ГОУ ВПО «Московский физико-технический институт (государственный университет)»
Физтех-школа Прикладной Математики и Информатики
кафедра проблем передачи информации и анализа данных

Работа допущена к защите
зав. кафедрой

_____ Соболевский А.Н.

«____» _____ 2020 г.

Выпускная квалификационная работа на соискание степени

МАГИСТРА

Тема: **Безградиентные и стохастические методы оптимизации, децентрализованная распределённая оптимизация**

Направление: 03.04.01 – Прикладные математика и физика (магистратура)

Выполнил студент гр. М05-875 _____ Горбунов Эдуард Александрович

Научный руководитель,
д.ф.-м.н. _____ Гасников Александр Владимирович

Москва – 2020

Contents

Глава 1. Аннотация 5

Глава 2. Introduction 6

Глава 3. Notations and Definitions 7

Глава 4. Optimal Decentralized Distributed Algorithms for Stochastic Convex Optimization 9

4.1. Introduction 9

4.1.1. Contributions 11

4.2. Optimal Bounds for Stochastic Convex Optimization 12

4.3. Similar Triangles Method with Inexact Proximal Step 14

4.4. Stochastic Convex Optimization with Affine Constraints: Primal Approach 18

4.5. Stochastic Convex Optimization with Affine Constraints: Dual Approach . 20

4.5.1. Convex Dual Function 21

4.5.2. Strongly Convex Dual Functions and Restarts Technique 23

4.5.3. Direct Acceleration for Strongly Convex Dual Function 30

4.6. Applications to Decentralized Distributed Optimization 36

4.7. Discussion 44

4.7.1. Possible Extensions 45

4.8. Application for Population Wasserstein Barycenter Calculation 46

4.8.1. Definitions and Properties 46

4.8.2. SA Approach 48

4.8.3. SAA Approach 49

4.8.4. SA vs SAA comparison 52

4.9. Missing Proofs, Technical Lemmas and Auxiliary Results 53

4.9.1. Basic Facts 53

4.9.2. Useful Facts about Duality 54

4.9.3. Auxiliary Results 55

4.9.4. Missing Proofs from Section 4.3 56

Proof of Lemma 4.3.1 56

Proof of Lemma 4.3.2 . . . 56
Proof of Theorem 4.3.1 . . . 59
Proof of Corollary 4.3.1 . . . 60
4.9.5. Missing Proofs from Section 4.4 . . . 60
Proof of Theorem 4.4.1 . . . 60
Proof of Theorem 4.4.2 . . . 61
4.9.6. Missing Lemmas and Proofs from Section 4.5.1 . . . 62
Lemmas . . . 62
Proof of Theorem 4.5.1 . . . 72
4.9.7. Missing Proofs from Section 4.5.2 . . . 84
Proof of Theorem 4.5.5 . . . 84
Proof of Corollary 4.5.3 . . . 86
4.9.8. Missing Proofs from Section 4.5.3 . . . 87
Proof of Lemma 4.5.1 . . . 87
Proof of Lemma 4.5.2 . . . 89
Proof of Theorem 4.5.6 . . . 93
Proof of Corollary 4.5.5 . . . 97
4.9.9. Technical Results . . . 99

Глава 5. Stochastic Derivative Free Optimization Methods with Momentum 104
5.1. Introduction . . . 104
5.2. Stochastic Momentum Three Points (SMTP) . . . 107
5.2.1. Non-Convex Case . . . 109
5.2.2. Convex Case . . . 110
5.2.3. Strongly Convex Case . . . 111
5.3. Stochastic Momentum Three Points with Importance Sampling (SMTP_IS) 113
5.3.1. Non-convex Case . . . 113
5.3.2. Convex Case . . . 115
5.3.3. Strongly Convex Case . . . 116
5.4. Comparison of SMTP and SMTP_IS . . . 117
5.5. Experiments . . . 118
5.6. Conclusion . . . 120
5.7. Missing Proofs, Technical Lemmas and Auxiliary Results . . . 120

5.7.1. Preliminaries 120
5.7.2. Missing Proofs from Section 5.2 122
Non-Convex Case 122
Convex Case 123
Strongly Convex Case 125
5.7.3. Missing Proofs from Section 5.3 128
Non-convex Case 129
Convex Case 131
Strongly Convex Case 133
5.7.4. Auxiliary results 137

References 139

Chapter 1

Аннотация

Оптимизация является одним из ключевых инструментов во многих приложениях. В частности, задачи оптимизации возникают в огромном числе задач машинного обучения и анализа данных. В последние годы безградиентные методы оптимизации стали основным инструментом в приложениях обучения с подкреплением и оптимального управления. Кроме того, огромный интерес исследователей привлекает распределённая оптимизация: обучение многих глубоких нейросетевых моделей практически не возможно или требует слишком больших вычислительных и временных ресурсов, если делать это не распределённо, а на одной компьютере/сервере. В этой диссертации предлагаются новые безградиентные стохастические методы оптимизации, а также новые ускоренные стохастические методы распределённой оптимизации.

В первой части диссертации рассматривается задача стохастической децентрализованной оптимизации. Предлагаются новые методы, использующие детерминированный прямой оракул и стохастический двойственный оракул, а также доказываются оценки скорости сходимости с большой вероятностью на классах выпуклых и сильно выпуклых гладких функций. На примере вычисления популяционного барицентра Вассерштейна сравниваются прямой и двойственный подходы к решению этой задаче, на основе новых результатов, полученных в данной работе.

Во второй части диссертации предлагается новый безградиентный метод (SMTP) с моментным членом в форме «тяжёлого шарика» и анализируется его скорость сходимости по математическому ожиданию для невыпуклых, выпуклых и сильно выпуклых функций. Кроме того, предлагается модификация метода (SMTP_IS), которая использует неравномерное сэмплирование направлений поиска, учитывающее изменение свойств гладкости функции вдоль разных направлений, что позволяет улучшить оценки скорости сходимости в предположении покомпонентной гладкости целевой функции.

Chapter 2

Introduction

Optimization plays a central role in different applications. In particular, optimization tasks appear in a huge number of machine learning problems. In recent years derivative-free methods became a key tool in reinforcement learning. Moreover, distributed optimization attracts a lot of attention from the machine learning community since the training of deep neural networks is often impossible or takes prohibitively long time while training is performed on a single machine. In this dissertation, we propose new derivative-free methods, as well as novel accelerated stochastic distributed methods.

In Chapter 4 we focus on stochastic decentralized distributed optimization problems. We propose new methods based on deterministic primal first-order oracle and stochastic dual first-order oracle and derive optimal convergence rates with high probability for smooth convex and strongly convex objectives. To illustrate the difference between the two approaches, we consider the problem of the population Wasserstein barycenter calculation.

In Chapter 5, we focus on the problems when the objective function is available only through the zeroth-order oracle. For this problem, we develop two new methods — `SMTP` and `SMTP_IS` — and analyze their convergence for non-convex, convex, and strongly convex objectives. Both methods are based on the heavy-ball method, and `SMTP_IS` uses coordinate-wise smoothness of the objective function and importance sampling trick.

In both chapters, we provide a detailed introduction to the topic and literature review. Full proofs of the proposed results are at the ends of the corresponding chapters as well as technical lemmas and auxiliary results. All notations and definitions are introduced in Chapter 3.

Chapter 3

Notations and Definitions

To denote standard inner product between two vectors $x, y \in \mathbb{R}^n$ we use $\langle x, y \rangle \stackrel{\text{def}}{=} \sum_{i=1}^n x_i y_i$, where x_i is i-th coordinate of vector x, $i = 1, \ldots, n$. Standard Euclidean norm of vector $x \in \mathbb{R}^n$ is defined as $\|x\|_2 \stackrel{\text{def}}{=} \sqrt{\langle x, x \rangle}$ and we use $\|\cdot\|_p$ to define ℓ_p-norm of the vector $x \in \mathbb{R}^d$: $\|x\|_p \stackrel{\text{def}}{=} \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$ for $p \geq 1$ and $\|x\|_1 \stackrel{\text{def}}{=} \max_{i \in [d]} |x_i|$. We use $\|\cdot\|^*$ to define the conjugate norm for the norm $\|\cdot\|$: $\|x\|^* \stackrel{\text{def}}{=} \max \{\langle a, x \rangle \mid a \in \mathbb{R}^n, \|a\| \leq 1\}$. By $\lambda_{\max}(A)$ and $\lambda^+_{\min}(A)$ we mean maximal and minimal positive eigenvalues of matrix $A \in \mathbb{R}^{n \times n}$ respectively and we use $\chi(A) \stackrel{\text{def}}{=} \lambda_{\max}(A) / \lambda^+_{\min}(A)$ to denote condition number of A. To define vector of ones we use $\mathbf{1}_n \stackrel{\text{def}}{=} (1, \ldots, 1)^\top \in \mathbb{R}^n$ and omit the subscript n when one can recover the dimension from the context. Moreover, we use $\widetilde{O}(\cdot)$, $\widetilde{\Omega}(\cdot)$ and $\widetilde{\Theta}(\cdot)$ that define exactly the same as $O(\cdot)$, $\Omega(\cdot)$ and $\Theta(\cdot)$ but besides constants factors they can hide polylogarithmical factors of the parameters of the method or the problem. Operator $\mathbf{E}[\cdot]$ denotes mathematical expectation with respect to all randomness and $\mathbf{E}_{s \sim \mathcal{D}}[\cdot]$ denotes conditional expectation w.r.t. randomness coming from random vector s which is sampled from probability distribution $\mathcal{D}$ on $\mathbb{R}^n$. Conditional mathematical expectation with respect to all randomness coming from random variable ξ is denoted by $\mathbf{E}_\xi[\cdot]$. We use $B_r(y) \subset \mathbb{R}^n$ to denote Euclidean ball centered at $y \in \mathbb{R}^n$ with radius r: $B_r(y) \stackrel{\text{def}}{=} \{x \in \mathbb{R}^n \mid \|x - y\|_2 \leq r\}$. The Kronecker product of two matrices $A \in \mathbb{R}^{m \times m}$ with elements A_{ij}, $i, j = 1, \ldots, m$ and $B \in \mathbb{R}^{n \times n}$ is such $mn \times mn$ matrix $C \stackrel{\text{def}}{=} A \otimes B$ that

$$C = \begin{bmatrix} A_{11}B & A_{12}B & A_{13}B & \ldots & A_{1m}B \\ A_{21}B & A_{22}B & A_{23}B & \ldots & A_{2m}B \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ A_{m1}B & A_{m2}B & A_{m3}B & \ldots & A_{mm}B \end{bmatrix}. \tag{3.1}$$

By I_n we denote $n \times n$ identity matrix and omit the subscript when the size of the matrix is obvious from the context.

Below we list some classical definitions for optimization (see, for example, [1] for the details).

Definition 3.0.1 (L-smoothness). *Function f is called L-smooth in $Q \subset \mathbb{R}^n$ with $L > 0$*

when it is di erentiable and its gradient is L-Lipschitz continuous in Q, i.e.

$$k\nabla f(x) - \nabla f(y)k_2 \leq Lkx - yk_2, \quad \forall x, y \in Q. \tag{3.2}$$

From this definition one can obtain

$$jf(y) - f(x) - h\nabla f(x), y - xij \leq \frac{L}{2}ky - xk_2^2, \quad \forall x, y \in \mathbb{R}^d, \tag{3.3}$$

and if additionally f is convex, i.e. $f(y) \geq f(x) + h\nabla f(x), y - xi$, we have

$$k\nabla f(x)k_2^2 \leq 2L(f(x) - f(x^*)), \quad \forall x \in \mathbb{R}^d. \tag{3.4}$$

Definition 3.0.2 (μ-strong convexity). *Di erentiable function f is called μ-strongly convex in $Q \subseteq \mathbb{R}^n$ with $\mu \geq 0$ if*

$$f(x) \geq f(y) + h\nabla f(y), x - yi + \frac{\mu}{2}kx - yk_2^2, \quad \forall x, y \in Q. \tag{3.5}$$

If $\mu > 0$ then there exists unique minimizer of f on Q which we denote by x^*, except the situations when we explicitly specify x^* in a different way. In the case when $\mu = 0$, i.e. f is convex, we assume that there exists at least one minimizer x^* of f on Q and in the case when the set of minimizers of f on the set Q is not a singleton we choose x^* to be either arbitrary or closest to the starting point of a method. When we consider some optimization method with a starting point x^0 we use R or R_0 to denote the Euclidean distance between x^0 and x^*.

Chapter 4

Optimal Decentralized Distributed Algorithms for Stochastic Convex Optimization

The results proposed in this chapter were obtained by the author of this thesis in [2].

4.1. Introduction

In this chapter we are interested in the convex optimization problem

$$\min_{x \in Q \subseteq \mathbb{R}^n} f(x), \tag{4.1}$$

where f is a convex function and Q is closed and convex subset of $\mathbb{R}^n$. More precisely, we study particular case of (4.1) when the objective function f could be represented as a mathematical expectation

$$f(x) = \mathbf{E}\left[f(x, \xi)\right], \tag{4.2}$$

where ξ is a random variable. Problems of this type play central role in a bunch of applications of machine learning [3, 4] and mathematical statistics [5]. Typically x represents feature vector defining the model, only samples of ξ are available and the distribution of ξ is unknown. One possible way to minimize generalization error (4.2) is to solve empirical risk minimization or finite-sum minimization problem instead, i.e. solve (4.1) with the objective

$$\hat{f}(x) = \frac{1}{m} \sum_{i=1}^{m} f(x, \xi_i), \tag{4.3}$$

where m should be sufficiently large to approximate the initial problem. Indeed, if $f(x, \xi)$ is convex and M-Lipschitz continuous for all ξ, Q has finite diameter D and $\hat{x} = \operatorname{argmin}_{x \in Q} \hat{f}(x)$, then (see [6, 7]) with probability at least $1 - \beta$

$$f(\hat{x}) - \min_{x \in Q} f(x) = O\left(\sqrt{\frac{M^2 D^2 n \ln(m) \ln(n/\beta)}{m}}\right), \tag{4.4}$$

and if additionally $f(x, \xi)$ is μ-strongly convex for all ξ, then (see [8]) with probability at least $1 - \beta$

$$f(\hat{x}) - \min_{x \in Q} f(x) = O\left(\frac{M^2 D^2 \ln(m) \ln(m/\beta)}{\mu m} + \sqrt{\frac{M^2 D^2 \ln(1/\beta)}{m}}\right). \tag{4.5}$$

In other words, to solve (4.1)+(4.2) with ε functional accuracy via minimization of empirical risk (4.3) it is needed to have $m = \widetilde{(M^2D^2n/\varepsilon^2)}$ in the convex case and $m = \widetilde{(\max\{M^2D^2/\varepsilon, M^2D^2/\varepsilon^2\mu\})}$ in the μ-strongly convex case where $\widetilde{(\cdot)}$ hides a constant factor, a logarithmic factor of $1/\varepsilon$ and a polylogarithmic factor of $1/\varepsilon$.

Stochastic first-order methods such as Stochastic Gradient Descent (SGD) [9–13] or its accelerated variants like AC-SA [14] or Similar Triangles Method (STM) [15–17] are very popular choice to solve either (4.1)+(4.2) or (4.1)+(4.3). In contrast with their cheap iterations in terms of computational cost, these methods converge only to the neighbourhood of the solution, i.e. to the ball centered at the optimality and radius proportional to the standard deviation of the stochastic estimator. For the particular case of finite-sum minimization problem one can solve this issue via variance-reduction trick [18–21] and its accelerated variants [22–24]. Unfortunately, this technique is not applicable in general for the problems of type (4.1)+(4.2). Another possible way to reduce the variance is mini-batching. When the objective function is L-smooth one can accelerate computations of batches using parallelization [16, 25–27] and it is one of the examples where centralized distributed optimization appears naturally [28].

