2}-\frac{k+2}{2L\eta^2}-\frac{k^2+2k+1}{4L\eta^2} = -\frac{1}{4L\eta^2}.$$

Thus, we get the following

$$\sum_{k=0}^{N-1}\left(f(x^*)\alpha_{k+1}+\sigma^2\alpha_{k+1}^2\right) \geq \underbrace{\sum_{k=0}^{N-1}\left(\alpha_{k+1}^2 L\eta^2\mathbb{E}[f(y^{k+1})]-\alpha_k^2 L\eta^2\mathbb{E}[f(y^k)]\right)}_{\alpha_N^2 L\eta^2\mathbb{E}[f(y^N)]-\frac{1}{4L\eta^2}f(y^0)} + \frac{1}{4L\eta^2}\sum_{k=0}^{N-1}\mathbb{E}[f(y^k)] + \frac{1}{2}\mathbb{E}\left[\|z^N-x^*\|_2^2\right]-\frac{1}{2}\|x^0-x^*\|_2^2.$$

Then we use α_{k+1} to estimate the summ in the L.H.S

$$\sum_{k=0}^{N-1}\alpha_{k+1} = \frac{1}{2L\eta^2}\sum_{k=0}^{N-1}(k+2) = \frac{N(N+3)}{4L\eta^2},$$
$$\sum_{k=0}^{N-1}\alpha_{k+1}^2 = \frac{1}{4L^2\eta^4}\sum_{k=0}^{N-1}(k+2)^2 \leq \frac{(N+1)^3}{4L^2\eta^4}$$

and get

$$\frac{N(N+3)}{4L\eta^2}f(x^*)+\frac{\sigma^2(N+1)^3}{4L^2\eta^4} \geq \frac{(N+1)^2}{4L\eta^2}\mathbb{E}[f(y^N)]+\frac{1}{4L\eta^2}\sum_{k=1}^{N-1}\mathbb{E}\underbrace{[f(y^k)]}_{\geq f(x^*)} + \frac{1}{2}\underbrace{\mathbb{E}\left[\|z^N-x^*\|_2^2\right]}_{\geq 0}-\frac{1}{2}\|x^0-x^*\|_2^2.$$

Then we rewrite this as follows

$$\frac{(N+1)^2}{4L\eta^2}\mathbb{E}[f(y^N)] \leq \left(\frac{N(N+3)}{4L\eta^2} - \frac{N-1}{4L\eta^2}\right) f(x^*) + \frac{\|x^0 - x^*\|_2^2}{2} + \frac{\sigma^2(N+1)^3}{4L^2\eta^4}$$
$$= \frac{(N+1)^2}{4L\eta^2} f(x^*) + \frac{\|x^0 - x^*\|_2^2}{2} + \frac{\sigma^2(N+1)^3}{4L^2\eta^4}.$$

Rearranging the terms we get the desired statement

$$\mathbb{E}[f(y^N)] - f(x^*) \leq \frac{2L\eta^2\|x^0 - x^*\|_2^2}{(N+1)^2} + \frac{\sigma^2(N+1)}{L\eta^2}.$$

5.2 Gradient-free optimization

In this section, we suppose that instead of first-order oracle we are given zeroth-order oracle. Moreover, we suppose objective $f(x, \xi)$ can be observed trough its noisy approximation

$$\varphi(x, \xi) \triangleq f(x, \xi) + \delta(x). \tag{5.9}$$

Now stochastic gradient $g(x, \xi)$ can be approximated by by the function evaluations in two random points closed to x variance [113]:

$$g(x, \xi, e) = \frac{n}{2\tau}\left(\varphi(x + \tau e, \xi) - \varphi(x - \tau e, \xi)\right) e \tag{5.10}$$

where vector e is picked uniformly at random from the Euclidean unit sphere $\{e : \|e\|_2 = 1\}$, and $\tau > 0$ is some constant. Next, we will show that stochastic gradient estimation $g(x, \xi, e)$ satisfies the SGC (Assuption 5.1).

Now we introduce the following smooth approximation for $f(\cdot)$

$$f^\tau(x) \triangleq \mathbb{E}_{\tilde{e}} f(x + \tau\tilde{e}), \tag{5.11}$$

where $\tau > 0$ and $\tilde{e}$ is a vector picked uniformly at random from the Euclidean unit ball: $\{\tilde{e} : \|\tilde{e}\|_2 \leq 1\}$. Function $f^\tau(x)$ can be referred as a smooth approximation of $f(x)$ and it will be used only for deriving the convergence rate of proposed algorithm. Here $f(x) \triangleq \mathbb{E} f(x, \xi)$.

Assumption 5.2 (Boundedness of the noise) For all $x \in \mathcal{X}$, it holds $|\delta(x)| \leq \Delta$.

Lemma 5.1 *For $g(x, \xi, e)$ from* (5.10)*, the following holds Assumption 5.2*

$$\mathbb{E}_{\xi,e}\left[\|g(x, \xi, e)\|_q^2\right] \leq \sqrt{3}\|\nabla f(x)\|_2^2 \min\{2q - 1, 32\ln n - 8\} n^{\frac{2}{q}}$$
$$+ 3n^2L^2\tau^2\mathbb{E}\left[\|e\|_q^2\right] + \frac{n^2\Delta^2}{2\tau^2}\mathbb{E}\left[\|e\|_q^2\right].$$

Proof Let us consider

$$\begin{aligned}\mathbb{E}_{\xi,e}\left[\|g(x,\xi,e)\|_q^2\right] &= \mathbb{E}_{\xi,e}\left[\left\|\frac{n}{2\tau}\left(\varphi(x+\tau e,\xi)-\varphi(x-\tau e,\xi)\right)e\right\|_q^2\right]\\ &= \frac{n^2}{4\tau^2}\mathbb{E}_{\xi,e}\left[\|e\|_q^2\left(f(x+\tau e,\xi)+\delta(x+\tau e)-f(x-\tau e,\xi)-\delta(x-\tau e)\right)^2\right]\\ &= \frac{n^2}{4\tau^2}\mathbb{E}_{\xi,e}\left[\|e\|_q^2\left(f(x+\tau e,\xi)-f(x-\tau e,\xi)+\langle\nabla f(x-\tau e,\xi),2\tau e\rangle\right.\right.\\ &\quad\left.\left.-\langle\nabla f(x-\tau e,\xi),2\tau e\rangle+\delta(x+\tau e)-\delta(x-\tau e)\right)^2\right]. \end{aligned} \tag{5.12}$$