In other words, in some situations, e.g. when the number of samples m is too big, it is preferable in practice to split the data into q blocks, assign each block to the separate worker, e.g. processor, and organize computation of the gradient or stochastic gradient in the parallel or distributed manner. Moreover, in view of (4.4)-(4.5) sometimes to solve an expectation minimization problem it is needed to have such a big number of samples that corresponding information (e.g. some objects like images, videos and etc.) cannot be stored on 1 machine because of the memory limitations (see Section 4.8 for the detailed example of such a situation). Then, we can rewrite the objective function in the following form

$$f(x) = \frac{1}{q}\sum_{i=1}^{q} f_i(x), \quad f_i(x) = \mathbf{E}_{\xi_i}[f(x,\xi_i)] \text{ or } f_i(x) = \frac{1}{s_i}\sum_{j=1}^{s_i} f(x,\xi_{ij}). \tag{4.6}$$

Here f_i corresponds to the loss on the i-th data block and could be also represented as an expectation or a finite sum. So, the general idea for parallel optimization is to compute gradients or stochastic gradients by each worker, then aggregate the results by the master node and broadcast new iterate or needed information to obtain the new iterate back to the workers.

The visual simplicity of the parallel scheme hides synchronization drawback and high

requirement to master node [29]. The big line of works is aimed to solve this issue via periodical synchronization [30–33], error-compensation [34, 35], quantization [36–40] or combination of these techniques [41, 42].

However, in this chapter we mainly focus on another approach to deal with aforementioned drawbacks — decentralized distributed optimization [28, 43]. It is based on two basic principles: every node communicates only with its neighbours and communications are performed simultaneously. Moreover, this architecture is more robust, e.g. it can be applied to time-varying (wireless) communication networks [44].

4.1.1. Contributions

One can consider this chapter as a continuation of work [45] where authors mentioned the key ideas that form a basis of this work. However, in this chapter we provide formal proofs of some results announced in [45] together with couple of new results that were not mentioned. Our contributions include:

- **Accelerated primal-dual method with biased stochastic dual oracle for convex and smooth dual problem.** We extent the result from the recent work [46] to the case when we have an access to the biased stochastic gradients. We emphasize that our analysis works for the minimization on whole space and we do not assume that the sequence generated by the method is bounded. It creates extra difficulties in the analysis, but we handle it via advanced technique for estimating recurrences (see also [46, 47]).

- **Two accelerated methods with stochastic dual oracle for strongly convex and smooth dual problem.** For the case when the dual function is strongly convex with Lipschitz continuous gradient we analyze two methods: one is R-RRMA-AC-SA^2 and another is SSTM_sc. The first one was described in [46], but in this dissertation we formally state the method and prove high probability bounds for its convergence rate. The second method is also well-known, but to the best of our knowledge there were no convergence results for it in such generality that we handle. That is, we consider SSTM_sc with *biased* stochastic oracle applied to the *unconstrained* smooth and strongly convex minimization problem and prove high probability bounds for its convergence rate together with the bound for the noise level. As for the convex case, we also do not assume that the sequence generated by the method is bounded. Then

we show how it can be applied to solve stochastic optimization problem with affine constraints using dual oracle.

- **Analysis of STM applied to convex smooth minimization problem with smooth convex composite term and inexact proximal step for unconstrained minimization.** Surprisingly, but before this work there were no analysis for STM in this case. The closest work to ours in this topic is [48], but in [48] authors considered optimization problems on bounded sets.

4.2. Optimal Bounds for Stochastic Convex Optimization

In this section our goal is to present the overview of the optimal methods and their convergence rates for the stochastic convex optimization problem (4.1)+(4.2) in the case when the gradient of the objective function is available only through (possibly biased) stochastic estimators with "light tails" or, equivalently, with σ^2-subgaussian variance. That is, we are interested in the situation when for an arbitrary $x \in Q$ one can get such stochastic gradient $\nabla f(x, \xi)$ that

$$\|\mathbf{E}[\nabla f(x,\xi)] - \nabla f(x)\|_2 \leq \delta, \tag{4.7}$$

$$\mathbf{E}\left[\exp\left(\frac{\|\nabla f(x,\xi) - \mathbf{E}[\nabla f(x,\xi)]\|_2^2}{\sigma^2}\right)\right] \leq \exp(1), \tag{4.8}$$

where $\delta \geq 0$ and $\sigma \geq 0$. If $\sigma = 0$, let us suppose that $\nabla f(x, \xi) = \mathbf{E}[\nabla f(x, \xi)]$ almost surely in ξ. When $\sigma = \delta = 0$ we get that $\nabla f(x, \xi) = \nabla f(x)$ almost surely in ξ which is equivalent to the deterministic first-order oracle. For clarity, we start with this simplest case of stochastic oracle and provide an overview of the state-of-the-art results for this particular case in Table 4.1. Note that for the methods mentioned in the table number of oracle calls and number of iterations are identical. In the case when the gradient of f is bounded it is often enough to assume this only in some ball centered at the optimality point x^* with radius proportional to R [17, 49, 50].

In this chapter we are mainly focus on smooth optimization problems and use different modifications of Similar Triangles Method (STM) since it gives optimal rates in this case and it is easy enough to analyze at least in the deterministic case. For convenience, we state the method in this section as Algorithm 1. Interestingly, if we run STM with $\mu > 0$ to solve (4.1) with μ-strongly convex and L-smooth objective, it will return x^N such that

Assumptions on f	Method	Citation	# of oracle calls
μ-strongly convex, L-smooth	R-STM	[16] [17]	$O\left(\sqrt{\frac{L}{\varepsilon}} \ln\left(\frac{R^2}{\varepsilon}\right)\right)$
L-smooth	STM	[16] [17]	$O\left(\sqrt{\frac{LR^2}{\varepsilon}}\right)$
μ-strongly convex, $\|\nabla f(x)\|_2 \le M$	MD	[51] [52]	$O\left(\frac{M^2}{\varepsilon}\right)$
$\|\nabla f(x)\|_2 \le M$	MD	[51] [52]	$O\left(\frac{M^2R^2}{\varepsilon^2}\right)$

Table 4.1: Optimal number N of deterministic first-order oracle calls in order to get such a point x^N that $f(x^N) - f(x^*) \le \varepsilon$. First column contains assumptions on f in addition to the convexity. MD states for Mirror Descent.

Algorithm 1 Similar Triangles Methods (STM), the case when $Q = \mathbb{R}^n$

Require: $x^0 = z^0 = x^0$, number of iterations N, $\alpha_0 = A_0 = 0$

1: **for** $k = 0, \ldots, N$ **do**

2: Set $\alpha_{k+1} = (1 + A_k)/2L + \sqrt{(1 + A_k)/4L^2 + A_k(1 + A_k)/L}$, $A_{k+1} = A_k + \alpha_{k+1}$

3: $x^{k+1} = (A_k x^k + \alpha_{k+1} z^k)/A_{k+1}$

4: $z^{k+1} = z^k - \left(\nabla f(x^{k+1}) - \mu x^{k+1}\right) \alpha_{k+1}/(1 + \mu\alpha_{k+1})$

5: $x^{k+1} = (A_k x^k + \alpha_{k+1} z^{k+1})/A_{k+1}$

6: **end for**

Ensure: x^N

$f(x^N) - f(x^*) \le \varepsilon$ after $N = O\left(\sqrt{L/\varepsilon} \ln\left(LR^2/\varepsilon\right)\right)$ iterations which is not optimal, see[1] Table 4.1. To match the optimal bound in this case one should use classical restart of STM which is run with $\mu = 0$ [16].

We notice that another highly widespread in machine learning applications type of problems is regularized or composite optimization problem

$$\min_{x \in Q} f(x) + h(x), \tag{4.9}$$

[1] In some places we put references not to the first work where this bound was shown but to the works where this complexity bound was shown for either more convenient or more relevant to our work method.

where h is a convex proximable function. For this case STM can be generalized via modifying the update rule in the following way [16, 17]:

$$z^{k+1} = \operatorname*{argmin}_{z \in Q} \left\{ \frac{1}{2}\|z - z^0\|_2^2 + \sum_{l=0}^{k+1} \alpha_l \left(\langle \nabla f(x^l), z - x^l \rangle + h(z) + \frac{\mu}{2}\|z - x^l\|_2^2 \right) \right\}. \quad (4.10)$$

We address such problems with L_h-smooth composite term in the Appendix, see Section 4.3 for the details.

Next, we go back to the problem (4.1)+(4.2) and consider more general case when $\delta = 0$ and $\sigma^2 > 0$. In this case one can construct unbiased estimator

$$\nabla f(x, \{\xi_i\}_{i=1}^r) = \frac{1}{r} \sum_{i=1}^{r} \nabla f(x, \xi_i),$$

where $\xi_1, \ldots, \xi_r$ are i.i.d. samples and $\nabla f(x, \{\xi_i\}_{i=1}^r)$ has r times smaller variance than $\nabla f(x, \xi_i)$:

$$\mathbf{E}_{\xi_1, \ldots, \xi_r} \left[\exp\left(\frac{\|\nabla f(x, \{\xi_i\}_{i=1}^r) - \nabla f(x)\|_2^2}{\sigma^2/r} \right) \right] \le \exp(1).$$

Then in order to get such a point x^N that $f(x^N) - f(x^*) \le \varepsilon$ with probability at least $1 - \beta$ where $\beta \in (0, 1)$ and f is μ-strongly convex ($\mu \ge 0$) and L-smooth one can run STM for

$$N = O\left(\min\left\{ \sqrt{\frac{LR^2}{\varepsilon}}, \sqrt{\frac{L}{\mu}} \ln\left(\frac{LR^2}{\varepsilon} \right) \right\} \right) \quad (4.11)$$

iterations with small modification: instead of using $\nabla f(x^{k+1})$ the method uses mini-batched stochastic approximation $\nabla f(x^{k+1}, \{\xi_i\}_{i=1}^{r_{k+1}})$ where the batch size is

$$r_{k+1} = \left(\max\left\{ 1, \frac{\sigma^2 \alpha_{k+1} \ln \frac{N}{\beta}}{(1 + A_{k+1}\mu)\varepsilon} \right\} \right). \quad (4.12)$$

The total number of oracle calls is

$$\sum_{k=1}^{N} r_k = O\left(N + \min\left\{ \frac{\sigma^2 R^2}{\varepsilon^2} \ln\left(\frac{\sqrt{LR^2/\varepsilon}}{\beta} \right), \frac{\sigma^2}{\mu\varepsilon} \ln\left(\frac{LR^2}{\varepsilon} \right) \ln\left(\frac{\sqrt{L/\varepsilon}}{\beta} \right) \right\} \right) \quad (4.13)$$

which is optimal up to logarithmic factors. We call this modification Stochastic STM (SSTM). As for the deterministic case we summarize the state-of-the-art results for this case in Table 4.2.

4.3. Similar Triangles Method with Inexact Proximal Step

In this section we focus on the composite optimization problem. i.e. problems of the type

$$\min_{x \in \mathbb{R}^n} F(x) = f(x) + h(x), \quad (4.14)$$

Assumptions on f	Method	Citation	# of oracle calls
μ-strongly convex, L-smooth	R-SSTM	[16] [53] [17]	$\widetilde{O}\left(\max\left\{\sqrt{\frac{L}{\mu}}\ln\left(\frac{R^2}{\varepsilon}\right), \frac{\sigma^2}{\mu\varepsilon}\right\}\right)$
L-smooth	SSTM	[16] [53] [17]	$\widetilde{O}\left(\max\left\{\sqrt{\frac{LR^2}{\varepsilon}}, \frac{\sigma^2R^2}{\varepsilon^2}\right\}\right)$
μ-strongly convex, $\mathbf{E}_\xi\left[\|\nabla f(x;\xi)\|_2^2\right] \le M^2$	MD	[51] [52]	$O\left(\frac{M^2}{\mu\varepsilon}\right)$
$\mathbf{E}_\xi\left[\|\nabla f(x;\xi)\|_2^2\right] \le M^2$	MD	[51] [52]	$O\left(\frac{M^2R^2}{\varepsilon^2}\right)$

Table 4.2: Optimal (up to logarithmic factors) number of stochastic unbiased first-order oracle calls in order to get such a point x^N that $f(x^N) - f(x^*) \le \varepsilon$ with probability at least $1-\beta$, $\beta \in (0,1)$ and f is defined in (4.2). First column contains assumptions on f in addition to the convexity. Blue terms in the last column correspond to the number of iterations of the method.

where $f(x)$ is convex and L-smooth and $h(x)$ is convex and L_h-smooth. Before we present our method, let us introduce new notation.

Definition 4.3.1. *Assume that function $g(x)$ defined on $\mathbb{R}^n$ is such that there exists (possibly non-unique) x^* satisfying $g(x^*) = \min_{x\in\mathbb{R}^n} g(x)$. Then for arbitrary $\delta > 0$ we say that $\hat{x}$ is δ-solution of the problem $g(x) \to \min_{x\in\mathbb{R}^n}$ and write $\hat{x} = \operatorname{argmin}_{x\in\mathbb{R}^n} g(x)$ if $g(\hat{x}) - g(x^*) \le \delta$.*

Note that δ-solution could be non-unique, but for our purposes in such cases it is enough to use any point from the set of δ-solutions. In the analysis we will need the following result.

Lemma 4.3.1 (See also Theorem 9 from [48]). *Let $g(x)$ be convex, L-smooth, x^* is such that $g(x^*) = \min_{x\in\mathbb{R}^n} g(x)$ and $\hat{x} = \operatorname{argmin}_{x\in\mathbb{R}^n} g(x)$ for some $\delta > 0$. Then for all $x \in \mathbb{R}^n$*

$$\langle \nabla g(\hat{x}), \hat{x} - x\rangle \le \sqrt{2L\delta}\|\hat{x} - x\|_2. \tag{4.15}$$

The main method of this section is stated as Algorithm 2. In the STM_IPS we use functions $g_{k+1}(z)$ which are defined for all $k = 0, 1, \ldots$ as follows:

$$g_{k+1}(z) = \frac{1}{2}\|z^k - z\|_2^2 + \alpha_{k+1}\left(f(x^{k+1}) + \langle \nabla f(x^{k+1}), z - x^{k+1}\rangle + h(z)\right). \tag{4.16}$$

Algorithm 2 Similar Triangles Methods with Inexact Proximal Step (STM_IPS)

Require: $x^0 = z^0 = x^0$ — starting point, N — number of iterations

1: Set $\alpha_0 = A_0 = 0$

2: **for** $k = 0, 1, \ldots, N-1$ **do**

3: Choose α_{k+1} such that $A_k + \alpha_{k+1} = 2L\alpha_{k+1}^2$, $A_{k+1} = A_k + \alpha_{k+1}$

4: $x^{k+1} = (A_k x^k + \alpha_{k+1} z^k)/A_{k+1}$

5: $z^{k+1} = \operatorname{argmin}_{z \in \mathbb{R}^n}^{k+1} g_{k+1}(z)$, where $g_{k+1}(z)$ is defined in (4.16) and $\delta_{k+1} = \delta \|z^k - \hat{z}^{k+1}\|_2^2$

6: $x^{k+1} = (A_k x^k + \alpha_{k+1} z^{k+1})/A_{k+1}$

7: **end for**

Ensure: x^N

Each $g_{k+1}(z)$ is 1-strongly convex function with, as a consequence, unique minimizer $\hat{z}^{k+1} \stackrel{\text{def}}{=} \operatorname{argmin}_{z \in \mathbb{R}^n} g_{k+1}(z)$.

Let us discuss a little bit the proposed method. First of all, if we slightly modify the method and choose $\delta_{k+1} = 0$, then we will get STM which is well-studied in the literature. Secondly, it may seem that in order to run the method we need to know $\|z^k - \hat{z}^{k+1}\|_2$, but in fact we do not need it. If $g_{k+1}(z)$ is L_{k+1}-smooth and μ_{k+1}-strongly convex, then one can run STP for $T = O\left(\sqrt{L_{k+1}/\mu_{k+1}} \ln L_{k+1}/\varepsilon\right)$ iterations with z^k as a starting point to solve the problem $g_{k+1}(z) \to \min_{z \in \mathbb{R}^n}$ and get $z^{k+1} = \operatorname{argmin}_{z \in \mathbb{R}^n}^{k+1} g_{k+1}(z)$. Note that in this case we do not need to know $\hat{z}^{k+1}$. Moreover, we do not assume that iterates of STM_IPS are bounded and instead of assuming it we prove such result which makes the analysis a little bit more technical then ones for STP. Finally, we notice that one can prove the results we present below even with such α_{k+1} that $A_{k+1} = A_k + \alpha_{k+1} = L\alpha_{k+1}^2$. It improves numerical constants in the upper bounds a little bit, but for simplicity we use the same choice of α_{k+1} as for the stochastic case.

We start our analysis with the following lemma.