Then we use that $(a+b+c)^2 \le 3a^2+3b^2+3c^2$ for (5.12) and obtain

$$\begin{aligned}&\mathbb{E}_{\xi,e}\left[\|g(x,\xi,e)\|_q^2\right]\\ &\quad\le \frac{3n^2}{4\tau^2}\mathbb{E}_{\xi,e}\left[\|e\|_q^2\left(f(x+\tau e,\xi)-f(x-\tau e,\xi)-\langle\nabla f(x-\tau e,\xi),2\tau e\rangle\right)^2\right]\\ &\quad+\frac{3n^2}{4\tau^2}\mathbb{E}_{\xi,e}\left[\|e\|_q^2\langle\nabla f(x-\tau e,\xi),2\tau e\rangle^2\right]\\ &\quad+\frac{n^2}{4\tau^2}\mathbb{E}_{\xi,e}\left[\|e\|_q^2\left(\delta(x+\tau e)-\delta(x-\tau e)\right)^2\right]. \end{aligned} \tag{5.13}$$

Now for the first term in the R.H.S. of (5.13) we use $f(x)$ is L-Lipschitz smooth

$$\begin{aligned}&\frac{3n^2}{4\tau^2}\mathbb{E}_{\xi,e}\left[\|e\|_q^2\left(f(x+\tau e,\xi)-f(x-\tau e,\xi)-\langle\nabla f(x-\tau e,\xi),2\tau e\rangle\right)^2\right]\\ &\qquad\le \frac{3n^2}{4\tau^2}\mathbb{E}\left[\|e\|_q^2\right]\cdot(2L\tau^2)^2\mathbb{E}\left[\|e\|_2^2\right]=3n^2L^2\tau^2\mathbb{E}\left[\|e\|_q^2\right]. \end{aligned} \tag{5.14}$$

For the second term in the R.H.S. of (5.13) we have from [47]

$$\frac{3n^2}{4\tau^2}\mathbb{E}_{\xi,e}\left[\|e\|_q^2\langle\nabla f(x-\tau e,\xi),2\tau e\rangle^2\right]\le 3\sqrt{3}\|\nabla f(x)\|_2^2\min\{2q-1,32\ln n-8\}n^{\frac{2}{q}}. \tag{5.15}$$

Then for the third term in the R.H.S. of (5.13) we use Assumption 5.2 and get the following

$$\frac{n^2}{4\tau^2}\mathbb{E}_e\left[\|e\|_q^2(\delta(x+\tau e)-\delta(x-\tau e))^2\right]\le\frac{n^2\Delta^2}{\tau^2}\mathbb{E}\left[\|e\|_q^2\right] \tag{5.16}$$

Next we use (5.14), (5.15) and (5.16) for (5.13) and get the statement of the lemma.

$$\begin{aligned}\mathbb{E}_{\xi,e}\left[\|g(x,\xi,e)\|_q^2\right] &\le \sqrt{3}\|\nabla f(x)\|_2^2\min\{2q-1,32\ln n-8\}n^{\frac{2}{q}}\\ &\quad+3n^2L^2\tau^2\mathbb{E}\left[\|e\|_q^2\right]+\frac{n^2\Delta^2}{\tau^2}\mathbb{E}\left[\|e\|_q^2\right].\end{aligned}$$

Lemma 5.1 for the Euclidean case (q=2)

For the Euclidean case, Lemma 5.1 can be simplified

$$\mathbb{E}_{\xi,e}\left[\|g(x,\xi,e)\|_2^2\right] \leq 3\sqrt{3}n\|\nabla f(x)\|_2^2 + 3n^2L^2\tau^2 + \frac{n^2\Delta^2}{\tau^2}.$$

Thus, for the Euclidean case, Lemma (5.1) implies that Assumption 5.1 (SGC) holds with $\eta = 3\sqrt{3}n$ and $\sigma^2 = 3n^2L^2\tau^2 + \frac{n^2\Delta^2}{\tau^2}$.

5.3 Component linear coupling

Appendix A
Concentration inequalities

Concentration inequalities are fundamental tools in probabilistic combinatorics and theoretical computer science for proving that random functions are near their means.

A.1 Azuma–Hoeffding inequality

A.2 Bernstein Inequality

A.3 McDiarmid's inequality

McDiarmid's inequality (or bounded differences inequality) [**?**] is one of the concentration inequalities, which provide bounds on how a random variable deviates from its expected value. It shows how the values of a bounded function of independent random variables concentrate about its mean.

McDiarmid's inequality states that if $X_1, \ldots, X_n$ are given independent random variables in some measurable space $\mathcal{X}$, and let $f : \mathcal{X}^n \to \mathbb{R}$ be a function of $X_1, \ldots, X_n$, for all $i = 1, \ldots, n$, there is $c_i \geq 0$, such that

$$|f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x', x_{i+1}, \ldots, x_n)| \leqslant c_i,$$

for all $x_1, \ldots, x_n, x' \in \mathcal{X}$. Then

$$\mathbb{P}\left(|f(X_1, \ldots, X_n) - \mathbb{E}f(X_1, \ldots, X_n)| \geq \epsilon\right) \leqslant 2\exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right). \tag{A.1}$$

McDiarmid inequality is a generalization of Hoeffding's inequality, which can be obtained by assuming $\mathcal{X} = [a, b]$ and choosing $f(X_1, \ldots, X_n) = \sum_{i=1}^n X_i$.

Appendix B
Main Results of Convex Analysis and Convex Optimization

B.1 Convex Analysis Tools

Let $(\mathbf{E}, \|\cdot\|)$ be a normed finite-dimensional vector space, with an arbitrary norm $\|\cdot\|$, and $\mathbf{E}^*$ be the conjugate space of $\mathbf{E}$ with the following norm

$$\|y\|_* = \max_{x \in \mathbf{E}} \{\langle y, x\rangle : \|x\| \leq 1\},$$

where $\langle y, x\rangle$ is the value of the continuous linear functional $y \in \mathbf{E}^*$ at $x \in \mathbf{E}$.

If $\mathbf{E} = \mathbb{R}^n$, then for $1 \leq p < \infty$, the most popular norms are so-called l_p-norms

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}, \quad x = (x_1, \ldots, x_n) \in \mathbb{R}^n.$$

Among them, there are three norms, which are commonly used

- Euclidean norm: $\|x\|_2 = \left(\sum_{i=1}^{n} x_i^2\right)^{1/2}$.
- l_1-norm: $\|x\|_1 = \sum\limits_{i=1}^{n} |x_i|$.
- l_∞-norm (it is also called Chebyshev norm): $\|x\|_\infty = \max\limits_{1 \leq i \leq n} |x_i|$.

B.1.1 Convex sets

Definition B.1 A set $Q \subseteq \mathbf{E}$ is convex if $\lambda x + (1 - \lambda) y \in Q$ for any $x, y \in Q$ and $\lambda \in [0, 1]$.