Lemma 4.3.2 (see also Theorem 1 from [54]). *Assume that $f(x)$ is convex and L-smooth, $h(x)$ is convex and L_h-smooth and $\delta < \frac{1}{2}$. Then after $N \geq 1$ iterations of Algorithm 2 we have*

$$A_N \left(F(x^N) - F(x^*)\right) \leq \frac{1}{2}R_0^2 - \frac{1}{2}R_N^2 + \delta \sum_{k=0}^{N-1} \sqrt{k+2}\,\widetilde{R}_{k+1}^2, \tag{4.17}$$

where x^* is the solution of (4.14) closest to the starting point z^0, $R_{k+1} \stackrel{\text{def}}{=} \|x^* - z^{k+1}\|_2$, $\widetilde{R}_0 \stackrel{\text{def}}{=}$

$R_0 \overset{def}{=} \|x^* - z^0\|_2$, $\widetilde{R}_{k+1} \overset{def}{=} \max\{\widetilde{R}_k, R_{k+1}\}$ *for* $k = 0, 1, \ldots, N-1$ *and* $\hat{\delta} \overset{def}{=} \sqrt{\frac{(L_h+2L)}{(1-\sqrt{\delta/2})^2 L}}$.

Below we state our main result of this section.

Theorem 4.3.1. *Let $f(x)$ be convex and L-smooth, $h(x)$ be convex and L_h-smooth and $\delta \le \frac{1}{4}$. Assume that for a given number of iterations $N \ge 1$ the number $\hat{\delta} \overset{def}{=} 2\sqrt{\frac{(L_h+2L)}{(1-\sqrt{\delta/2})^2 L}}$ satisfies $\hat{\delta} \le \frac{C}{(N+1)^{3/2}}$ with some positive constant $C \in (0, 1/4)$. Then after N iteration of Algorithm 2 we have*

$$F(x^N) - F(x^*) \le \frac{3R_0^2}{2A_N}. \tag{4.18}$$

Corollary 4.3.1. *Under assumptions of Theorem 4.3.1 we get that for an arbitrary $\varepsilon > 0$ after*

$$N = O\left(\sqrt{\frac{LR_0^2}{\varepsilon}}\right) \tag{4.19}$$

iterations of Algorithm 2 we have $F(x^N) - F(x^*) \le \varepsilon$. Moreover, we get that δ should satisfy

$$\delta = O\left(\frac{L}{(L_h + L)N^3}\right). \tag{4.20}$$

That is, if the auxiliary problem $g_{k+1}(z) \to \min_{z \in \mathbb{R}^n}$ is solved with good enough accuracy, then STM_IPS requires the same number of iterations as STM to achieve $F(x^N) - \min_{x \in \mathbb{R}^n} F(x) \le \varepsilon$.

Finally, we notice that one can set δ_{k+1} in Algorithm 2 in a different way in order to get the same convergence guarantees, e.g. one can use $\delta_{k+1} = \delta\widetilde{R}_{k+1}^2$ and the order of δ given by (4.20) will be the same. In this case inequalities (4.128) and (4.130) transform to

$$\langle g_{k+1}(z^{k+1}), z^{k+1} - x^* \rangle \le \sqrt{2(\alpha_{k+1}L_h + 1)\delta\widetilde{R}_{k+1}^2} \cdot \|z^{k+1} - x^*\|_2$$

and

$$\langle z^{k+1} - z^k + \alpha_{k+1}\nabla f(x^{k+1}) + \alpha_{k+1}\nabla h(z^{k+1}), z^{k+1} - x^* \rangle \le \hat{\delta}\sqrt{k+2}\widetilde{R}_{k+1}^2,$$

respectively, where $\hat{\delta} \overset{def}{=} 2\sqrt{\frac{(L_h+2L)}{L}}$. Then the remaining part of the proof remains the same and gives the same result up to small changes in the numerical constants.

4.4. Stochastic Convex Optimization with Affine Constraints: Primal Approach

Now, we are going to make the next step towards decentralized distributed optimization and consider convex optimization problem with affine constraints:

$$\min_{Ax=0;x\in Q} f(x), \tag{4.21}$$

where $A \neq 0$ and $\mathrm{Ker}A \not\supseteq \{0\}$. Up to a sign we can define the dual problem in the following way

$$\min_y \psi(y), \quad \text{where} \tag{4.22}$$

$$\varphi(y) = \max_{x\in Q} \{\langle y, x\rangle - f(x)\}, \tag{4.23}$$

$$\psi(y) = \varphi(A^\top y) = \max_{x\in Q} \{\langle y, Ax\rangle - f(x)\} = \langle A^\top y, x(A^\top y)\rangle - f(x(A^\top y)), \tag{4.24}$$

where $x(y) \stackrel{\text{def}}{=} \mathrm{argmax}_{x\in Q} \{\langle y, x\rangle - f(x)\}$. Since $\mathrm{Ker}A \not\supseteq \{0\}$ the solution of the dual problem (4.22) is not unique. We use y^* to denote the solution of (4.22) with the smallest ℓ_2-norm $R_y \stackrel{\text{def}}{=} \|y^*\|_2$.

However, in this section we are interested only in primal approaches to solve (4.21) and, in particular, the main goal of this section is to present first-order methods that are optimal both in terms of $\nabla f(x)$ and $A^\top Ax$ calculations. Before we start our analysis let us notice that typically in decentralized optimization matrix A from (4.21) is chosen as a square root of Laplacian matrix W of communication network [29] (see Section 4.6 for the details). In asynchronous case the square root $\sqrt{W}$ is replaced by incidence matrix M [55] ($W = M^\top M$). Then in asynchronous case instead of accelerated methods for (4.22) one should use accelerated block-coordinate descent methods [15, 55–57].

To solve problem (4.21) we use the following trick [45, 49]: instead of (4.21) we consider penalized problem

$$\min_{x\in Q} F(x) = f(x) + \frac{R_y^2}{\varepsilon}\|Ax\|_2^2, \tag{4.25}$$

where $\varepsilon > 0$ is the desired accuracy of the solution in terms of $f(x)$ that we want to achieve. The motivation behind this trick is revealed in the following theorem.

Theorem 4.4.1 (See also Remark 4.3 from [49]). *Assume that $x^N \in Q$ is such that*

$$F(x^N) - \min_{x\in Q} F(x) \le \varepsilon. \tag{4.26}$$

Then

$$f(x^N) - \min_{Ax=0;x\in Q} f(x) \le \varepsilon, \quad \|Ax^N\|_2 \le \frac{2\varepsilon}{R_y}. \tag{4.27}$$

We start with the analysis of the case when f is L-smooth and convex.

Theorem 4.4.2. *Let f be convex and L-smooth, $Q = \mathbb{R}^n$ and $h(x) = R_y^2\|Ax\|_2^2/\varepsilon$. Assume that full gradients of f and h are available. Then* STM_IPS *(see Algorithm 2, Section 4.3) applied to solve problem* (4.25) *requires*

$$O\left(\sqrt{\frac{LR^2}{\varepsilon}}\right) \quad \text{calculations of } \nabla f(x), \tag{4.28}$$

$$\widetilde{O}\left(\sqrt{\frac{LR^2}{\varepsilon}\chi(A^\top A)}\right) \quad \text{calculations of } A^\top Ax \tag{4.29}$$

to produce point x^N such that (4.26) *holds.*

That is, number of $A^\top Ax$ calculations matches the optimal bound for deterministic convex and L-smooth problems of type (4.1) multiplied by $\sqrt{\chi(A^\top A)}$ up to logarithmic factors (see Table 4.1).

We believe that using the same recurrence technique that we use in Sections 4.3 and 4.5 one can generalize this result for the case when instead of $\nabla f(x)$ only stochastic gradient $\nabla f(x,\xi)$ (see inequalities (4.7)-(4.8)) is available. To the best of our knowledge it is not done in the literature for the case when $Q = \mathbb{R}^n$. Moreover, it is also possible to extend our approach to handle strongly convex case via variants of STM.

We conjecture that the same technique in the case when f is μ-strongly convex and L-smooth gives the method that requires such number of $A^\top Ax$ calculations that matches the second rows of Tables 4.1 and 4.2 in the corresponding cases with additional factor $\sqrt{\chi(A^\top A)}$ and logarithmic factors. Recently such bounds were shown in [58] for the distributed version of Multistage Accelerated Stochastic Gradient method from [59]. However, this bounds were shown for the case when the stochastic gradient is unbiased.

Next, we assume that Q is closed and convex and f is μ-strongly convex, but possibly non-smooth function with bounded gradients: $\|\nabla f(x)\|_2 \le M$ for all $x \in Q$. Let us start with the case $\mu = 0$. Then, to achieve (4.26) one can run Sliding method from [53, 60] considering $f(x)$ as a composite term. In this case Sliding requires

$$O\left(\sqrt{\frac{\lambda_{\max}(A^\top A)R_y^2R^2}{\varepsilon^2}}\right) \text{ calculations of } A^\top Ax, \tag{4.30}$$

$$O\left(\frac{M^2R^2}{\varepsilon^2}\right) \text{ calculations of } \nabla f(x). \tag{4.31}$$

In the case when Q is a compact set and $\nabla f(x)$ is not available and unbiased stochastic gradient $\nabla f(x,\xi)$ is used instead (see inequalities (4.7)-(4.8) with $\delta = 0$) one can show [53, 60] that Stochastic Sliding (S-Sliding) method can achieve (4.26) with probability at least $1-\beta$, $\beta \in (0,1)$, and it requires the same number of calculations of $A^\top Ax$ as in (4.30) up to logarithmic factors and

$$\widetilde{O}\left(\frac{(M^2+\sigma^2)R^2}{\varepsilon^2}\right) \text{ calculations of } \nabla f(x,\xi). \tag{4.32}$$

When $\mu > 0$ one can apply restarts technique on top of S-Sliding (RS-Sliding) [45, 61] and get that to guarantee (4.26) with probability at least $1-\beta$, $\beta \in (0,1)$ RS-Sliding requires

$$\widetilde{O}\left(\sqrt{\frac{\lambda_{\max}(A^\top A)R_y^2}{\mu\varepsilon}}\right) \text{ calculations of } A^\top Ax, \tag{4.33}$$

$$\widetilde{O}\left(\frac{M^2+\sigma^2}{\mu\varepsilon}\right) \text{ calculations of } \nabla f(x,\xi). \tag{4.34}$$

We notice that bounds presented above for the non-smooth case are proved only for the case when Q is bounded. For the case of unbounded Q the convergence results with such rates were proved only in expectation. Moreover, it would be interesting to study S-Sliding and RS-Sliding in the case when $\delta > 0$, i.e. stochastic gradient is biased, but we leave these questions for future works.

4.5. Stochastic Convex Optimization with Affine Constraints: Dual Approach

In this section we assume that one can construct a dual problem for (4.21). If f is μ-strongly convex in ℓ_2-norm, then ψ and φ have L –Lipschitz continuous and L'–Lipschitz continuous in ℓ_2-norm gradients respectively [62, 63], where $L = \lambda_{\max}(A^\top A)/\mu$ and $L' = 1/\mu$. In our proofs we often use Demyanov–Danskin theorem [63] which states that

$$\nabla\psi(y) = Ax(A^\top y), \quad \nabla\varphi(y) = x(y). \tag{4.35}$$

We notice that in this section we do not assume that A is symmetric or positive semidefinite.

Below we propose a primal-dual method for the case when f is additionally Lipschitz continuous on some ball and two methods for the problems when the primal function is also L-smooth and Lipschitz continuous on some ball. In the subsections below we assume that $Q = \mathbb{R}^n$.

4.5.1. Convex Dual Function

In this section we assume that the dual function $\varphi(y)$ could be rewritten as an expectation, i.e. $\varphi(y) = \mathbf{E}\ [\varphi(y,\xi)]$, where stochastic realisations $\varphi(y,\xi)$ are differentiable in y functions almost surely in ξ. Then, we can also represent $\psi(y)$ as an expectation: $\psi(y) = \mathbf{E}\ [\psi(y,\xi)]$. Consider the stochastic function $f(x,\xi)$ which is defined implicitly as follows:

$$\varphi(y,\xi) = \max_{x \in \mathbb{R}^n} \{\langle y, x\rangle - f(x,\xi)\}. \tag{4.36}$$

Similarly to the deterministic case we introduce $x(y,\xi) \stackrel{\text{def}}{=} \operatorname{argmax}_{x \in \mathbb{R}^n} \{\langle y, x\rangle - f(x,\xi)\}$ which satisfies $\nabla \varphi(y,\xi) = x(y,\xi)$ due to Demyanov-Danskin theorem, where the gradient is taken w.r.t. y. As a simple corollary, we get $\nabla \psi(y,\xi) = Ax(A^\top y)$. Finally, introduced notations and obtained relations imply that $x(y) = \mathbf{E}\ [x(y,\xi)]$ and $\nabla \psi(y) = \mathbf{E}\ [\nabla \psi(y,\xi)]$.

Consider the situation when $x(y,\xi)$ is known only through the noisy observations $\tilde{x}(y,\xi) = x(y,\xi) + \delta(y,\xi)$ and assume that the noise is bounded in expectation, i.e. there exists non-negative deterministic constant $\delta_y \geq 0$, such that

$$\|\mathbf{E}\ [\delta(y,\xi)]\|_2 \leq \delta_y, \quad \forall y \in \mathbb{R}^n. \tag{4.37}$$

Assume additionally that $x(y,\xi)$ satisfies so-called "light-tails" inequality:

$$\mathbf{E}\ \left[\exp\left(\frac{\|\tilde{x}(y,\xi) - \mathbf{E}\ [\tilde{x}(y,\xi)]\|_2^2}{\sigma_x^2}\right)\right] \leq \exp(1), \quad \forall y \in \mathbb{R}^n, \tag{4.38}$$

where σ_x is some positive constant. It implies that we have an access to the biased gradient $\tilde{\nabla} \psi(y,\xi) \stackrel{\text{def}}{=} A\tilde{x}(y,\xi)$ which satisfies following relations:

$$\left\|\mathbf{E}\ \left[\tilde{\nabla} \psi(y,\xi)\right] - \nabla \psi(y)\right\|_2 \leq \delta, \quad \forall y \in \mathbb{R}^n, \tag{4.39}$$

$$\mathbf{E}\ \left[\exp\left(\frac{\left\|\tilde{\nabla} \psi(y,\xi) - \mathbf{E}\ \left[\tilde{\nabla} \psi(y,\xi)\right]\right\|_2^2}{\sigma^2}\right)\right] \leq \exp(1), \quad \forall y \in \mathbb{R}^d, \tag{4.40}$$

where $\delta \stackrel{\text{def}}{=} \sqrt{\lambda_{\max}(A^\top A)}\delta_y$ and $\sigma \stackrel{\text{def}}{=} \sqrt{\lambda_{\max}(A^\top A)}\sigma_x$. We will use $\nabla(y, \xi^k)$ to denote batched stochastic gradient:

$$\nabla(y, \xi^k) = \frac{1}{r_k}\sum_{l=1}^{r_k} \nabla\psi(y, \xi^l), \quad \nabla x(y, \xi^k) = \frac{1}{r_k}\sum_{l=1}^{r_k} \nabla x(y, \xi^l) \tag{4.41}$$

The size of the batch r_k could always be restored from the context, so, we do not specify it here. Note that the batch version satisfies

$$\left\|\mathbf{E}\left[\nabla(x, \xi^k)\right] - \nabla\psi(x)\right\|_2 \leq \delta, \quad \forall x \in \mathbb{R}^n, \tag{4.42}$$

$$\mathbf{E}\left[\exp\left(\frac{\left\|\nabla(x, \xi^k) - \mathbf{E}\left[\nabla(x, \xi^k)\right]\right\|_2^2}{O(\sigma_\psi^2/r_k^2)}\right)\right] \leq \exp(1), \quad \forall x \in \mathbb{R}^n, \tag{4.43}$$

where in the last inequality we used combination of Lemmas 4.9.3 and 4.9.5 (see two inequalities after (4.161) for the details). We call this approach SPDSTM (Stochastic Primal-Dual Similar Triangles Method, see Algorithm 3). Note that Algorithm 4 from [46] is a special case of SPDSTM when $\delta = 0$, i.e. stochastic gradient is unbiased, up to a factor 2 in the choice of $\tilde{L}$.

Algorithm 3 SPDSTM

Require: $\hat{y}^0 = z^0 = y^0 = 0$, number of iterations N, $\alpha_0 = A_0 = 0$

1: **for** $k = 0, \ldots, N$ **do**

2: Set $\tilde{L} = 2L$

3: Set $A_{k+1} = A_k + \alpha_{k+1}$, where $2\tilde{L}\alpha_{k+1}^2 = A_k + \alpha_{k+1}$

4: $\hat{y}^{k+1} = (A_k y^k + \alpha_{k+1} z^k)/A_{k+1}$

5: $z^{k+1} = z^k - \alpha_{k+1}\nabla(\hat{y}^{k+1}, \xi^k)$

6: $y^{k+1} = (A_k y^k + \alpha_{k+1} z^{k+1})/A_{k+1}$

7: **end for**

Ensure: y^N, $x^N = \frac{1}{A_N}\sum_{k=0}^N \alpha_k \nabla x(A^\top \hat{y}^k, \xi^k)$.

Below we present the main convergence result of this section.

Theorem 4.5.1 (see also Theorem 2 from [46]). *Assume that f is μ-strongly convex and $\|\nabla f(x^*)\|_2 = M_f$. Let $\varepsilon > 0$ be a desired accuracy. Next, assume that f is L_f-Lipschitz continuous on the ball $B_{R_f}(0)$ with $R_f = \tilde{\Theta}\left(\max\left\{\frac{\sqrt{\lambda_{\max}(A^\top A)}R_y}{A_N}, \sqrt{\lambda_{\max}(A^\top A)}R_y, R_x\right\}\right)$, where R_y is such that $\|y^*\|_2 \leq R_y$, y^* is the solution of the dual problem (4.22), and*

$R_x = \|x(A^\top y^*)\|_2$. *Assume that at iteration k of Algorithm 3 batch size is chosen according to the formula* $r_k \geq \max\left\{1, \frac{\widetilde{\sigma}_k^2 \ln(N/\beta)}{\hat{C}\varepsilon}\right\}$, *where* $\widetilde{\alpha}_k = \frac{k+1}{2L}$, $0 < \varepsilon \leq \frac{HLR_0^2}{N^2}$, $0 \leq \delta \leq \frac{GLR_0}{(N+1)^2}$ *and* $N \geq 1$ *for some numeric constant* $H > 0$, $G > 0$ *and* $\hat{C} > 0$. *Then with probability* $\geq 1 - 4\beta$, *where* $\beta \in (0, 1/4)$ *is such that* $\frac{1+\sqrt{\ln\frac{1}{\beta}}}{\sqrt{\ln\frac{N}{\beta}}} \leq 2$, *after* $N = \widetilde{O}\left(\sqrt{\frac{M_f}{\varepsilon}\chi(A^\top A)}\right)$ *iterations where* $\chi(A^\top A) = \frac{\max(A^\top A)}{\lambda^+_{\min}(A^\top A)}$, *the outputs* x^N *and* y^N *of Algorithm 3 satisfy the following condition*

$$f(x^N) - f(x^*) \leq f(x^N) + \psi(y^N) \leq \varepsilon, \quad \|Ax^N\|_2 \leq \frac{\varepsilon}{R_y} \tag{4.44}$$

with probability at least $1 - 4\beta$. *What is more, to guarantee* (4.44) *with probability at least* $1 - 4\beta$ *Algorithm 3 requires*

$$\widetilde{O}\left(\max\left\{\frac{\sigma_x^2 M_f^2}{\varepsilon^2}\chi(A^\top A)\ln\left(\frac{1}{\beta}\sqrt{\frac{M_f}{\mu\varepsilon}\chi(A^\top A)}\right), \sqrt{\frac{M_f}{\mu\varepsilon}\chi(A^\top A)}\right\}\right) \tag{4.45}$$

calls of the biased stochastic oracle $\nabla\psi(y, \xi)$, *i.e.* $x(y, \xi)$.

4.5.2. Strongly Convex Dual Functions and Restarts Technique

In this section we assume that primal functional f is additionally L-smooth. It implies that the dual function ψ in (4.22) is additionally μ -strongly convex in $y^0 + (\mathrm{Ker}A^\top)^\perp$ where $\mu = \lambda^+_{\min}(A^\top A)/L$ [62, 63] and $\lambda^+_{\min}(A^\top A)$ is the minimal positive eigenvalue of $A^\top A$.

From weak duality $-f(x^*) \leq \psi(y^*)$ and (4.24) we get the key relation of this section (see also [64–66])

$$f(x(A^\top y)) - f(x^*) \leq \langle \nabla\psi(y), y\rangle = \langle Ax(A^\top y), y\rangle \tag{4.46}$$

This inequality implies the following theorem.

Theorem 4.5.2. *Consider function* f *and its dual function* ψ *defined in* (4.24) *such that problems* (4.21) *and* (4.22) *have solutions. Assume that* y^N *is such that* $\|\nabla\psi(y^N)\|_2 \leq \varepsilon/R_y$ *and* $\|y^N\| \leq 2R_y$, *where* $\varepsilon > 0$ *is some positive number and* $R_y = \|y^*\|_2$ *where* y^* *is any minimizer of* ψ. *Then for* $x^N = x(A^\top y^N)$ *following relations hold:*

$$f(x^N) - f(x^*) \leq 2\varepsilon, \quad \|Ax^N\|_2 \leq \frac{\varepsilon}{R_y}, \tag{4.47}$$

where x^* *is any minimizer of* f.

Proof. Applying Cauchy-Schwarz inequality to (4.46) we get

$$f(x^N) - f(x^*) \overset{(4.46)}{\leq} \|r\psi(y^N)\|_2 \cdot \|y^N\|_2 \leq \frac{\varepsilon}{R_y} \cdot 2R_y = 2\varepsilon.$$

The second part (4.47) immediately follows from $\|r\psi(y^N)\|_2 \leq \varepsilon/R_y$ and Demyanov-Danskin theorem which implies $r\psi(y^N) = Ax^N$. □

That is why, in this section we mainly focus on the methods that provides optimal convergence rates for the gradient norm. In particular, we consider Recursive Regularization Meta-Algorithm from (see Algorithm 4) [67] with AC-SA^2 (see Algorithm 6) as a subroutine (i.e. RRMA-AC-SA^2) which is based on AC-SA algorithm (see Algorithm 5) from [68]. We notice that RRMA-AC-SA^2 is applied for a regularized dual function

$$\tilde{\psi}(y) = \psi(y) + \frac{\lambda}{2}\|y - y^0\|_2^2, \tag{4.48}$$

where $\lambda > 0$ is some positive number which will be defined further. Function $\tilde{\psi}$ is λ-strongly convex and $\tilde{L}$ -smooth in $\mathbb{R}^n$ where $\tilde{L} = L + \lambda$. For now, we just assume w.l.o.g. that $\tilde{\psi}$ is $(\mu + \lambda)$-strongly convex in $\mathbb{R}^n$, but we will go back to this question further.

In this section we consider the same oracle as in Section 4.5, but we additionally assume that $\delta = 0$, i.e. stochastic first-order oracle is unbiased. To define batched version of the stochastic gradient we will use the following notation:

$$r(y, \{\xi^l\}_{l=1}^{r_t}) = \frac{1}{r_t}\sum_{l=1}^{r_t} r\psi(y, \xi^l), \quad x(y, \{\xi^l\}_{l=1}^{r_t}) = \frac{1}{r_t}\sum_{l=1}^{r_t} x(y, \xi^l). \tag{4.49}$$

As before in the cases when the batch-size r_t can be restored from the context, we will use simplified notation $r(y, \{\xi^l\}_{l=1}^t)$ and $x(y, \{\xi^l\}_{l=1}^t)$. In the AC-SA algorithm we use batched

Algorithm 4 RRMA-AC-SA^2 [67]

Require: y^0 — starting point, m — total number of iterations

1: $\psi_0 \equiv \tilde{\psi}$, $\mathcal{Y}^0 \equiv y^0$, $T = \left\lfloor \log_2 \frac{L_\psi}{\lambda} \right\rfloor$

2: **for** $k = 1, \dots, T$ **do**

3: Run AC-SA^2 for m/T iterations to optimize ψ_{k-1} with $\mathcal{Y}^{k-1}$ as a starting point and get the output $\mathcal{Y}^k$

4: $\psi_k(y) \equiv \tilde{\psi}(y) + \lambda \sum_{l=1}^k 2^{l-1}\|y - \mathcal{Y}^l\|_2^2$

5: **end for**

Ensure: $\mathcal{Y}^T$.

stochastic gradients of functions ψ_k which are defined as follows:

$$r \quad _k(y, \quad ^t) \quad = \quad \frac{1}{r_t}\sum_{l=1}^{r_t} r\,\psi_k(y,\xi^l), \tag{4.50}$$

$$r\,\psi_k(y,\xi) \quad = \quad r\,\psi(y,\xi) + \lambda(y - y^0) + \lambda\sum_{l=1}^{k} 2^l(y - \hat{y}^l).$$

Algorithm 5 AC-SA [68]

Require: z^0 — starting point, m — number of iterations, ψ_k — objective function

1: $y_{ag}^0 \quad z^0$, $y_{md}^0 \quad z^0$

2: **for** $t = 1, \ldots, m$ **do**

3: $\quad \alpha_t \quad \frac{2}{t+1}$, $\gamma_t \quad \frac{4L_\psi}{t(t+1)}$

4: $\quad y_{md}^t \quad \frac{(1- \ _t)(\ + \ _t)}{\ _t+(1- \ _t^2)} y_{ag}^{t-1} + \frac{\ _t((1- \ _t)\ + \ _t)}{\ _t+(1- \ _t^2)} z^{t-1}$

5: $\quad z^t \quad \frac{\ _t}{\ + \ _t} y_{md}^t + \frac{(1- \ _t)\ + \ _t}{\ + \ _t} z^{t-1} - \frac{\ _t}{\ + \ _t} r \quad _k(y_{md}^t, \quad ^t)$

6: $\quad y_{ag}^t \quad \alpha_t z^t + (1 - \alpha_t) x_{ag}^{t-1}$

7: **end for**

Ensure: y_{ag}^m.

Algorithm 6 AC-SA^2 [67]

Require: z^0 — starting point, m — number of iterations, ψ_k — objective function

1: Run AC-SA for $m/2$ iterations to optimize ψ_k with z^0 as a starting point and get the output y^1

2: Run AC-SA for $m/2$ iterations to optimize ψ_k with y^1 as a starting point and get the output y^2

Ensure: y^2.

The following theorem states the main result for RRMA-AC-SA^2 that we need in the section.

Theorem 4.5.3 (Corollary 1 from [67]). *Let ψ be L -smooth and μ -strongly convex function and $\lambda = \left(\left(L_\psi \ln^2 N\right)/N^2\right)$ for some $N > 1$. If the Algorithm 4 performs N iterations in total[2] with batch size r for all iterations, then it will provide such a point $\hat{y}$ that*

$$\mathbf{E}\left[k\,r\,\psi(\hat{y})k_2^2 \,j\, y^0, r\right] \le C\left(\frac{L^2 k y^0 - y^* k_2^2 \ln^4 N}{N^4} + \frac{\sigma^2 \ln^6 N}{rN}\right), \tag{4.51}$$

[2] It means that the overall number of performed iterations preformed during the calls of AC-SA^2 equals N.

where $C > 0$ is some positive constant and y^* is a solution of the dual problem (4.22).

Let us show that w.l.o.g. we can assume in this section that function ψ defined in (4.24) is μ -strongly convex everywhere with $\mu = \lambda_{\min}^{+}(AA^\top)/L$. In fact, from L-smoothness of f we have only that ψ is μ -strongly convex in $y^0 + \left(\mathrm{Ker}(A^\top)\right)^\perp$ (see [62, 63] for the details). However, the structure of the considered here methods is such that all points generated by the RRMA-AC-SA^2 and, in particular, AC-SA lie in $y^0 + \left(\mathrm{Ker}(A^\top)\right)^\perp$.

Theorem 4.5.4. *Assume that Algorithm 5 is run for the objective $\psi_k(y) = \tilde{\psi}(y) + \lambda \sum_{l=1}^{k} 2^{l-1}\|y - \hat{y}^l\|_2^2$ with z^0 as a starting point, where $z^0, \hat{y}^1, \ldots, \hat{y}^k$ are some points from $y^0 + \left(\mathrm{Ker}(A^\top)\right)^\perp$ and $y^0 \in \mathbb{R}^n$. Then for all $t \geq 0$ we have $y_{md}^t, z^t, y_{ag}^t \in y^0 + \left(\mathrm{Ker}(A^\top)\right)^\perp$.*

Proof. We prove the statement of the theorem by induction. For $t = 0$ the statement is trivial, since $y_{md}^0 = y_{ag}^0 = z^0 \in y_0 + \left(\mathrm{Ker}(A^\top)\right)^\perp$. Assume that $y_{md}^t, z^t, y_{ag}^t \in y^0 + \left(\mathrm{Ker}(A^\top)\right)^\perp$ for some $t \geq 0$ and prove it for $t + 1$. Since $y_0 + \left(\mathrm{Ker}(A^\top)\right)^\perp$ is a convex set and y_{md}^{t+1} is a convex combination of y_{ag}^t and z^t we have $y_{md}^{t+1} \in y^0 + \left(\mathrm{Ker}(A^\top)\right)^\perp$. Next, the point $\frac{\alpha_t}{\alpha_t + \gamma_t} y_{md}^{t+1} + \frac{(1-\alpha_t)\alpha_t + \gamma_t}{\alpha_t + \gamma_t} z^t$ also lies in $y^0 + \left(\mathrm{Ker}(A^\top)\right)^\perp$ since it is convex combination of the points lying in this set. Due to (4.48), (4.49) and (4.50) we have that $r_{\beta_k}(y_{md}^{t+1}, \gamma_t) = Ax(A^\top y_{md}^{t+1}, \gamma_t) + \lambda(y_{md}^{t+1} - y^0) + \lambda \sum_{l=1}^{k} 2^l (y_{md}^{t+1} - \hat{y}^l)$. The first term lies in $\left(\mathrm{Ker}(A^\top)\right)^\perp$ since $\mathrm{Im}(A) = \left(\mathrm{Ker}(A^\top)\right)^\perp$ and the second and the third terms also lie in $\left(\mathrm{Ker}(A^\top)\right)^\perp$ since $y_{md}^{t+1}, y^0, \hat{y}^1, \ldots, \hat{y}^k \in y^0 + \left(\mathrm{Ker}(A^\top)\right)^\perp$. Putting all together we get $z^{t+1} \in y^0 + \left(\mathrm{Ker}(A^\top)\right)^\perp$. Finally, y_{ag}^{t+1} lies in $y^0 + \left(\mathrm{Ker}(A^\top)\right)^\perp$ as a convex combination of points from this set. □

Corollary 4.5.1. *Assume that Algorithm 4 is run for the objective $\psi_k(y) = \tilde{\psi}(y) + \lambda \sum_{l=1}^{k} 2^{l-1}\|y - \hat{y}^l\|_2^2$ with y^0 as a starting point. Then for all $k \geq 0$ we have $\hat{y}^k \in y^0 + \left(\mathrm{Ker}(A^\top)\right)^\perp$.*

Proof. We prove this result by induction. For $t = 0$ the statement is trivial since $\hat{y}^0 = y^0$. Next, assume that $\hat{y}^0, \hat{y}^1, \ldots, \hat{y}^k \in y^0 + \left(\mathrm{Ker}(A^\top)\right)^\perp$ and prove that $\hat{y}^{k+1} \in y^0 + \left(\mathrm{Ker}(A^\top)\right)^\perp$. Our assumption implies that the assumptions from Theorem 4.5.4 and applying the result of the theorem we get that y^1 and y^2 from the method AC-SA^2 applied to the ψ_k also lie in $y^0 + \left(\mathrm{Ker}(A^\top)\right)^\perp$. That is, the output of AC-SA^2 applied for ψ_k lies in $y^0 + \left(\mathrm{Ker}(A^\top)\right)^\perp$. □

Now we are ready to present our approach which was sketched in [45] of constructing an accelerated method for the strongly convex dual problem using restarts of RRMA-AC-SA^2.