Definition B.2 Let $Q \subset \mathbf{E}$ be a closed convex set. For any $y \in \mathbf{E}$, we define the closest point to y in Q as

$$\mathrm{Proj}_Q(y) = \arg\min_{x \in Q} \|y - x\|.$$

$\text{Proj}_Q(y)$ is called the projection of y onto Q.

Proposition B.1 *Let $Q \subset \mathbf{E}$ be a closed convex set and $y \in \mathbf{E}$ be given. Then the* $\text{Proj}_Q(y)$ *must exist and unique.*

Remark B.1 In many cases when the set Q is relatively simple, we can compute $\text{Proj}_Q(y)$ explicitly. Let us see the following two examples

- Let $H := \left\{x \in \mathbf{E} : a^T x + b = 0\right\}$, where $a \in \mathbf{E}$ and $b \in \mathbb{R}$, be a given hyperplane. Then the projection of any $y \in \mathbf{E}$ onto H is given by

$$\text{Proj}_H(y) = y - \frac{(a^T y + b)a}{\|a\|^2}.$$

- Let Q be the unit ball. Then the projection of any $y \in \mathbf{E}$ onto Q is given by

$$\text{Proj}_Q(y) = \frac{y}{\|y\|}.$$

B.1.2 Differentiable convex functions

Let f be a function. By $\text{dom } f = \{x \in \mathbf{E} : |f(x)| < \infty\}$ we denote the domain of the function f . We always assume that f is proper, i.e. $\text{dom } f \neq \emptyset$.

Definition B.3 Let $Q \subseteq \mathbf{E}$ be a nonempty convex set and $f : Q \to \mathbb{R}$ be a given function. The function f is called *convex* on the set Q if

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) \quad \forall\, x, y \in Q \text{ and } \lambda \in (0,1). \tag{B.1}$$

A function f is called *strictly convex* if the inequality (B.1) is strict, whenever $x \neq y$ and $\lambda \in (0,1)$. We say that f is *concave* if $-f$ is convex. Similarly f is *strictly concave* if $-f$ is strictly convex. Moreover, let $\mu > 0$, the function f is called *μ–strongly convex* if and only if

$$f(\lambda x + (1-\lambda)y) + \lambda(1-\lambda)\frac{\mu}{2}\|x-y\|^2 \leq \lambda f(x) + (1-\lambda) f(y), \tag{B.2}$$

for all $x, y \in Q$ and $\lambda \in (0,1)$. The parameter μ is called the *strong convexity parameter* of f. Note that, the convex function is strongly convex with parameter $\mu = 0$.

When the function f is differentiable, then the Definition B.3 is equivalent to the following definition

Definition B.4 A continuously differentiable function $f : Q \to \mathbb{R}$ is called *convex* on the convex set Q if

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle \quad \forall\, x, y \in Q, \tag{B.3}$$

where $\nabla f(x)$ denotes the gradient of f at the point x.

Similar to the Definition B.4, f will be *strictly convex* if the inequality (B.3) is strict, f is *concave* if the inequality (B.3) is reversed, and f is *strictly concave* if the reverse inequality is strict.

Moreover, a continuously differentiable function f is called *μ–strongly convex* with the strong convexity parameter $\mu > 0$ if and only if

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2 \quad \forall\, x, y \in Q. \tag{B.4}$$

Theorem B.1 *Let $Q \subseteq \mathbf{E}$ be an open set. A twice continuously differentiable function $f : Q \to \mathbb{R}$ is convex if and only its Hessian is positive semidefinite, i.e.,*

$$\nabla^2 f(x) \geq 0 \quad \forall\, x \in Q.$$

Analogously, we say that, the twice continuously differentiable function f is *strictly convex* if its Hessian is positive definite, *concave* if its Hessian is negative semidefinite, and *strictly concave* if its Hessian is negative definite.

B.1.3 Non-differentiable convex functions

Note that the convex functions are not always differentiable everywhere over their domain. For the non-differentiable functions, there is an important notion, which is called *subgradients*. This notion is the generalization of the gradients for differentiable functions.

Definition B.5 A vector $g \in \mathbf{E}^*$ is called a *subgradient* of the function f at the point $x \in \operatorname{dom} f$ if for any $y \in \operatorname{dom} f$ we have

$$f(y) \geq f(x) + \langle g, y - x \rangle. \tag{B.5}$$

The set of all subgradients of f at x is called the *subdifferential* of the function f at the point x and is denoted by $\partial f(x)$.

Remark B.2 In the case when the function f is non-differentiable, the strongly convex condition (B.2) is equivalent to the following

$$f(y) \geq f(x) + \langle g, y - x \rangle + \frac{\mu}{2}\|y - x\|^2 \quad \forall\, x, y \in Q. \tag{B.6}$$

where $g \in \partial f(x)$ is any subgradient of f at x.

Let $f_1, f_2 : Q \to \mathbb{R}$ be strongly convex functions with parameters μ_1, μ_2, respectively. Then the function $\alpha_1 f_1 + \alpha_2 f_2$, for any $\alpha_1, \alpha_2 \geq 0$ is $(\alpha_1\mu_1 + \alpha_2\mu_2)$-strongly convex, and the function $f(x) = \max\{f_1, f_2\}$ is $(\min\{\mu_1, \mu_2\})$-strongly convex.

B.1.4 Lipschitz continuity

Definition B.6 Let $Q \subseteq \mathbf{E}$ be an open convex set and $f : Q \to \mathbb{R}$. Then f is called *Lipschitz-continuous* if there exists a constant $M > 0$, such that

$$|f(x) - f(y)| \leq M\|x - y\| \quad \forall\, x, y \in Q. \tag{B.7}$$

This means that at every point $x \in Q$, there is a subgradient $g \in \partial f(x)$, such that $\|g\|_* \leq M$.

Definition B.7 Let $Q \subseteq \mathbf{E}$ be an open convex set and $f : Q \to \mathbb{R}$ is differentiable function in Q, we say that f *has a Lipschitz continuous gradient* (or *the gradient of f is Lipschitz-continuous* or f is *L–smooth*) if there exists a constant $L > 0$, such that

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\| \quad \forall\, x, y \in Q. \tag{B.8}$$

Theorem B.2 *Let $Q \subseteq \mathbf{E}$ be an open convex set and $f : Q \to \mathbb{R}$ is L–smooth convex function, then $\forall x, y \in Q$*

$$0 \leq f(y) - f(x) - \langle \nabla f(x), y - x\rangle \leq \frac{L}{2}\|x - y\|^2, \tag{B.9}$$

$$\|\nabla f(x) - \nabla f(y)\|^2 \leq 2L \underbrace{\left(f(x) - f(y) - \langle \nabla f(y), x - y\rangle\right)}_{V_f(x,y)}. \tag{B.10}$$