To explain the main idea we start with the simplest case: $\sigma^2 = 0$, $r = 0$. It means that there is no stochasticity in the method and the bound (4.51) can be rewritten in the following form:

$$\|\nabla\psi(\hat{y})\|_2 \le \frac{\sqrt{2}\overline{C}L\,\|y^0 - y^*\|_2 \ln^2 N}{N^2} \le \frac{\sqrt{2}\overline{C}L\,\|\nabla\psi(y^0)\|_2 \ln^2 N}{\mu\,N^2}, \tag{4.52}$$

where we used inequality $\|\nabla\psi(y^0)\| \ge \mu\,\|y^0 - y^*\|$ which follows from the μ -strong convexity of ψ. It implies that after $N = \mathcal{O}(\sqrt{L_\psi/\mu})$ iterations of RRMA-AC-SA^2 the method returns such $y^1 = \hat{y}$ that $\|\nabla\psi(y^1)\|_2 \le \frac{1}{2}\|\nabla\psi(y^0)\|_2$. Next, applying RRMA-AC-SA^2 with y^1 as a starting point for the same number of iterations we will get new point y^2 such that $\|\nabla\psi(y^2)\|_2 \le \frac{1}{2}\|\nabla\psi(y^1)\|_2 \le \frac{1}{4}\|\nabla\psi(y^0)\|_2$. Then, after $l = O(\ln(R_y\|\nabla\psi(y^0)\|_2/\varepsilon))$ of such restarts we can get the point y^l such that $\|\nabla\psi(y^l)\|_2 \le \varepsilon/R_y$ with total number of gradients computations $Nl = \mathcal{O}\left(\sqrt{L_\psi/\mu}\ln(R_y\|\nabla\psi(y^0)\|_2/\varepsilon)\right)$.

In the case when $\sigma^2 \neq 0$ we need to modify this approach. The first ingredient to handle the stochasticity is large enough batch size for the l-th restart: r_l should be $\left(\sigma^2/(\bar{N}\|\nabla\psi(\bar{y}^{l-1})\|_2^2)\right)$. However, in the stochastic case we do not have an access to the $\nabla\psi(y^{l-1})$, so, such batch size is impractical. One possible way to fix this issue is to independently sample large enough number $\hat{r}_l \sim R_y^2/\varepsilon^2$ of stochastic gradients additionally, which is the second ingredient of our approach, in order to get good enough approximation $\nabla\psi(y^{l-1}, \ldots, \hat{r}_l)$ of $\nabla\psi(y^{l-1})$ and use the norm of such an approximation which is close to the norm of the true gradient with big enough probability in order to estimate needed batch size r^l for the optimization procedure. Using this, we can get the bound of the following form:

$$\mathbf{E}\left[\|\nabla\psi(y^l)\|_2^2 \mid y^{l-1}, r_l, \hat{r}_l\right] \le A_l \stackrel{\text{def}}{=} \frac{\|\nabla\psi(y^{l-1})\|_2^2}{8} + \frac{\|\nabla\psi(y^{l-1}, \ldots, \hat{r}_l) - \nabla\psi(y^{l-1})\|_2^2}{32}.$$

The third ingredient is the amplification trick: we run $p_l = \ln(1/\beta)$ independent trajectories of RRMA-AC-SA^2, get points $y^{l;1}, \ldots, y^{l;p_l}$ and choose such $y^{l;p(l)}$ among of them that $\|\nabla\psi(y^{l;p(l)})\|_2$ is *close enough* to $\min_{p=1,\ldots,p_l}\|\nabla\psi(y^{l;p})\|_2$ with high probability, i.e. $\|\nabla\psi(y^{l;p(l)})\|_2^2 \le 2\min_{p=1,\ldots,p_l}\|\nabla\psi(y^{l;p})\|_2^2 + \varepsilon^2/8R_y^2$ with probability at least $1-\beta$ for fixed $\nabla\psi(y^{l-1}, \ldots, \hat{r}_l)$. We achieve it due to additional sampling of $r_l \sim R_y^2/\varepsilon^2$ stochastic gradients at $y^{l;p}$ for each trajectory and choosing such $p(l)$ corresponding to the smallest norm of the obtained batched stochastic gradient. By Markov's inequality for all $p = 1, \ldots, p_l$

$$\mathbf{P}\left\{\|\nabla\psi(y^{l;p})\|_2^2 \ge 2A_l \mid y^{l-1}, r_l, r_l\right\} \le \frac{1}{2},$$

hence

$$\mathbf{P}\left\{\min_{p=1,\dots,p_l} \|\nabla\psi(y^{l,p})\|_2^2 \geq 2A_l\,|\,y^{l-1}, r_l, r_l\right\} \leq \frac{1}{2^{p_l}}.$$

That is, for $p_l = \log_2(1/\ )$ we have that with probability at least $1-2\beta$

$$\|\nabla\psi(y^{l,p(l)})\|_2^2 \leq \frac{\|\nabla\psi(y^{l-1})\|_2^2}{2} + \frac{\|\nabla\ \ (y^{l-1},\ \ ^{l-1}, \hat{f}_l) - \nabla\psi(y^{l-1})\|_2^2}{8} + \frac{\varepsilon^2}{8R_y^2}$$

for fixed $\nabla\ \ (y^{l-1},\ \ ^{l-1}, \hat{f}_l)$ which means that

$$\|\nabla\psi(y^{l,p(l)})\|_2^2 \leq \frac{\|\nabla\psi(y^{l-1})\|_2^2}{2} + \frac{\varepsilon^2}{4R_y^2}$$

with probability at least $1-3\beta$. Therefore, after $l = \log_2(2R_y^2\|\nabla\ \ (y^0)\|_2^2/\varepsilon^2)$ of such restarts our method provide the point $y^{l,p(l)}$ such that with probability at least $1-3l\beta$

$$\|\nabla\psi(y^{l,p(l)})\|_2^2 \leq \frac{\|\nabla\psi(y^0)\|_2^2}{2^l} + \frac{\varepsilon^2}{4R_y^2}\sum_{k=0}^{l-1} 2^{-k} \leq \frac{\varepsilon^2}{2R_y^2} + \frac{\varepsilon^2}{4R_y^2}\cdot 2 = \frac{\varepsilon^2}{R_y^2}.$$

The approach informally described above is stated as Algorithm 7.

Algorithm 7 Restarted-RRMA-AC-SA^2

Require: y^0 — starting point, l — number of restarts, $\{\hat{f}_k\}_{k=1}^l$, $\{r_k\}_{k=1}^l$ — batch-sizes, $\{p_k\}_{k=1}^l$ — amplification parameters

1: Choose the smallest integer $N > 1$ such that $\frac{CL_\psi^2 \ln^4 \bar{N}}{\mu_\psi^2 N^4} \leq \frac{1}{32}$

2: $y^{0,p(0)} \quad y^0$

3: **for** $k = 1, \dots, l$ **do**

4: Compute $\nabla\ \ (y^{k-1,p(k-1)},\ \ ^{k-1,p(k-1)}, \hat{f}_k)$

5: $r_k \quad \max\left\{1, \frac{64C\ \ _\psi^2 \ln^6 \bar{N}}{\bar{N}\|\nabla\ \ (\bar{y}^{k-1,p(k-1)};\ \ ^{k-1,p(k-1)}; \hat{f}_k)\|_2^2}\right\}$

6: Run p_k independent trajectories of RRMA-AC-SA^2 for N iterations with batch-size r_k with $y^{k-1,p(k-1)}$ as a starting point and get outputs $y^{k,1}, \dots, y^{k,p_k}$

7: Compute $\nabla\ \ (y^{k,1},\ \ ^{k,1}, r_k), \dots, \nabla\ \ (y^{k,p_k},\ \ ^{k,p_k}, r_k)$

8: $p(k) \quad \operatorname{argmin}_{p=1,\dots,p_k} \|\nabla\ \ (y^{k,p},\ \ ^{k,p}, r_k)\|_2$

9: **end for**

Ensure: $y^{l,p(l)}$.

Theorem 4.5.5. *Assume that ψ is μ -strongly convex and L -smooth. If Algorithm 7 is*

run with

$$
\begin{aligned}
l &= \max\left\{1, \log_2 \frac{2R_y^2 k\, r\, \psi(y^0) k_2^2}{\varepsilon^2}\right\} \\
\hat{\ell}_k &= \max\left\{1, \frac{4\sigma^2 \left(1 + \sqrt{3 \ln \frac{l}{\beta}}\right)^2 R_y^2}{\varepsilon^2}\right\}, \\
r_k &= \max\left\{1, \frac{64 C \sigma^2 \ln^6 N}{N k\, r \quad (y^{k-1;p(k-1)}, \quad ^{k-1;p(k-1)}, \hat{\ell}_k) k_2^2}\right\}, \\
p_k &= \max\left\{1, \log_2 \frac{l}{\beta}\right\} \\
r_k &= \max\left\{1, \frac{128 \sigma^2 \left(1 + \sqrt{3 \ln \frac{l p_k}{\beta}}\right)^2 R_y^2}{\varepsilon^2}\right\}
\end{aligned} \tag{4.53}
$$

for all $k = 1, \ldots, l$ *where* $N > 1$ *is such that* $\frac{C L_\psi^2 \ln^4 \bar{N}}{\frac{2}{\psi} \bar{N}^4} \le \frac{1}{32}$, β *2* $(0, 1/3)$ *and* $\varepsilon > 0$*, then with probability at least* $1 - 3\beta$

$$
k\, r\, \psi(y^{l;p(l)}) k_2 \le \frac{\varepsilon}{R_y} \tag{4.54}
$$

and the total number of the oracle calls equals

$$
\sum_{k=1}^{l} (\hat{\ell}_k + N p_k r_k + p_k r_k) = \widetilde{O}\left(\max\left\{\sqrt{\frac{L}{\mu}}, \frac{\sigma^2 R_y^2}{\varepsilon^2}\right\}\right). \tag{4.55}
$$

Corollary 4.5.2. *Under assumptions of Theorem 4.5.5 we get that with probability at least* $1 - 3\beta$

$$
k y^{l;p(l)} - y^* k_2 \le \frac{\varepsilon}{\mu \ R_y}, \tag{4.56}
$$

where β *2* $(0, 1/3)$ *the total number of the oracle calls is defined in* (4.55).

Proof. Inequalities (4.54) and $\mu \ k y - y^* k_2 \le k\, r\, \psi(y) k_2$ which follows from μ -strong convexity of ψ imply that

$$
k y^{l;p(l)} - y^* k_2 \le \frac{k\, r\, \psi(y^{l;p(l)}) k_2}{\mu} \overset{(4.54)}{\le} \frac{\varepsilon}{\mu \ R_y}.
$$

□

Now we are ready to present convergence guarantees for the primal function and variables.

Corollary 4.5.3. *Let the assumptions of Theorem 4.5.5 hold. Assume that f is L_f-Lipschitz continuous on $B_{R_f}(0)$ where*

$$R_f = \left(\frac{\mu}{8\sqrt{\lambda_{\max}(A^\top A)}} + \frac{\sqrt{\lambda_{\max}(A^\top A)}}{\mu} + \frac{R_x}{R_y} \right) R_y$$

and $R_x = \|x(A^\top y^*)\|_2$. Then, with probability at least $1 - 4\beta$

$$f(x^l) - f(x^*) \le \left(2 + \frac{L_f}{8R_y\sqrt{\lambda_{\max}(A^\top A)}} \right) \varepsilon, \quad \|Ax^l\| \le \frac{9\varepsilon}{8R_y}, \tag{4.57}$$

where $\beta \in (0, 1/4)$, $\varepsilon \in (0, \mu R_y^2)$ $x^l \overset{def}{=} x(A^\top y^{l;p(l)}, \ell^{l;p(l)}, r_l)$ and to achieve it we need the total number of oracle calls equals

$$\sum_{k=1}^{l} (\hat{N}_k + N p_k r_k + p_k r_k) = \widetilde{O}\left(\max\left\{ \sqrt{\frac{L}{\mu}\chi(A^\top A)}, \frac{\sigma_x^2 M^2}{\varepsilon^2}\chi(A^\top A) \right\} \right) \tag{4.58}$$

where $M = \|\nabla f(x^*)\|_2$.

4.5.3. Direct Acceleration for Strongly Convex Dual Function

We consider first the following minimization problem:

$$\min_{y \in \mathbb{R}^n} \psi(y), \tag{4.59}$$

where $\psi(y)$ is μ -strongly convex and L -smooth. We use the same notation to define the objective in (4.59) as for the dual function from (4.22) because later in the section we apply the algorithm introduced below to the (4.22), but for now it is not important that ψ is a dual function for (4.21) and we prefer to consider more general situation. As in Section 4.5.1, we do not assume that we have an access to the exact gradient of $\psi(y)$ and consider instead of it biased stochastic gradient $\nabla\psi(y, \xi)$ satisfying inequalities (4.39) and (4.40) with $\delta \ge 0$ and $\sigma \ge 0$. In the main method of this section batched version of the stochastic gradient is used:

$$\nabla \quad (y, \quad ^k) = \frac{1}{r_k} \sum_{l=1}^{r_k} \nabla\psi(y, \xi^l), \tag{4.60}$$

where r_k is the batch-size that we leave unspecified for now. Note that $\nabla \quad (y, \quad ^k)$ satisfies inequalities (4.42) and (4.43).

We use Stochastic Similar Triangles Method which is stated in this section as Algorithm 8 to solve problem (4.59). To define the iterate z^{k+1} we use the following

sequence of functions:

$$\begin{aligned} g_0(z) &\stackrel{\text{def}}{=} \frac{1}{2}\|z - z^0\|_2^2 + \alpha_0\left(\psi(y^0) + \langle \nabla \psi(y^0), z - y^0\rangle + \frac{\mu}{2}\|z - y^0\|_2^2\right), \\ g_{k+1}(z) &\stackrel{\text{def}}{=} g_k(z) + \alpha_{k+1}\left(\psi(y^{k+1}) + \langle \nabla \psi(y^{k+1}), z - y^{k+1}\rangle + \frac{\mu}{2}\|z - y^{k+1}\|_2^2\right) \\ &= \frac{1}{2}\|z - z^0\|_2^2 + \sum_{l=0}^{k+1} \alpha_l \left(\psi(y^l) + \langle \nabla \psi(y^l), z - y^l\rangle + \frac{\mu}{2}\|z - y^l\|_2^2\right) \quad (4.61) \end{aligned}$$

We notice that $g_k(z)$ is $(1 + A_k\mu)$-strongly convex.

Algorithm 8 Stochastic Similar Triangles Methods for strongly convex problems (SSTM_sc)

Require: $y^0 = z^0 = y^0$ — starting point, N — number of iterations

1: Set $\alpha_0 = A_0 = 1/L_\psi$

2: Get $\nabla \psi(y^0, \cdot)$ to define $g_0(z)$

3: **for** $k = 0, 1, \ldots, N-1$ **do**

4: Choose α_{k+1} such that $A_{k+1} = A_k + \alpha_{k+1}$, $A_{k+1}(1 + A_k\mu) = \alpha_{k+1}^2 L$

5: $y^{k+1} = (A_k y^k + \alpha_{k+1} z^k)/A_{k+1}$

6: $z^{k+1} = \operatorname{argmin}_{z \in \mathbb{R}^n} g_{k+1}(z)$, where $g_{k+1}(z)$ is defined in (4.61)

7: $y^{k+1} = (A_k y^k + \alpha_{k+1} z^{k+1})/A_{k+1}$

8: **end for**

Ensure: x^N

Lemma 4.5.1. *Assume that Algorithm 8 is run to solve problem* (4.59) *with* $\psi(y)$ *being* μ *-strongly convex and* L *-smooth. Then, for all* $k \geq 0$ *we have*

$$\begin{aligned} A_k \psi(y^k) \leq\ & g_k(z^k) - \sum_{l=0}^{k-1} \frac{A_l \mu}{2}\|y^l - y^{l+1}\|_2^2 \\ & + \sum_{l=0}^{k} \frac{\alpha_l}{2\mu}\left\|\nabla \psi(y^l, \cdot) - \nabla \psi(y^l)\right\|_2^2. \quad (4.62) \end{aligned}$$

Lemma 4.5.2. *Let the sequences of non-negative numbers* $\{\alpha_k\}_{k \geq 0}$*, random non-negative variables* $\{R_k\}_{k \geq -1}$*,* $\{\widetilde{R}_k\}_{k \geq -1}$ *and random vectors* $\{\eta^k\}_{k \geq 0}$*,* $\{a^k\}_{k \geq 0}$*,* $\{a'^k\}_{k \geq 0}$ *satisfy inequality*

$$\begin{aligned} A_l R_l^2 + \sum_{k=0}^{l-1} A_k \widetilde{R}_k^2 \leq\ & A + h\delta \sum_{k=0}^{l} \alpha_k (R_{k-1} + \widetilde{R}_k) \\ & + u \sum_{k=0}^{l-1} \alpha_{k+1} \langle \eta^k, a^k + a'^k \rangle + c \sum_{k=0}^{l-1} \alpha_{k+1} \|\eta^k\|_2^2, \quad (4.63) \end{aligned}$$

for all $l = 1, \ldots, N$, *where* h, δ, u *and* c *are some non-negative constants and* $A_{k+1} = A_k + \alpha_{k+1}$, $\alpha_{k+1} \le DA_k$ *for some* $D \ge 1$, $A_0 = \alpha_0 > 0$. *Assume that for each* $k \ge 1$ *vector* a^k *is a function of* $\eta^0, \ldots, \eta^{k-1}$, a^0 *is a deterministic vector,* $u \ge 1$, *sequence of random vectors* $\{\eta^k\}_{k \ge 0}$ *satisfy*

$$\mathbf{E}\left[\eta^k \mid \eta^0, \ldots, \eta^{k-1}\right] = 0, \quad \mathbf{E}\left[\exp\left(\frac{\|\eta^k\|_2^2}{\sigma_k^2}\right) \mid \eta^0, \ldots, \eta^{k-1}\right] \le \exp(1), \tag{4.64}$$

$\forall k \ge 0$, $\sigma_k^2 \le \frac{C''}{N^2\left(1+\sqrt{3\ln\frac{N}{\beta}}\right)^2}$ *for some* $C > 0$, $\varepsilon > 0$, $\beta \in (0,1)$, *sequences* $\{a^k\}_{k \ge 0}$ *and* $\{\alpha^k\}_{k \ge 0}$ *are such that* $\|a^k\|_2 \le R_k$ *and* $\|\alpha^k\|_2 \le \widetilde{R}_k$, R_k *and* $\widetilde{R}_k$ *depend only on* $\eta_0, \ldots, \eta^k$ *and* $\widetilde{R}_0 = 0$. *If additionally* $\delta \le \frac{GR_0}{N\sqrt{A_N}}$ *and* $\varepsilon \le \frac{HR_0^2}{A_N}$ *Then with probability at least* $1 - 2\beta$ *the inequalities*

$$R_l \le \frac{JR_0}{\sqrt{A_l}}, \quad \widetilde{R}_{l-1} \le \frac{JR_0}{\sqrt{A_{l-1}}} \tag{4.65}$$

and

$$h\delta \sum_{k=0}^{l-1} \alpha_{k+1}(R_k + \widetilde{R}_k) + u\sum_{k=0}^{l-1} \alpha_{k+1}\langle \eta^k, a^k + \alpha^k\rangle + c\sum_{k=0}^{l-1} \alpha_{k+1}\|\eta^k\|_2^2 \le \left(2cHC + 2JD\left(hG + uC_1\sqrt{2HCg(N)}\right)\right) R_0^2 \tag{4.66}$$

hold for all $l = 1, \ldots, N$ *simultaneously, where* C_1 *is some positive constant,* $g(N) = \frac{\ln\left(\frac{N}{\beta}\right) + \ln\ln\left(\frac{B}{b}\right)}{\left(1+\sqrt{3\ln\left(\frac{N}{\beta}\right)}\right)^2}$,

$$B = 8HCDR_0^2\left(N\left(\frac{3}{2}\right)^N + 1\right)\left(A + 2Dh^2G^2R_0^2 + 2C\left(c + 2Du^2\right)HR_0^2\right),$$

$b = 2\sigma_0^2\alpha_1^2R_0^2$ *and*

$$J = \max\left\{\sqrt{A_0}, \frac{3B_1D + \sqrt{9B_1^2D^2 + \frac{4A}{R_0^2} + 8cHC}}{2}\right\},$$

$$B_1 = hG + uC_1\sqrt{2HCg(N)}.$$

Theorem 4.5.6. *Assume that the function* ψ *is* μ *-strongly convex and* L *-smooth,*

$$r_k = \left(\max\left\{1, \left(\frac{\mu}{L}\right)^{3/2}\frac{N^2\sigma^2\ln\frac{N}{\varepsilon}}{\varepsilon}\right\}\right),$$

i.e. $r_k \ge \frac{1}{C}\max\left\{1, \left(\frac{\psi}{L_\psi}\right)^{3/2}\frac{N^2\frac{2}{\psi}\left(1+\sqrt{3\ln\frac{N}{\beta}}\right)^2}{\varepsilon}\right\}$ *with positive constants* $C > 0$, $\varepsilon > 0$ *and* $N \ge 1$. *If additionally* $\delta \le \frac{GR_0}{N\sqrt{A_N}}$ *and* $\varepsilon \le \frac{HR_0^2}{A_N}$ *where* $R_0 = \|y^* - y^0\|_2$ *and Algorithm 8 is run for* N *iterations, then with probability at least* $1 - 3\beta$

$$\|y^N - y^*\|_2^2 \le \frac{J^2R_0^2}{A_N}, \tag{4.67}$$

where $\beta \in (0, 1/3)$,

$$\hat{g}(N) = \frac{\ln\left(\frac{N}{\beta}\right) + \ln\ln\left(\frac{\hat{B}}{b}\right)}{\left(1 + \sqrt{3\ln\left(\frac{N}{\beta}\right)}\right)^2}, \; b = \frac{2\sigma_1^2\alpha_1^2 R_0^2}{r_1}, \; D \overset{(4.229)}{=} 1 + \frac{\mu}{L} + \sqrt{1 + \frac{\mu}{L}},$$

$$\hat{B} = 8HC\left(\frac{L}{\mu}\right)^{3/2} DR_0^4\left(N\left(\frac{3}{2}\right)^N + 1\right)\left(\hat{A} + 2Dh^2G^2 + 2C\left(\frac{L}{\mu}\right)^{3/2}(c + 2Du^2)\,H\right),$$

$$h = u = \frac{2}{\mu}, \; c = \frac{2}{\mu^2},$$

$$\hat{A} = \frac{1}{\mu} + \frac{2G}{L\;\mu\;N\sqrt{A_N}} + \frac{2G^2}{\mu^2 N^2} + \left(\frac{L}{\mu}\right)^{3/4}\frac{2\sqrt{2CH}}{L\;\mu\;N\sqrt{A_N}} + \left(\frac{L}{\mu}\right)^{3/2}\frac{4CH}{L\;\mu^2 N^2 A_N},$$

$$\hat{\gamma} = \max\left\{\sqrt{\frac{1}{L}}, \frac{3\hat{B}_1 D + \sqrt{9\hat{B}_1^2 D^2 + 4\hat{A} + 8cHC\left(\frac{L_\psi}{\psi}\right)^{3/2}}}{2}\right\},$$

$$\hat{B}_1 = hG + uC_1\sqrt{2HC\left(\frac{L}{\mu}\right)^{3/2}\hat{g}(N)}$$

and C_1 *is some positive constant. In other words, to achieve* $\|y^N - y^*\|_2^2 \le \varepsilon$ *with probability at least* $1 - 3\beta$ *Algorithm 8 needs* $N = \widetilde{O}\left(\sqrt{\frac{L_\psi}{\psi}}\right)$ *iterations and* $\widetilde{O}\left(\max\left\{\sqrt{\frac{L_\psi}{\psi}}, \frac{2}{\psi}\right\}\right)$ *oracle calls where* $\widetilde{O}(\cdot)$ *hides polylogarithmic factors depending on* L, μ, R_0, ε *and* β.

Next, we apply the SSTM_sc for the problem (4.22) when the objective of the primal problem (4.21) is L-smooth, μ-strongly convex and L_f-Lipschitz continuous on some ball which will be specified next, i.e. we consider the same setup as in Section 4.5 but we additionally assume that the primal functional f has L-Lipschitz continuous gradient. As in Section 4.5 we also consider the case when the gradient of the dual functional is known only through biased stochastic estimators, see (4.36)–(4.43) and the paragraphs containing these formulas.

In Section 4.5 and 4.5.2 we mentioned that in the considered case dual function ψ is L-smooth on $\mathbb{R}^n$ and μ-strongly convex on $y^0 + (\mathrm{Ker}A^\top)^\perp$ where $L = \lambda_{\max}(AA^\top)/\alpha$ and $\mu = \lambda_{\min}^+(AA^\top)/\alpha$. Using the same technique as in the proof of Theorem 4.5.4 we show next that w.l.o.g. one can assume that ψ is μ-strongly convex on $\mathbb{R}^n$ since $\tilde{r}(y, y^k)$ lies

in $\mathrm{Im}A = (\mathrm{Ker}A^\top)^\perp$ by definition of $\pi_{}(y, {}^k)$. For this purposes we need the explicit formula for z^{k+1} which follows from the equation $r\, g_{k+1}(z^{k+1}) = 0$:

$$z^{k+1} = \frac{z^0}{1 + A_{k+1}\mu} + \sum_{l=0}^{k+1} \frac{\alpha_l \mu}{1 + A_{k+1}\mu} \nabla f^l - \frac{1}{1 + A_{k+1}\mu} \sum_{l=0}^{k+1} \alpha_l \pi_{}(\nabla f^l, {}^l). \tag{4.68}$$

Theorem 4.5.7. *For all $k \geq 0$ we have that the iterates of Algorithm 8 $\nabla f^k, z^k, y^k$ lie in $y^0 + \left(Ker(A^\top)\right)^\perp$.*

Proof. We prove the statement of the theorem by induction. For $k = 0$ the statement is trivial, since $\nabla f^0 = z^0 = y^0$. Assume that for some $k \geq 0$ we have $\nabla f^t, z^t, y^t \in y^0 + \left(\mathrm{Ker}(A^\top)\right)^\perp$ for all $0 \leq t \leq k$ and prove it for $k + 1$. Since $y_0 + \left(\mathrm{Ker}(A^\top)\right)^\perp$ is a convex set and ∇f^{k+1} is a convex combination of y^k and z^k we have $\nabla f^{k+1} \in y^0 + \left(\mathrm{Ker}(A^\top)\right)^\perp$. Next, the point $\frac{z^0}{1+A_{k+1}\ \psi} + \sum_{l=0}^{k+1} \frac{l\ \psi}{1+A_{k+1}\ \psi} \nabla f^l$ also lies in $y^0 + \left(\mathrm{Ker}(A^\top)\right)^\perp$ since it is convex combination of the points lying in this set which follows from $A_{k+1} = \sum_{l=0}^{k+1} \alpha_l$. By definition $\pi_{}(\nabla f^l, {}^l)$ of we have that $\pi_{}(\nabla f^l, {}^l)$ lies in $\mathrm{Im}A = (\mathrm{Ker}A^\top)^\perp$ for all ∇f^l. Putting all together and using (4.68) we get $z^{k+1} \in y^0 + \left(\mathrm{Ker}(A^\top)\right)^\perp$. Finally, y^{k+1} lies in $y^0 + \left(\mathrm{Ker}(A^\top)\right)^\perp$ as a convex combination of points from this set. □

This theorem makes it possible to apply the result from Theorem 4.5.6 for SSTM_sc which is run on the problem (4.22).

Corollary 4.5.4. *Under assumptions of Theorem 4.5.6 we get that after $N = \widetilde{O}\left(\sqrt{\frac{L_\psi}{\psi}} \ln \frac{1}{\varepsilon}\right)$ iterations of Algorithm 8 which is run on the problem (4.22) with probability at least $1 - 3\beta$*

$$\|\nabla \psi(y^N)\|_2 \leq \frac{\varepsilon}{R_y}, \tag{4.69}$$

where $\beta \in (0, 1/3)$ and the total number of oracles calls equals

$$\widetilde{O}\left(\max\left\{\sqrt{\frac{L}{\mu}}, \frac{\sigma^2 R_y^2}{\varepsilon^2}\right\}\right). \tag{4.70}$$

If additionally $\varepsilon \leq \mu\ R_y^2$, then with probability at least $1 - 3\beta$

$$\|y^N - y^*\|_2 \leq \frac{\varepsilon}{\mu\ R_y}, \tag{4.71}$$

$$\|y^N\|_2 \leq 2R_y \tag{4.72}$$

Proof. Theorem 4.5.6 implies that with probability at least $1 - 3\beta$ we have

$$\|y^N - y^*\|_2^2 \leq \frac{\hat{\mathcal{J}}^2 R_0^2}{A_N}.$$

Using this and L -smoothness of ψ we get that with probability $\geq 1 - 3\beta$

$$\|\nabla \psi(y^N)\|_2^2 = \|\nabla \psi(y^N) - \nabla \psi(y^*)\|_2^2 \leq L^2 \|y^N - y^*\|_2^2 \leq \frac{L^2 \hat{J}^2 R_0^2}{A_N}.$$

Since $A \overset{(4.228)}{\geq} \frac{1}{L_\psi}\left(1 + \frac{1}{2}\sqrt{\frac{\psi}{L_\psi}}\right)^{2k}$, it implies that after $N = \widetilde{O}\left(\sqrt{\frac{L_\psi}{\psi}} \ln \frac{1}{\varepsilon}\right)$ iterations of SSTM_sc we will get (4.69) with probability at least $1 - 3\beta$ and the number of oracle calls will be

$$\sum_{k=0}^{N} r_k = \widetilde{O}\left(\max\left\{\sqrt{\frac{L}{\mu}}, \frac{\sigma^2 R_y^2}{\varepsilon^2}\right\}\right).$$

Next, from μ -strong convexity of $\psi(y)$ we have that with probability at least $1 - 3\beta$

$$\|y^N - y^*\|_2 \leq \frac{\|\nabla \psi(y^N)\|_2}{\mu} \leq \frac{\varepsilon}{\mu R_y}$$

and from this we obtain that with probability at least $1 - 3\beta$

$$\|y^N\|_2 \leq \|y^N - y^*\|_2 + \|y^*\|_2 \leq \frac{\varepsilon}{\mu R_y} + R_y \leq 2R_y.$$

□

Corollary 4.5.5. *Let the assumptions of Theorem 4.5.6 hold. Assume that f is L_f-Lipschitz continuous on $B_{R_f}(0)$ where*

$$R_f = \left(\sqrt{\frac{2C}{\lambda_{\max}(A^\top A)}} + G_1 + \frac{\sqrt{\lambda_{\max}(A^\top A)}}{\mu}\right)\frac{\varepsilon}{R_y} + R_x,$$

$R_x = \|x(A^\top y^*)\|_2$, $\varepsilon \leq \mu R_y^2$ *and* $\delta_y \leq \frac{G_1 \varepsilon}{N R_y}$ *for some positive constant* G_1. *Assume additionally that the last batch-size r_N is slightly bigger than other batch-sizes, i.e.*

$$r_N \geq \frac{1}{C}\max\left\{1, \left(\frac{\mu}{L}\right)^{3/2} \frac{N^2\sigma^2\left(1 + \sqrt{3\ln\frac{N}{\beta}}\right)^2 R_y^2}{\varepsilon^2}, \frac{\sigma^2\left(1 + \sqrt{3\ln\frac{N}{\beta}}\right)^2 R_y^2}{\varepsilon^2}\right\}. \tag{4.73}$$

Then, with probability at least $1 - 4\beta$

$$f(x^N) - f(x^*) \leq \left(2 + \left(\sqrt{\frac{2C}{\lambda_{\max}(A^\top A)}} + G_1\right)\frac{L_f}{R_y}\right)\varepsilon, \tag{4.74}$$

$$\|Ax^N\|_2 \leq \left(1 + \sqrt{2C} + G_1\sqrt{\lambda_{\max}(A^\top A)}\right)\frac{\varepsilon}{R_y}, \tag{4.75}$$

where $\beta \in (0, 1/4)$, $x^N \stackrel{def}{=} x(A^\top y^N, \ldots, r_N)$ *and to achieve it we need the total number of oracle calls including the cost of computing* x^N *equals*

$$\widetilde{O}\left(\max\left\{\sqrt{\frac{L}{\mu}\chi(A^\top A)}, \frac{\sigma_x^2 M^2}{\varepsilon^2}\chi(A^\top A)\right\}\right) \tag{4.76}$$

where $M = \|f(x^*)\|_2$.