B.2 Convex Optimization Tools

Let $(\mathbf{E}, \|\cdot\|)$ be a normed finite-dimensional vector space. Let we consider the following constrained optimization problem

$$\min_{x \in Q} f(x), \tag{B.11}$$

For this problem, we have the following concepts

- The dimension of the problem is defined by the dimension of $\mathbf{E}$.
- A feasible solution is any point x that belongs to the feasible set Q.
- For a given accuracy $\varepsilon > 0$. An ε–solution of the problem (B.11), is the point $x^* \in Q$ (which is called a minimizer of the function f), such that $f(x) - f(x^*) \leq \varepsilon$, for any $x \in Q$. Note that, x^* is not necessarily unique.
- A global optimal solution is a feasible solution x^* such that

$$f(x^*) \leq f(x) \quad \forall\, x \in Q.$$

- A local optimal solution is a feasible solution x^* for which there exists $r > 0$ such that
$$f(x^*) \leq f(x) \quad \forall\, x \in \{x \in \mathbf{E} : \|x - x^*\| < r\} \cap Q.$$
- The optimal value of the objective function f is defined as $f^* = \min_{x \in Q} f(x)$.
- The set of optimal solutions is defined as $X_* = \{x^* \in Q : f(x^*) = f^*\}$.

The problem (B.11) is very general and covers very different types of objective functions and feasible sets. Unfortunately, solving the optimization problem (B.11) is a big challenge. For a majority of optimization problems, there is no hope to find a solution analytically (i.e. find a closed-form to an optimal solution). In general, we cannot guarantee whether one can find an optimal solution, and if so, how much computational effort one needs.

However, it turns out that we can provide such guarantees for a special class of problems, namely convex optimization problems, where the set Q is convex and the function f is also convex.

B.2.1 Properties of convex optimization problems

In this subsection we mention to a review of some basic and fundamental properties of convex optimization problem (B.11).

Theorem B.3 *Let f be a convex function on a convex set Q and let $x^* \in Q \cap \operatorname{dom} f$ be a local minimizer of f on Q. Then x^* is a global minimizer of f on Q. Moreover, the set X_* of all minimizers of f on Q is convex.*

When the objective function f is strongly convex, then we have the following theorem

Theorem B.4 *If f is strongly convex, then the optimal solution set X_* is either empty or a singleton.*

Concerning the existence of an optimal solution, we have the following theorem

Theorem B.5 *Let f be a convex function on a closed convex set Q and the set $\{x \in Q : f(x) \leq \alpha\}$ is nonempty and compact, for some $\alpha \in \mathbb{R}$. Then the optimal solution set X_* is nonempty, compact and convex.*

In particular, we can guarantee the existence and uniqueness of an optimal solution, when f is a differentiable strongly convex function and Q is a closed convex set.

Another important property of convex problems is the existence of necessary and sufficient optimality conditions, which we can mention in the following theorem

Theorem B.6 *Assume that the problem* (B.11) *is unconstrained, i.e. $Q = \mathbf{E}$ and that f is convex and differentiable. Then the point $x^* \in \mathbf{E}$ is an optimal solution of* (B.11) *if and only if $\nabla f(x^*) = \boldsymbol{0}$.*

When the problem is constrained, i.e. $Q \neq \mathbf{E}$, the necessary and sufficient optimality condition in Theorem B.6 becomes

Theorem B.7 *Let f be a differentiable convex function and Q be a closed convex set. The point $x^* \in Q$ is an optimal solution of* (B.11) *if and only if*

$$\langle \nabla f(x^*), x - x^* \rangle \geq 0 \quad \forall\, x \in Q. \tag{B.12}$$

For non-smooth problems, i.e. when the objective function f is non-differentiable, we cannot use the gradient in the optimality condition (Theorems B.6 and B.7). The equivalent results can be obtained by replacing the gradient ∇f with an arbitrary subgradient of f.

More precisely, the analogue of Theorem B.6 has the following form

Theorem B.8 *Assume that $Q = \mathbf{E}$ and that f is convex. Then $x^* \in \mathbf{E}$ is an optimal solution of* (B.11) *if and only if $\mathbf{0} \in \partial f(x^*)$.*

In the case of constrained problems, the analogue of Theorem B.7 has the following form

Theorem B.9 *Let f be a convex function and Q be a closed convex set. The point $x^* \in Q$ is an optimal solution of* (B.11) *if and only if there exists $g \in \partial f(x^*)$, such that*

$$\langle g, x - x^* \rangle \geq 0, \quad \forall\, x \in Q. \tag{B.13}$$

B.3 Numerical methods for convex optimization problems

B.3.1 The concept of black-box

As a rule, in mathematical optimization, numerical methods are developed for solving many problems of the same type with similar characteristics (problems of the same class). Therefore, the effectiveness of the method on some classes of problems can be considered an important characteristic of the quality of the method.

The information about the problem, which is known in advance to the numerical method, is called the *model* of the problem being solved. The efficiency (therefore the optimality) of the method on the class of problems will be understood in the sense of the number of calls to the oracle [**?**], where the *oracle* is understood as a unit that answers the successive questions of the methods. In depending on the level of the available informations about the model of the problem under consideration we have: *zero-order oracle*, which returns the value of the objective function, *first-order oracle* which returns the value of the objective function and its subgradient.

The number of calls of the oracle which is necessary to solve an optimization problem up to accuracy ε is called the complexity (more precisely it is called the analytical complexity) of the problem. There is one standard assumption on the oracle which allows us to obtain the majority of results on analytical complexity for

optimization schemes. This assumption, called the *Local Black Box* Concept, which is [89]

1. The only information available for the numerical scheme is the answer of the oracle.
2. The oracle is local, that is a small variation of the problem far enough from the test point x, which is compatible with the description of the problem class, does not change the answer at x.

This concept is one of the most useful inventions in numerical analysis. In particular, it allows one to obtain the corresponding upper and lower bounds for the oracle complexity of solving optimization problems for various sub-classes of convex functions.

Let x_* be a solution of (B.11) and $\varepsilon > 0$ be a given accuracy parameter. We call $\hat{x}$ an ε-solution of (B.11) if $f(\hat{x}) - f(x_*) \leq \varepsilon$. In 1983, Nemirovsky and Yudin in their monograph [**?**], derived the optimal worst-case complexities of first-order methods for several classes of convex problems. If a first-order method attains the worst-case complexity of a class of problems, it is called **optimal** (see Table B.1). A special feature of these methods is that the corresponding complexity does not depend explicitly on the problem dimension n.

Table B.1: Optimal complexities of the first-order methods for several classes of problems, with $N \leq n$, where N is the number of calls of the oracle, n is the dimension of the problem and R is the distance between the initial point and exact solution of the problem.

	f is M-Lipschitz	∇f is L-Lipschitz
f is convex	$O\left(\frac{M^2R^2}{\varepsilon^2}\right)$	$O\left(\sqrt{\frac{LR^2}{\varepsilon}}\right)$
f is μ-strongly convex	$O\left(\frac{M^2}{\mu\varepsilon}\right)$	$O\left(\sqrt{\frac{L}{\mu}}\ln\left(\frac{\mu R^2}{\varepsilon}\right)\right)$, $\forall N$

Moreover, the lower bounds in the non-smooth case (f is Lipschitz) are achieved using *subgradient methods*, and in the smooth case (f has a Lipschitz gradient) are achieved using the *fast accelerated gradient method* proposed by Yu. Nesterov in 1983 [**?**].

B.3.2 Convex optimization methods for lower dimensional problems

For deterministic low-dimensional minimization problems, cutting plane (or center of gravity type) methods are arguably most efficient as they achieve linear convergence rate while imposing very mild assumptions [**?**]. Prominent examples of such methods are the Vaidya's cutting plane method [**?**, **?**] and ellipsoid method [95].

In this section, we review some fundamental results of these methods, for the lower dimensional problems.

B.3.2.1 Vaidya's cutting plane method

Let we focus on the problem (B.11), where $Q \subset \mathbb{R}^n$ is a convex compact set with a nonempty interior and the function $f : Q \longrightarrow \mathbb{R}$ is continuous and convex. Let $P = \{x \in \mathbb{R}^n : \ Ax \geq b\}$ be a bounded n-dimensional polyhedron, where $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. The logarithmic barrier of the set P is defined as

$$\mathrm{Barr}(x) = -\sum_{i=1}^{m} \ln\left(a_i^\top x - b_i\right),$$

where $a_i^\top$ is the *ith* row of the matrix A. The Hessian $H(x)$ of the function $\mathrm{Barr}(x)$ is

$$H(x) = \sum_{i=1}^{m} \frac{a_i a_i^\top}{\left(a_i^\top x - b_i\right)^2}.$$

The matrix $H(x)$ is positive definite for all x in the interior of P. The volumetric barrier $\mathcal{V}$ for P is defined as

$$\mathcal{V}(x) = \frac{1}{2}\ln(\det(H(x))),$$

where $\det(H(x))$ designates the determinant of $H(x)$. The point of minimum of the function $\mathcal{V}$ on P will be referred to as the volumetric center of the set P. Denote

$$\sigma_i(x) = \frac{a_i^\top (H(x))^{-1} a_i}{\left(a_i^\top x - b_i\right)^2}, \quad 1 \leq i \leq m. \tag{B.14}$$

Then the gradient of the volumetric barrier $\mathcal{V}$ can be written as

$$\nabla\mathcal{V}(x) = -\sum_{i=1}^{m} \sigma_i(x) \frac{a_i}{a_i^\top x - b_i}.$$

Let $Q(x)$ be defined as

$$Q(x) = \sum_{i=1}^{m} \sigma_i(x) \frac{a_i a_i^\top}{\left(a_i^\top x - b_i\right)^2}.$$

Note that $Q(x)$ is positive definite on the interior of P and also that $Q(x)$ is a good approximation to the Hessian of the function $\mathcal{V}(x)$; i.e., $\nabla^2\mathcal{V}(x)$.

Vaidya's method generates a sequence of pairs $(A_k, b_k) \in \mathbb{R}^{m\times n}\times\mathbb{R}^m$ such that the corresponding polyhedron contain the solution. For the initial polyhedron, defined by the pair (A_0, b_0), one usually takes a simplex (the algorithm can start from any convex bounded n-dimensional polyhedron that easily yields to the calculation of its volumetric center, for example, from the n-rectangle).

Let $x_k (k \geq 0)$ denote the volumetric center of the polyhedron defined by the pair (A_k, b_k), and suppose that the quantities $\{\sigma_i(x_k)\}_{1\leq i\leq m}$ have been calculated for

this polyhedron, by (B.14). The next polyhedron (A_{k+1}, b_{k+1}) is obtained from the current one as a result of either joining or removing a constraint:

1. If for some $i \in \{1, \ldots, m\}$ one has $\sigma_i(x_k) = \min_{1 \le j \le m} \sigma_j(x_k) < \gamma$, then (A_{k+1}, b_{k+1}) is obtained by eliminating the *ith* row from (A_k, b_k).
2. Otherwise, the oracle called up at the current point x_k returns a vector c_k such that $f(x) \leqslant f(x_k)\, \forall x \in \left\{z \in Q_x : c_k^\top z \ge c_k^\top x_k\right\}$; i.e., $c_k \in -\partial f(x_k)$. Select, $\beta_k \in \mathbb{R}$ such that
$$\frac{c_k^\top \left(H(x_k)\right)^{-1} c_k}{\left(x_k^\top c_k - \beta_k\right)^2} = \frac{\sqrt{\gamma}}{5}.$$
Determine (A_{k+1}, b_{k+1}) by adding the row (c_k, β_k) to (A_k, b_k).

The volumetric barrier $\mathcal{V}_k$ is a self-concordant function; therefore, it can be efficiently minimized by the Newton method—one step of the Newton method for $\mathcal{V}_k$ made from x_{k-1} is sufficient. For more details and theoretical analysis, refer to [**?**, **?**].

The following theorem gives an estimate for the complexity of Vaidya's algorithm.

Theorem B.10 *Let $\mathcal{B}_\rho$ and $\mathcal{B}_R$ be some Euclidean balls of radii ρ and R, respectively, such that $\mathcal{B}_\rho \subset Q \subset \mathcal{B}_R$ and let $M > 0$ be a number such that $|f(x_1) - f(x_2)| \leqslant M; \forall x_1, x_2 \in Q$.Then Vaidya's method finds an ε-solution of problem* (B.11) *in $O\left(n \ln \frac{nMR}{\rho\varepsilon}\right)$ iterations.*

B.3.2.2 The ellipsoid method

Recall that an ellipsoid is a convex set of the form

$$\mathcal{E} = \left\{x \in \mathbb{R}^n : (x - c)^\top H^{-1}(x - c) \le 1\right\}$$

where $c \in \mathbb{R}^n$ and H is a symmetric positive definite matrix.

In this subsection we mention to the ellipsoid method with δ-subgradient (listed as Algorithm 7, below. Note that when $\delta = 0$ then the δ-ellipsoid method coincide with the usual ellipsoid method [**?**]), for problem (B.11), where in each iteration of this method we use the δ-subgradient of the objective function. The concept of δ-subgradient can be defined as follows

Definition B.8 Let $\delta \ge 0$, we call $v \in \mathbb{R}^n$ a δ-subgradient of the convex function $f : Q \longrightarrow \mathbb{R}$ at the point $y \in Q$, if $f(x) \ge f(y) + \langle v, x - y\rangle - \delta,\ \forall x \in Q$. The set of δ-subgradients of f at y denoted by $\partial_\delta f(y)$.