4.6. Applications to Decentralized Distributed Optimization

In this section we apply our results to the decentralized optimization problems. But let us consider first the centralized or parallel architecture. As we mentioned in the introduction, when the objective function is L-smooth one can compute batches in parallel [16, 25–27] in order to accelerate the work of the method and (4.11)-(4.13) imply that

$$O\left(\frac{2R^2/\varepsilon^2}{\sqrt{LR^2/\varepsilon}}\right) \text{ or } O\left(\frac{2/\varepsilon}{\sqrt{L/\mu}\ln(R^2/\varepsilon)}\right) \tag{4.77}$$

number of workers in such a parallel scheme gives the method with working time proportional to the number of iterations defined in (4.11). However, number of workers defined in (4.77) could be too big in order to use such an approach in practice. But still computing the batches in parallel even with much smaller number of workers could reduce the working time of the method if the communication is fast enough and it follows from (4.13).

Besides the computation of batches in parallel for the general type of problem (4.1)+(4.2), parallel optimization is often applied to the finite-sum minimization problems (4.1)+(4.3) or (4.1)+(4.6) that we rewrite here in the following form:

$$\min_{x \in Q \subseteq \mathbb{R}^n} f(x) = \frac{1}{m}\sum_{k=1}^{m} f_k(x). \tag{4.78}$$

We notice that in this section m is a number of workers and $f_k(x)$ is known only for the k-th worker. Consider the situation when workers are connected in a network and one can construct a spanning tree for this network. Assume that the diameter of the obtained graph equals d, i.e. the height of the tree — maximal distance (in terms of connections) between the root and a leaf [29]. If we run `STM` on such a spanning tree then we will get that the number of communication rounds will be d times larger than number of iterations defined in (4.11).

Now let us consider decentralized case when workers can communicate only with their neighbours. Next, we describe the method of how to reflect this restriction in the problem (4.78). Consider the Laplacian matrix $\overline{W} \in \mathbb{R}^{m\times m}$ of the network with vertices V and edges E which is defined as follows:

$$\overline{W}_{ij} = \begin{cases} -1, & \text{if } (i,j) \in E, \\ \deg(i), & \text{if } i = j, \\ 0 & \text{otherwise,} \end{cases} \tag{4.79}$$

where $\deg(i)$ is degree of i-th node, i.e. number of neighbours of the i-th worker. Since we consider only connected networks the matrix $\overline{W}$ has unique eigenvector $\mathbf{1}_m \stackrel{\text{def}}{=} (1,\ldots,1)^\top \in \mathbb{R}^m$ corresponding to the eigenvalue 0. It implies that for all vectors $a = (a_1,\ldots,a_m)^\top \in \mathbb{R}^m$ the following equivalence holds:

$$a_1 = \ldots = a_m \quad \Leftrightarrow \quad Wa = 0. \tag{4.80}$$

Now let us think about a_i as a number that i-th node stores. Then, using (4.80) we can use Laplacian matrix to express in the short matrix form the fact that all nodes of the network store the same number. In order to generalize it for the case when a_i are vectors from $\mathbb{R}^n$ we should consider the matrix $W \stackrel{\text{def}}{=} \overline{W} \otimes I_n$ where $\otimes$ represents the Kronecker product (see (3.1)). Indeed, if we consider vectors $x_1,\ldots,x_m \in \mathbb{R}^n$ and $\mathbf{x} = \left(x_1^\top,\ldots,x_m^\top\right) \in \mathbb{R}^{nm}$, then (4.80) implies

$$x_1 = \ldots = x_m \quad \Leftrightarrow \quad W\mathbf{x} = 0. \tag{4.81}$$

For simplicity, we also call W as a Laplacian matrix and it does not lead to misunderstanding since everywhere below we use W instead of $\overline{W}$. The key observation here that computation of Wx requires one round of communications when the k-th worker sends x_k to all its neighbours and receives x_j for all j such that $(k,j) \in E$, i.e. k-th worker gets vectors from all its neighbours. Note, that W is symmetric and positive semidefinite [29] and, as a consequence, $\sqrt{W}$ exists. Moreover, we can replace W by $\sqrt{W}$ in (4.81) and get the equivalent statement:

$$x_1 = \ldots = x_m \quad \Leftrightarrow \quad \sqrt{W}\mathbf{x} = 0. \tag{4.82}$$

Using this we can rewrite the problem (4.78) in the following way:

$$\min_{\substack{\sqrt{W}\mathbf{x}=0;\\ x_1,\ldots,x_m\in Q\subseteq\mathbb{R}^n}} f(\mathbf{x}) = \frac{1}{m}\sum_{k=1}^{m} f_k(x_k). \tag{4.83}$$

We are interested in the general case when $f_k(x_k) = \mathbf{E}_{\xi_k}[f_k(x_k, \xi_k)]$ where $\{\xi_k\}_{k=1}^m$ are independent. This type of objective can be considered as a special case of (4.6). Then, as it was mentioned in the introduction it is natural to use stochastic gradients $\nabla f_k(x_k, \xi_k)$ that satisfy

$$\|\mathbf{E}_{\xi_k}[\nabla f_k(x_k, \xi_k)] - \nabla f_k(x_k)\|_2 \leq \delta, \tag{4.84}$$

$$\mathbf{E}_{\xi_k}\left[\exp\left(\frac{\|\nabla f_k(x_k, \xi_k) - \mathbf{E}_{\xi_k}[\nabla f_k(x_k, \xi_k)]\|_2^2}{\sigma^2}\right)\right] \leq \exp(1). \tag{4.85}$$

Then, the stochastic gradient

$$\nabla f(\mathbf{x}, \xi) \overset{\text{def}}{=} \nabla f(\mathbf{x}, \{\xi_k\}_{k=1}^m) \overset{\text{def}}{=} \frac{1}{m}\sum_{k=1}^m \nabla f_k(x_k, \xi_k)$$

satisfies (see also (4.43))

$$\mathbf{E}\left[\exp\left(\frac{\|\nabla f(\mathbf{x}, \xi) - \mathbf{E}[\nabla f(\mathbf{x}, \xi)]\|_2^2}{\sigma_f^2}\right)\right] \leq \exp(1)$$

with $\sigma_f^2 = O(\sigma^2/m)$.

As always, we start with the smooth case with $Q = \mathbb{R}^n$ and assume that each f_k is L-smooth, μ-strongly convex and satisfies $\|\nabla_k f_k(x_k)\|_2 \leq M$ on some ball $B_{R_M}(x^*)$ where we use $\nabla_k f(x_k)$ to emphasize that f_k depends only on the k-th n-dimensional block of $\mathbf{x}$. Since the functional $f(\mathbf{x})$ in (4.83) has separable structure, it implies that f is L/m-smooth, μ/m-strongly convex and satisfies $\|\nabla f(\mathbf{x})\|_2 \leq M/m\sqrt{m}$ on $B_{\sqrt{m}R_M}(\mathbf{x}^*)$. Indeed, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$

$$\begin{aligned}
\|\mathbf{x} - \mathbf{y}\|_2^2 &= \sum_{k=1}^m \|x_k - y_k\|_2^2, \\
\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 &= \sqrt{\frac{1}{m^2}\sum_{k=1}^m \|\nabla_k f_k(x_k) - \nabla_k f_k(y_k)\|_2^2} \\
&\leq \sqrt{\frac{L^2}{m^2}\sum_{k=1}^m \|x_k - y_k\|_2^2} = \frac{L}{m}\|\mathbf{x} - \mathbf{y}\|_2, \\
f(\mathbf{x}) &= \frac{1}{m}\sum_{k=1}^m f_k(x_k) \geq \frac{1}{m}\sum_{k=1}^m \left(f(y_k) + \langle \nabla_k f_k(y_k), x_k - y_k\rangle + \frac{\mu}{2}\|x^k - y^k\|_2^2\right) \\
&= f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle + \frac{\mu}{2m}\|\mathbf{x} - \mathbf{y}\|_2^2, \\
\|\nabla f(\mathbf{x})\|_2^2 &= \frac{1}{m^2}\sum_{k=1}^m \|\nabla_k f_k(x_k)\|_2^2.
\end{aligned}$$

Therefore, one can consider the problem (4.83) as (4.21) with $A = \sqrt{W}$ and $Q = \mathbb{R}^{nm}$. Next, if the starting point $\mathbf{x}^0$ is such that $\mathbf{x}^0 = (x^0, \ldots, x^0)^\top$ then

$$\mathbf{R}^2 \stackrel{\text{def}}{=} \|\mathbf{x}^0 - \mathbf{x}^*\|_2^2 = m\|x^0 - x^*\|_2^2 = mR^2, \quad R_{\mathbf{y}}^2 \stackrel{\text{def}}{=} \|\mathbf{y}^*\|_2^2 \leq \frac{\|\sqrt{W} f(\mathbf{x}^*)\|_2^2}{\lambda_{\min}^+(W)} \leq \frac{M^2}{m\lambda_{\min}^+(W)}.$$

Now it should become clear why in Section 4.4 we paid most of our attention on number of $A^\top A\mathbf{x}$ calculations. In this particular scenario $A^\top A\mathbf{x} = \sqrt{W}^\top \sqrt{W} x = Wx$ which can be computed via one round of communications of each node with its neighbours as it was mentioned earlier in this section. That is, for the primal approach we can simply use the results discussed in Section 4.4. For convenience, we summarize them in Tables 4.3 and 4.4 which are obtained via plugging the parameters that we obtained above in the bounds from Section 4.4. Note that the results presented in this match the lower bounds obtained in [69] in terms of the number of communication rounds up to logarithmic factors and and there is a conjecture [45] that these bounds are also optimal in terms of number of oracle calls per node for the class of methods that require optimal number of communication rounds. Recently, the very similar result about the optimal balance between number of oracle calls per node and number of communication round was proved for the case when the primal functional is convex and L-smooth and deterministic first-order oracle is available [70].

Finally, consider the situation when $Q = \mathbb{R}^n$ and each f_k from (4.83) is dual-friendly, i.e. one can construct dual problem for (4.83)

$$\min_{\mathbf{y} \in \mathbb{R}^{nm}} \quad \psi(\mathbf{y}), \qquad \text{where } \mathbf{y} = (y_1^\top, \ldots, y_m^\top)^\top \in \mathbb{R}^{nm},\ y_1, \ldots, y_m \in \mathbb{R}^n, \tag{4.86}$$

$$\varphi_k(y_k) = \max_{x_k \in \mathbb{R}^n} \{\langle y_k, x_k \rangle - f_k(x_k)\}, \tag{4.87}$$

$$\psi(\mathbf{y}) = \frac{1}{m}\sum_{k=1}^m \varphi_k(my_k), \quad \psi(\mathbf{y}) = \psi(\sqrt{W}\mathbf{y}) = \frac{1}{m}\sum_{k=1}^m \varphi_k(m[\sqrt{W}\mathbf{x}]_k), \tag{4.88}$$

where $[\sqrt{W}\mathbf{x}]_k$ is the k-th n-dimensional block of $\sqrt{W}x$. Note that

$$\begin{aligned}\max_{\mathbf{x} \in \mathbb{R}^{nm}} \{\langle \mathbf{y}, \mathbf{x} \rangle - f(\mathbf{x})\} &= \max_{\mathbf{x} \in \mathbb{R}^{nm}} \left\{\sum_{k=1}^m \langle y_k, x_k \rangle - \frac{1}{m}\sum_{k=1}^m f_k(x_k)\right\} \\ &= \frac{1}{m}\sum_{k=1}^m \max_{x_k \in \mathbb{R}^n} \{\langle my_k, x_k \rangle - f_k(x_k)\} = \frac{1}{m}\sum_{k=1}^m \varphi_k(my_k) = \psi(\mathbf{y}),\end{aligned}$$

so, $\psi(\mathbf{y})$ is a dual function for $f(\mathbf{x})$. As for the primal approach, we are interested in the general case when $\varphi_k(y_k) = \mathbf{E}_{\xi_k}[\varphi_k(y_k, \xi_k)]$ where $\{\xi_k\}_{k=1}^m$ are independent and stochastic

Assumptions on f_k	Method	# of communication rounds	# of $\nabla f_k(x)$ oracle calls per node
μ-strongly convex, L-smooth	D-MASG, $Q = \mathbb{R}^n$, [58]	$\widetilde{O}\left(\sqrt{\frac{L}{\mu}}\chi\right)$	$\widetilde{O}\left(\sqrt{\frac{L}{\mu}}\right)$
L-smooth	STP_IPS with STP as a subroutine, $Q = \mathbb{R}^n$, [This paper]	$\widetilde{O}\left(\sqrt{\frac{LR^2}{\varepsilon}}\chi\right)$	$\widetilde{O}\left(\sqrt{\frac{LR^2}{\varepsilon}}\right)$
μ-strongly convex, $\|\nabla f_k(x)\|_2 \le M$	R-Sliding, [45] [53] [60, 71]	$\widetilde{O}\left(\sqrt{\frac{M^2}{\mu\varepsilon}}\chi\right)$	$\widetilde{O}\left(\frac{M^2}{\mu\varepsilon}\right)$
$\|\nabla f_k(x)\|_2 \le M$	Sliding, [53, 60] [71]	$O\left(\sqrt{\frac{M^2R^2}{\varepsilon^2}}\chi\right)$	$O\left(\frac{M^2R^2}{\varepsilon^2}\right)$

Table 4.3: Summary of the covered results in this paper for solving (4.83) using primal deterministic approach from Section 4.4. First column contains assumptions on f_k, $k = 1, \ldots, m$ in addition to the convexity, $\chi = \chi(W)$. All methods except D-MASG should be applied to solve (4.25).

gradients $\nabla\varphi_k(x_k, \xi_k)$ satisfy

$$\|\mathbf{E}_k[\nabla\varphi_k(y_k, \xi_k)] - \nabla\varphi_k(y_k)\|_2 \le \delta, \tag{4.89}$$

$$\mathbf{E}_k\left[\exp\left(\frac{\|\nabla\varphi_k(y_k, \xi_k) - \mathbf{E}_k[\nabla\varphi_k(y_k, \xi_k)]\|_2^2}{\sigma^2}\right)\right] \le \exp(1). \tag{4.90}$$

Consider the stochastic function $f_k(x_k, \xi_k)$ which is defined implicitly as follows:

$$\varphi_k(y_k, \xi_k) = \max_{x_k \in \mathbb{R}^n} \{\langle y_k, x_k\rangle - f(x_k, \xi_k)\}. \tag{4.91}$$

Since

$$\nabla\varphi^*(\mathbf{y}) = \sum_{k=1}^m \nabla\varphi_k^*(my_k) \overset{(4.35)}{=} \sum_{k=1}^m x_k(my_k) \overset{\text{def}}{=} \mathbf{x}(\mathbf{y}), \quad x_k(y_k) \overset{\text{def}}{=} \operatorname*{argmax}_{x_k \in \mathbb{R}^n} \{\langle y_k, x_k\rangle - f_k(x_k)\}$$

Assumptions on f_k	Method	# of communication rounds	# of $r f_k(x,\xi)$ oracle calls per node
μ-strongly convex, L-smooth	D-MASG, in expectation, $Q = \mathrm{R}^n$, [58]	$\widetilde{O}\left(\sqrt{\frac{L}{\mu}}\chi\right)$	$\widetilde{O}\left(\max\left\{\sqrt{\frac{L}{\mu}}, \frac{\sigma^2}{\mu\varepsilon}\right\}\right)$
L-smooth	SSTP_IPS with STP as a subroutine, $Q = \mathrm{R}^n$, conjecture, [This paper] [45]	$\widetilde{O}\left(\sqrt{\frac{LR^2}{\varepsilon}}\chi\right)$	$\widetilde{O}\left(\max\left\{\sqrt{\frac{LR^2}{\varepsilon}}, \frac{\sigma^2R^2}{\varepsilon^2}\right\}\right)$
μ-strongly convex, $\|r f_k(x)\|_2 \le M$	RS-Sliding Q is bounded, [45] [53] [60, 71]	$\widetilde{O}\left(\sqrt{\frac{M^2}{\mu\varepsilon}}\chi\right)$	$\widetilde{O}\left(\frac{M^2+\sigma^2}{\mu\varepsilon}\right)$
$\|r f_k(x)\|_2 \le M$	S-Sliding Q is bounded, [53, 60] [71]	$\widetilde{O}\left(\sqrt{\frac{M^2R^2}{\varepsilon^2}}\chi\right)$	$\widetilde{O}\left(\frac{(M^2+\sigma^2)R^2}{\varepsilon^2}\right)$