Note, that the δ-subgradient coincides with the usual subgradient when $\delta = 0$.

For Algorithm 7, in [**?**], it was proved the following result.

Theorem B.11 *Let Q be a compact convex set, which is contained in some Euclidean ball of radius R and includes some Euclidean ball of radius ρ, $f : Q \to \mathbb{R}$ is a*

Algorithm 7 Ellipsoid method with δ-subgradient for the problem (B.11).

Input: Number of iterations $N > 0$, $\delta \geqslant 0$, the ball $\mathcal{B}_{\mathcal{R}} \supseteq Q$, with center c and radius $\mathcal{R}$.

$\mathcal{E}_0 := \mathcal{B}_{\mathcal{R}}, \quad H_0 := \mathcal{R}^2 I_n, \quad c_0 := c.$
for $k = 0, \ldots, N-1$ **do**
  if $c_k \in Q$ **then**
    $v_k := v \in \partial_\delta f(c_k)$,
    if $v_k = 0$ **then**
      return c_k,
    end if
  else
    $v_k := v$, where $v \neq 0$, such that $Q \subset \{x \in \mathcal{E}_k : \langle v, x - c_k \rangle \leqslant 0\}$.
  end if
  $c_{k+1} := c_k - \frac{1}{n+1} \frac{H_k v_k}{\sqrt{v_k^T H_k v_k}}$,
  $H_{k+1} := \frac{n^2}{n^2-1} \left(H_k - \frac{2}{n+1} \frac{H_k v_k v_k^T H_k}{v_k^T H_k v_k} \right)$,
  $\mathcal{E}_{k+1} := \{x : (x - c_{k+1})^T H_{k+1}^{-1} (x - c_{k+1}) \leqslant 1\}$,
end for

Output: $x_N = \arg \min_{x \in \{c_0, \ldots, c_N\} \cap Q} f(x)$.

continuous convex function, $B > 0$ is a number such that $|f(x) - f(x')| \leq B \; \forall x, x' \in Q$. After $N \geq 2n^2 \ln\left(\frac{R}{\rho}\right)$ iterations of Algorithm 7, it holds the following inequality in the output point $x_N \in Q$,

$$f(x_N) - f(x^*) \leq \frac{BR}{\rho} \exp\left(-\frac{N}{2n^2}\right) + \delta,$$

where x^* is one of the exact solution of the problem B.11. Additionally, if the function f is μ-strongly convex, then we have

$$\|x_N - x^*\|_2^2 \leq \frac{2}{\mu} \left(\frac{BR}{\rho} \exp\left(-\frac{N}{2n^2}\right) + \delta \right).$$

B.3.3 Bregman divergence basics

Let $d : Q \to \mathbb{R}$ be a distance generating function (also called *prox-function*),which is continuously differentiable and 1–strongly convex with respect to the norm $\|\cdot\|$, i.e.

$$d(y) \geq d(x) + \langle \nabla d(x), y - x \rangle + \frac{1}{2}\|y - x\|^2 \quad \forall\, x, y \in Q.$$

For all $x, y \in Q \subset \mathbf{E}$, we consider the corresponding *Bregman divergence*,

$$V(y, x) = d(y) - d(x) - \langle \nabla d(x), y - x \rangle.$$

Depending on the formulation of a specific problem, different approaches are possible to determine the prox-function (proximal setup) of the problem and the corresponding Bregman divergence: standard proximal setup (i.e. Euclidean), entropy, ℓ_1/ℓ_2, simplex, nuclear norm and spectahedron can be found, for example in [**?**].

There are well-known examples of distance generating functions, let us denote ℓ_p norm by $\|\cdot\|_p$ and the standard unit simplex in $\mathbb{R}^n$ by

$$S_n(1) = \left\{ x \in \mathbb{R}^n_+ : \sum_{i=1}^n x_i = 1 \right\}.$$

Consider the following two cases

- **Entropy prox-function.** If $p = 1$, then for any $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n) \in Q \subseteq S_n(1)$

$$d(x) = \sum_{k=1}^n x_k \ln x_k \quad \text{and} \quad V(y,x) = \sum_{k=1}^n y_k \ln\left(\frac{y_k}{x_k}\right). \tag{B.15}$$

- **Standard Euclidean prox-function.** If $p = 2$, then for any $x, y \in Q$

$$d(x) = \frac{1}{2}\|x\|_2^2 \quad \text{and} \quad V(y,x) = \frac{1}{2}\|y - x\|_2^2. \tag{B.16}$$

B.3.4 First-order optimization methods

The first-order methods, which nowadays have the most attention in the optimization community, go back to 1847 with the work of Cauchy on the steepest descent method. With the increase in the number of applications that can be modeled as large-scale or even huge-scale optimization problems, first-order methods, which require low iteration cost as well as low memory storage, have received much interest over the past few decades in order to solve smooth and non-smooth convex optimization problems.

Historically, the gradient descent and subgradient methods were the first numerical schemes proposed to solve optimization problems with smooth and non-smooth convex objective functions, respectively. In order to solve the problem (B.11), with $Q = \mathbf{E}$ and f is smooth, the gradient descent method generates a sequence of iterations of the form

$$x^{k+1} = x^k - \gamma_k \nabla f(x^k), \quad k = 0, 1, 2, \ldots \tag{B.17}$$

where $\gamma_k \in (0, 1]$ is a certain step-size at the k-th iteration.

In order to solve the problem (B.11), with $Q = \mathbf{E}$ and f is non-smooth, subgradient methods generate the same sequence as (B.17), but instead $\nabla f(x^k)$ we take a subgradient $g(x^k) \in \partial f(x^k)$.

For the constrained case of the problem (B.11), i.e. when $Q \neq \mathbf{E}$, subgradient projection iterations are given by

$$x^{k+1} = \text{Proj}_Q\left(x^k - \gamma_k g(x^k)\right) := \arg\min_{x \in Q} \left\| x - \left(x^k - \gamma_k g(x^k)\right)\right\|_2, \quad k = 0, 1, 2, ... \tag{B.18}$$

The subgradient methods are strongly linked to the Euclidean structure of the space $\mathbf{E}$. More precisely, the construction of methods depend on the Euclidean projection (see (B.18)). The formulation (B.17) can be written in another form, which is called *proximal formulation*, as follows

$$x^{k+1} = \arg\min_{x \in \mathbb{R}^n} \left\{ f(x^k) + \left\langle \nabla f(x^k), x - x^k \right\rangle + \frac{1}{2\gamma_k} \left\| x - x^k \right\|_2^2 \right\}, \quad k = 0, 1, 2, ... \tag{B.19}$$

A substantial generalization of (B.19) which allows for adjusting the method to the possibly non-Euclidean geometry of the problem is a Mirror Descent method. It extends the standard projected subgradient methods by replacing the Euclidean proximal term in (B.19) with a Bregman divergence

$$x^{k+1} = \arg\min_{x \in \mathbb{R}^n} \left\{ f(x^k) + \left\langle \nabla f(x^k), x - x^k \right\rangle + \frac{1}{\gamma_k} V(x^k, x) \right\}, \quad k = 0, 1, 2, ... \tag{B.20}$$

B.3.5 The composite optimization problem

Let $Q \subseteq \mathbf{E}$ be a closed convex set, and $f : Q \longrightarrow \mathbb{R}$ be a convex L-smooth function. Let we consider the following optimization problem:

$$\min_{x \in Q} F(x) := f(x) + h(x), \tag{B.21}$$

where $h : Q \longrightarrow \mathbb{R}$ is an arbitrary simple closed convex (in general, non-smooth) function, which is called *a composite*. The problem (B.21) is called a *composite optimization problem* or *problem with a structure*.

Such problems often arise in many applications. One of the most famous examples is the so-called LASSO (Least Absolute Shrinkage and Selection Operator) problem, motivated by statistical problems.

Also, as a specific example, we can consider the problem of restoring the correspondence matrix by measuring flows on links (edges) in a large computer network (Minimal Mutual Information Model), which reduces to a composite optimization problem of the form

$$\min_{x \in S_n(1)} F(x) = \frac{1}{2}\|Ax - b\|_2^2 + \mu \sum_{k=1}^{n} x_k \ln x_k,$$

where $S_n(1)$ is the unit simplex in n-dimensional space.

In order to solve (B.21), we can use the *Similar Triangles Method* (STM) [89]. By this method, as a result, it will be a sequence of points $\{x^k\}_{k\geq 1}$, for which we have the following convergence rate

$$F\left(x^k\right) - F\left(x^*\right) \leq \frac{4Ld\left(x^*\right)}{k(k+1)} - \frac{2L}{k(k+1)}\left\|v^k - x^*\right\|^2; \quad \forall k \geq 1. \tag{B.22}$$

where x^* is an optimal solution to problem (B.21), $d(\cdot)$ is a prox function on Q, and $\{v^k\}_{k\geq 1}$ is a sequence generated by STM, by solving an auxiliary (easily solvable) minimization problem.

B.4 Lower complexity bounds for the variational inequalities and saddle point problems

For the following saddle point problem

$$\min_x \max_y F(x, y), \tag{B.23}$$

let we focus in the following classes of problems [**?**].

1. The first class denoted by $\mathcal{F}(L_{xx}, L_{yy}, L_{xy}, \mu_x, \mu_y)$, with $\mu_x > 0$ and $\mu_y > 0$. In this class, the function $F(\cdot, y)$ is μ_x-strongly convex for any fixed y and $F(x, \cdot)$ is μ_y-strongly concave for any fixed x.The function F is smooth and ∇F is Lipschitz for any x_1, x_2, y_1, y_2, i.e.

$$\begin{aligned}
\left\|\nabla_x F\left(x_1, y\right) - \nabla_x F\left(x_2, y\right)\right\| &\leq L_{xx}\left\|x_1 - x_2\right\|, \\
\left\|\nabla_y F\left(x, y_1\right) - \nabla_y F\left(x, y_2\right)\right\| &\leq L_{yy}\left\|y_1 - y_2\right\|, \\
\left\|\nabla_x F\left(x, y_1\right) - \nabla_x F\left(x, y_2\right)\right\| &\leq L_{xy}\left\|y_1 - y_2\right\|, \\
\left\|\nabla_y F\left(x_1, y\right) - \nabla_y F\left(x_2, y\right)\right\| &\leq L_{xy}\left\|x_1 - x_2\right\|.
\end{aligned} \tag{B.24}$$

   For this class of problems, we have the following lower complexity bound

$$\Omega\left(\sqrt{\frac{L_{xx}}{\mu_x} + \frac{L_{xy}^2}{\mu_x\mu_y} + \frac{L_{yy}}{\mu_y}}\ln\left(\frac{1}{\varepsilon}\right)\right). \tag{B.25}$$

2. The second class is a special class of the first. It is a bilinear class of problem denoted by $\mathcal{B}(L_{xy}, \mu_x, \mu_y)$, with $\mu_x > 0$ and $\mu_y > 0$. The problems in this class written as

$$\min_x \max_y F(x, y) := f(x) + x^\top A y - g(y), \tag{B.26}$$

   where $f(x)$ is μ_x-strongly convex and $g(y)$ is μ_y-strongly concave, and the matrix A satisfies $\|A\|_2 \leq L_{xy}$. For this class, we assume that the following

prox-operators are available, for some $\alpha > 0, \beta > 0$:

$$\text{prox}_f(v) := \underset{x}{\text{argmin}} \left\{ f(x) + \frac{1}{2\alpha}\|x - v\|^2 \right\},$$

$$\text{prox}_g(u) := \underset{y}{\text{argmin}} \left\{ g(y) + \frac{1}{2\beta}\|y - u\|^2 \right\}.$$

For this class of problems, we have the following lower complexity bound

$$\Omega\left(\sqrt{\frac{L_{xy}^2}{\mu_x \mu_y}} \ln\left(\frac{1}{\varepsilon}\right) \right). \tag{B.27}$$

An optimal algorithm for previously mentioned classes of problems was recently proposed in [66].

As a result from the previous lower bounds, for the general convex-concave saddle point problems (i.e. with $\mu_x = \mu_y = 0$), we have the following classes.

1. For the class $\mathcal{F}(L_{xx}, L_{yy}, L_{xy}, 0, 0)$, we have the following lower bound

$$\Omega\left(\sqrt{\frac{L_{xx}R_x^2}{\varepsilon}} + \frac{L_{xy}R_xR_y}{\varepsilon} + \sqrt{\frac{L_{yy}R_y^2}{\varepsilon}} \right). \tag{B.28}$$

   Optimal algorithms, for this class of problems, were proposed in [**?**, **?**, **?**].
2. For the class $\mathcal{B}(L_{xy}, 0, 0)$, we have the following lower bound

$$\Omega\left(\frac{L_{xy}R_xR_y}{\varepsilon} \right), \tag{B.29}$$

   where in (B.28) and (B.29), we have $\|x^*\| \leq R_x, \|y^*\| \leq R_y$ and $(x^*, y^*) = \text{argmin}_x \text{argmax}_y F(x, y)$.

Appendix C
?

Appendix D
?

Appendix E
Regularization and restarts in convex (stochastic) optimization and saddle point problems

The optimal results presented in Table B.1 are remarkable in that the estimate in one cell can be easily achieved from estimate in another cell by the same method with a slight modification, depending on this pair of cells. Let's demonstrate, for example, that optimal complexity for convex case (in both continuous and smooth cases) can be obtained from known optimal complexities for μ-strongly convex case. The trick is to consider regularized problem

$$\min_{x \in Q} f_\mu(x) := f(x) + \frac{\mu}{2}\|x - x^0\|^2 \tag{E.1}$$

with $\mu = \varepsilon^2/R$ and ε is desired accuracy. Now, one can use optimal algorithm for strongly convex optimization to solve this problem, and the result of its operation will be desired solution for original problem $\min_{x \in Q} f(x)$. Indeed, if it holds that

$$f_\mu(x) - \min_{x \in Q} f_\mu(x) \leq \frac{\varepsilon}{2} \tag{E.2}$$

for some point $x \in Q$, then

$$\begin{aligned} f(x) - f(x^*) &= (f(x) - f_\mu(x)) + (f_\mu(x) - \min_{x \in Q} f_\mu(x)) + (\min_{x \in Q} f_\mu(x) - f(x^*)) \\ &\leq 0 + \frac{\varepsilon}{2} + \frac{\varepsilon}{2R^2}\|x^0 - x^*\|^2 \leq \varepsilon. \end{aligned} \tag{E.3}$$

Note also, that convergence rates in Table B.1 for convex case are equal to convergence rates for strongly convex case if we substitute $\mu = \varepsilon/R^2$ to them, so the convergence rate of algorithm we used above in terms of function discrepancy for initial non-strongly convex problem corresponds to presented estimates. Thus, optimal convergence rate on the class of non-strongly convex functions cannot be worse than that presented in Table B.1 and obtained by our reduction.

Let us now describe the inverse reduction, that is how optimal convergence rate for μ-strongly convex case can be obtained from optimal convergence rate for not-strongly convex case and achieved after slight modification by the same optimal

not-strongly convex optimization algorithm. It can be done with a help of restarts technique. For example, in the case of Lipschitz continuous ∇f and convex f, function discrepancy after k iterations of optimal algorithm, in accordance with Table B.1, is

$$f(x^k) - f(x^*) \leq \frac{CL\|x^0 - x^*\|^2}{k^2}, \tag{E.4}$$

for some $C > 0$. On the other hand, due to μ-strong convexity, we have

$$\frac{\mu}{2}\|x^k - x^*\|^2 \leq f(x^k) - f(x^*). \tag{E.5}$$

Let's choose $k = 2\sqrt{CL/\mu}$. Then, it holds that

$$\|x^k - x^*\|^2 \leq \frac{1}{2}\|x^0 - x^*\|^2. \tag{E.6}$$

In other words, optimal convex optimization algorithm can half argument discrepancy in k iterations. After that, we restart the algorithm by setting $x^0 = x^k$ to half it again, and so on. To achieve the desired accuracy ε, it is sufficient to achieve $2\varepsilon/\mu$ argument discrepancy, so we need to perform $N = \log_2 \frac{\mu R^2}{2\varepsilon} + 1$ restarts. Each of them takes $2\sqrt{CL/\mu}$ iterations, that finally leads to convergence rate presented in Table B.1. We can similarly reduce convergence rate for Lipschitz continuous f case, but reasoning would be a little more cumbersome. Thus, optimal convergence rate on the class of strongly convex functions cannot be worse than that presented in Table B.1 and obtained by the described reduction. Together with previous paragraph this demonstrates that optimal convergence rates in, for example, strongly convex case completely determine optimal convergence rates in non-strongly convex case and vice versa.

Note that the same reasoning works for stochastic optimization problems as well. Indeed, if there is a method having optimal convergence on average, one can replace the corresponding discrepancies in the text above with $\mathbb{E}\left[f(x^k) - f(x^*)\right]$ and $\mathbb{E}\left[\|x^k - x^*\|^2\right]$, and obtain the same procedures and guarantees. Things are different if we analyse probabilities of large deviations. In a case of regularization, we need to change current reasoning only by adding a clause "with probability $1 - \beta$", because relation between original and regularized problem is deterministic. But to demonstrate that restarts work too, we need in more delicate argument. Let us demand that resulting probability of big deviations is lower than β. Assume that (E.4) holds with probability $1 - \beta/N$ for each restart, or, in other words, does not hold with probability β/N. A probability of that on at least one of N restarts it does not hold (union of events that it does not hold on i-th restart, $i = 1, ..., N$) can be upper bounded by $\beta/N + ... + \beta/N$ (N times). So, a probability of that desired accuracy is reached is greater than $1 - N \cdot \beta/N = 1 - \beta$, as was to be shown.

The constructions of regularization and restarts take important place in optimization theory, because some of the optimal convergence rates can be reached only with a help of reductions described above at the moment. Below, we present two such examples.

Nesterov's accelerated method and $\log \frac{\mu}{\varepsilon}$ in a strongly-convex case

Let us consider strongly-convex optimization problem $\min_{x \in Q} f(x)$. It is known that Nesterov's accelerated method with iteration as follows

$$x^{k+1} = y^k - \frac{1}{L}\nabla f(y^k) \tag{E.7}$$

$$y^{k+1} = x^{k+1} + \frac{k}{k+3}(x^{k+1} - x^k) \tag{E.8}$$

has optimal convergence rate $O\left(\sqrt{\frac{LR^2}{\varepsilon}}\right)$ on the class of non-strongly convex functions. One can apply restarts to this method to obtain optimal method of strongly-convex optimization with convergence rate $O\left(\sqrt{\frac{L}{\mu}}\log\frac{\mu R^2}{\varepsilon}\right)$. On the other hand, construction proposed by Nesterov allows to obtain method with following iteration

$$x^{k+1} = y^k - \frac{1}{L}\nabla f(y^k) \tag{E.9}$$

$$y^{k+1} = x^{k+1} + \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}(x^{k+1} - x^k) \tag{E.10}$$

such that it holds that

$$f(x^k) - f(x^*) \leq LR^2\left(1 - \sqrt{\frac{\mu}{L}}\right)^k, \tag{E.11}$$

which means that its convergence rate is $O\left(\sqrt{\frac{L}{\mu}}\log\frac{LR^2}{\varepsilon}\right)$. This convergence rate is not optimal because of presence of L under logarithm. Nevertheless, this method is usually referred as optimal because logarithmic factor grows slow enough to be ignored in practice (notation of $\tilde{O}(\cdot)$ asymptotic ignoring logarithms also became widespread in literature; presented method, thereby, has $\tilde{O}\left(\sqrt{\frac{L}{\mu}}\right)$ convergence rate). It is a common problem that L appears under logarithm, so numerous optimal results we know with only μ under logarithm can be achieved only by restarted methods.

Mirror-Prox in a strongly-convex strongly-concave case

Index