Table 4.4: Summary of the covered results in this paper for solving (4.83) using primal stochastic approach from Section 4.4 with the stochastic oracle satisfying (4.84)-(4.85) with $\delta = 0$. First column contains assumptions on f_k, $k = 1,\dots,m$ in addition to the convexity, $\chi = \chi(W)$. All methods except D-MASG should be applied to solve (4.25). The bounds from the last two rows hold even in the case when Q is unbounded, but in the expectation (see [72]).

it is natural to define the stochastic gradient $r\ (\mathbf{y},\xi)$ as follows:

$$r\ (\mathbf{y},\xi) \overset{\text{def}}{=} r\ (\mathbf{y}, f\xi_k g_{k=1}^m) \overset{\text{def}}{=} \sum_{k=1}^m r\,\varphi_k(my_k,\xi_k) \overset{(4.35)}{=} \sum_{k=1}^m x_k(my_k,\xi_k) \overset{\text{def}}{=} \mathbf{x}(\mathbf{y},\xi),$$

$$x_k(y_k,\xi_k) \overset{\text{def}}{=} \operatorname*{argmax}_{x_k 2 \mathrm{R}^n} f\langle y_k, x_k\rangle - f_k(x_k,\xi_k)g\,.$$

It satisfies (see also (4.43))

$$\begin{aligned} k\mathbf{E}\ [r\ (\mathbf{y},\xi)] - r\ (\mathbf{y})k_2 &\le \delta_\Phi, \\ \mathbf{E}\ \left[\exp\left(\frac{k\,r\ (\mathbf{y},\xi) - \mathbf{E}\ [r\ (\mathbf{y},\xi)]k_2^2}{\sigma_\Phi^2}\right)\right] &\le \exp(1) \end{aligned}$$

with $\delta_\Phi = m\delta$, and $\sigma_\Phi^2 = O\,(m\sigma^2)$. Using this, we define the stochastic gradient of $(\mathbf{y})$

as $r(\mathbf{y},\xi) \stackrel{\text{def}}{=} r(\sqrt{W}\mathbf{y},\xi) = r(\sqrt{W}\mathbf{x}(\sqrt{W}\mathbf{y},\xi)$ and, as a consequence, we get

$$\begin{aligned} \|\mathbf{E}[r(\mathbf{y},\xi)] - r(\mathbf{y})\|_2 &\leq \delta_\psi, \\ \mathbf{E}\left[\exp\left(\frac{\|r(\mathbf{y},\xi) - \mathbf{E}[r(\mathbf{y},\xi)]\|_2^2}{\sigma_\psi^2}\right)\right] &\leq \exp(1) \end{aligned}$$

with $\delta_\psi = \sqrt{\lambda_{\max}(W)}\delta_\Phi$ and $\sigma_\psi = \sqrt{\lambda_{\max}(W)}\sigma_\Phi$.

Taking all of this into account we conclude that problem (4.86) is a special case of (4.22) with $A = \sqrt{W}$. To make the algorithms from Section 4.5 distributed we should change the variables in those methods via multiplying them by $\sqrt{W}$ from the left [45, 46, 61], e.g. for the iterates of SPDSTM we will get

$$\mathcal{y}^{k+1} := \sqrt{W}\mathcal{y}^{k+1}, \quad z^{k+1} := \sqrt{W}z^{k+1}, \quad y^{k+1} := \sqrt{W}y^{k+1},$$

which means that it is needed to multiply lines 4-6 of Algorithm 3 by $\sqrt{W}$ from the left. After such a change of variables all methods from Section 4.5 become suitable to run them in the distributed fashion. Besides that, it does not spoil the ability of recovering the primal variables since before the change of variables all of the methods mentioned in Section 4.5 used $\mathbf{x}(\sqrt{W}\mathbf{y})$ or $\mathbf{x}(\sqrt{W}\mathbf{y},\xi)$ where points y were some dual iterates of those methods, so, after the change of variables we should use $\mathbf{x}(\mathbf{y})$ or $\mathbf{x}(\mathbf{y},\xi)$ respectively. Moreover, it is also possible to compute $\|\sqrt{W}x\|_2^2 = \langle\mathbf{x}, W\mathbf{x}\rangle$ in the distributed fashion using consensus type algorithms: one communication step is needed to compute $W\mathbf{x}$, then each worker computes $\langle x_k, [W\mathbf{x}]_k\rangle$ locally and after that it is needed to run consensus algorithm. We summarize the results for this case in Tables 4.5 and 4.6. Note that the proposed bounds are optimal in terms of the number of communication rounds up to polylogarithmic factors [29, 69, 73, 74]. Note that the lower bounds from [29, 73, 74] are presented for the convolution of two criteria: number of oracle calls per node and communication rounds. One can obtain lower bounds for the number of communication rounds itself using additional assumption that time needed for one communication is big enough and the term which corresponds to the number of oracle calls can be neglected. Regarding the number of oracle calls there is a conjecture [45] that the bounds that we present in this paper are also optimal up to polylogarithmic factors for the class of methods that require optimal number of communication rounds.

We would like to thank F. Bach, P. Dvurechensky, M. Gürbüzbalaban, D. Kovalev, A. Nemirovski, A. Olshevsky, N. Srebro, A. Taylor and C. Uribe for useful discussions.

Assumptions on f_k	Method	# of communication rounds	# of $r\,\varphi_k(y,\xi)$ oracle calls per node
μ-strongly convex, L-smooth, $k\,r\,f_k(x)k_2 \le M$	R-RRMA-AC-SA^2 (Algorithm 7), Corollary 4.5.3, SSTM_sc (Algorithm 8), Corollary 4.5.5	$\widetilde{O}\left(\sqrt{\frac{L}{"}}\chi\right)$	$\widetilde{O}\left(\max\left\{\sqrt{\frac{L}{"}}\chi, \frac{\frac{2}{\Phi}M^2}{"^2}\chi\right\}\right)$
μ-strongly convex, $k\,r\,f_k(x)k_2 \le M$	SPDSTM (Algorithm 3), Theorem 4.5.1	$\widetilde{O}\left(\sqrt{\frac{M^2}{"}}\chi\right)$	$\widetilde{O}\left(\max\left\{\sqrt{\frac{M^2}{"}}\chi, \frac{\frac{2}{\Phi}M^2}{"^2}\chi\right\}\right)$

Table 4.5: Summary of the covered results in this paper for solving (4.86) using dual stochastic approach from Section 4.5 with the stochastic oracle satisfying (4.84)-(4.85) with $\delta = 0$. First column contains assumptions on f_k, $k = 1,\dots,m$ in addition to the convexity, $\chi = \chi(W)$.

Assumptions on f_k	Method	# of communication rounds	# of $r\,\varphi_k(y,\xi)$ oracle calls per node
μ-strongly convex, L-smooth, $k\,r\,f_k(x)k_2 \le M$	SSTM_sc (Algorithm 8), Corollary 4.5.5	$\widetilde{O}\left(\sqrt{\frac{L}{"}}\chi\right)$	$\widetilde{O}\left(\max\left\{\sqrt{\frac{L}{"}}\chi, \frac{\frac{2}{\Phi}M^2}{"^2}\chi\right\}\right)$
μ-strongly convex, $k\,r\,f_k(x)k_2 \le M$	SPDSTM (Algorithm 3), Theorem 4.5.1	$\widetilde{O}\left(\sqrt{\frac{M^2}{"}}\chi\right)$	$\widetilde{O}\left(\max\left\{\sqrt{\frac{M^2}{"}}\chi, \frac{\frac{2}{\Phi}M^2}{"^2}\chi\right\}\right)$

Table 4.6: Summary of the covered results in this paper for solving (4.86) using **biased** dual stochastic approach from Section 4.5 with the stochastic oracle satisfying (4.84)-(4.85) with $\delta' > 0$. First column contains assumptions on f_k, $k = 1,\dots,m$ in addition to the convexity, $\chi = \chi(W)$. For both cases the noise level should satisfy $\delta' = \widetilde{O}("/M\sqrt{m})$.

The work of E. Gorbunov was supported by RFBR, project number 19-31-51001. The work of D. Dvinskikh was supported by Russian Science Foundation (project 18-71-10108). The work of A. Gasnikov was supported by RFBR, project number 19-31-51001 and by Yahoo! Research Faculty Engagement Program.

4.7. Discussion

In this section we want to discuss some aspects of the proposed results that were not covered in the main part of this paper. First of all, we should say that in the smooth case for the primal approach our bounds for the number of communication steps coincides with the optimal bounds for the number of communication steps for parallel optimization if we substitute the diameter d of the spanning tree in the bounds for parallel optimization by $\widetilde{O}(\sqrt{\chi(W)})$.

However, we want to discuss another interesting difference between parallel and decentralized optimization in terms of the complexity results which was noticed in [45]. From the line of works [75–78] it is known that for the problem (4.1)+(4.6) (here we use m instead of q and iterator k instead of i for consistency) with L-smooth and μ-strongly convex f_k for all $k = 1, \ldots, m$ the optimal number of oracle calls, i.e. calculations of of the stochastic gradients of f_k with σ^2-subgaussian variance is

$$\widetilde{O}\left(m + \sqrt{m\frac{L}{\mu}} + \frac{\sigma^2}{\mu\varepsilon}\right). \tag{4.92}$$

The bad news is that (4.92) does not work with full parallelization trick and the best possible way to parallelize it is described in [78]. However, standard accelerated scheme using mini-batched versions of the stochastic gradients without variance-reduction technique and incremental oracles which gives the bound

$$\widetilde{O}\left(m\sqrt{\frac{L}{\mu}} + \frac{\sigma^2}{\mu\varepsilon}\right) \tag{4.93}$$

for the number of oracle calls and it admits full parallelization. It means that in the parallel optimization setup when we have computational network with m nodes and the spanning tree for it with diameter d the number of oracle calls per node is

$$\widetilde{O}\left(\sqrt{\frac{L}{\mu}} + \frac{\sigma^2}{m\mu\varepsilon}\right) = \widetilde{O}\left(\max\left\{\sqrt{\frac{L}{\mu}}, \frac{\sigma^2}{m\mu\varepsilon}\right\}\right) \tag{4.94}$$

and the number of communication steps is

$$\widetilde{O}\left(d\sqrt{\frac{L}{\mu}}\right). \tag{4.95}$$

However, for the decentralized setup the second row of Table 4.4 states that the number of communication rounds is the same as in (4.95) up to substitution of d by $\sqrt{\chi(W)}$ and

the number of oracle calls per node is

$$\widetilde{O}\left(\max\left\{\sqrt{\frac{L}{\mu}}, \frac{\sigma^2}{\mu\varepsilon}\right\}\right) \tag{4.96}$$

which has m times bigger statistical term under the maximum than in (4.94). What is more, recently it was shown that there exists such a decentralized distributed method that requires

$$\widetilde{O}\left(\frac{\sigma^2}{m\mu\varepsilon}\right)$$

stochastic gradient oracle calls per node [79, 80], but it is not optimal in terms of the number of communications. Moreover, there is a hypothesis [45] that in the smooth case the bounds from Tables 4.3 and 4.4 (rows 2 and 3) are optimal in terms of the number of oracle calls per node *for the class of methods that require optimal number of communication rounds* up to polylogarithmic factors.

The same claim but for Table 4.5 was also presented in [45] as a hypothesis and in this paper we propose the same hypothesis for the result stated Table 4.6 up to polylogarithmic and additionally we hypothesise that the noise level that we obtained is also unimprovable up to polylogarithmic factors.

4.7.1. Possible Extensions

- As it was mentioned in Section 4.4, the recurrence technique that we use in Sections 4.3 and 4.5 can be very useful in the generalization of the results for STM from Section 4.4 for the case when instead of $\nabla f(x)$ only stochastic gradient $\nabla f(x, \xi)$ (see inequalities (4.7)-(4.8)) is available, f is L-smooth and proximal step is computed in an inexact manner. It would be nice also to compare proposed methods for the case when δ with the results from [58]. For the convex but non-strongly convex case one can also try to combine Nesterov's smoothing technique [61, 81, 82] with D-MASG from [58].

- We believe that the technique presented in the proofs of Lemmas 4.9.8 and 4.5.2 can also be extended or modified in order to be applied for different optimization methods to obtain high probability bounds in the case when $Q = \mathbb{R}^n$.

- We emphasize that in our results we assume that each f_i from (4.83) is L-smooth and μ-strongly convex. When each f_i is L_i-smooth and μ_i-strongly convex, it means that in order to satisfy the assumption we use in our paper we need to choose

$L = \max_{1 \le i \le m} L_i$ and $\mu = \min_{1 \le i \le m} \mu_i$. This choice can lead to a very slow rate in some situations, e.g. the worst-case L can be m times larger than L for f as for the case when $m = d$ and $f(x) = \|x\|_2^2/2m = \frac{1}{m} \sum_{i=1}^m f_i(x)$, $f_i(x) = x_i^2/2$ where $L_i = 1$ for all i but f is $\frac{1}{d}$-smooth [83]. It was shown [29, 61] that instead of worst-case μ and L one can use $\mu = \frac{1}{m} \sum_{i=1}^m \mu_i$ and $\hat{L}$ to be some weighted average of L_i, but such techniques can spoil number of communication rounds needed to achieve desired accuracy.

- It would be also interesting to generalize the proposed results for the case of more general stochastic gradients [9, 11, 13, 59].

4.8. Application for Population Wasserstein Barycenter Calculation

In this section we consider the problem of calculation of population Wasserstein barycenter since this example hides different interesting details connected with the theory discussed in this paper. In our presentation of this example we rely mostly on the recent work [84].

4.8.1. Definitions and Properties

We define the probability simplex in $\mathbb{R}^n$ as $S_n(1) = \left\{x \in \mathbb{R}^n_+ \mid \sum_{i=1}^n x_i = 1\right\}$. One can interpret the elements of $S_n(1)$ as discrete probability measures with n shared atoms. For an arbitrary pair of measures $p, q \in S_n(1)$ we introduce the set $\Pi(p, q) = \left\{\pi \in \mathbb{R}_+^{n \times n} \mid \pi \mathbf{1} = p,\ \pi^\top \mathbf{1} = q\right\}$ called transportation polytope. Optimal transportation (OT) problem between measures $p, q \in S_n(1)$ is defined as follows

$$W(p, q) = \min_{\pi \in \Pi(p;q)} \langle C, \pi \rangle = \min_{\pi \in \Pi(p;q)} \sum_{i,j=1}^n C_{ij} \pi_{ij} \tag{4.97}$$

where C is a transportation cost matrix. That is, (i, j)-th component C_{ij} of C is a cost of transportation of the unit mass from point x_i to the point x_j where points $x_1, \ldots, x_n \in \mathbb{R}$ are atoms of measures from $S_n(1)$.

Next, we consider the entropic OT problem (see [85, 86])

$$W_\mu(p, q) = \min_{\pi \in \Pi(p;q)} \sum_{i,j=1}^n \left(C_{ij} \pi_{ij} + \mu \pi_{ij} \ln \pi_{ij}\right). \tag{4.98}$$

Consider some probability measure $\mathbf{P}$ on $S_n(1)$. Then one can define population barycenter of measures from $S_n(1)$ as

$$p^* = \operatorname*{argmin}_{p \in S_n(1)} \int_{q \in S_n(1)} W(p,q) d\mathbf{P}(q) = \operatorname*{argmin}_{p \in S_n(1)} \underbrace{\mathbf{E}_q [W(p,q)]}_{W_\mu(p)}. \tag{4.99}$$

For a given set of samples $q^1, \ldots, q^m$ we introduce empirical barycenter as

$$\hat{p}^* = \operatorname*{argmin}_{p \in S_n(1)} \underbrace{\frac{1}{m} \sum_{i=1}^{m} W(p, q^i)}_{\hat{W}(p)}. \tag{4.100}$$

We consider the problem (4.99) of finding population barycenter with some accuracy and discuss possible approaches to solve this problem in the following subsections.

However, before that, we need to mention some useful properties of $W(p,q)$. First of all, one can write explicitly the dual function of $W(p,q)$ for a fixed $q \in S_n(1)$ (see [84, 87]):

$$W(p,q) = \max_{\lambda \in \mathbb{R}^n} \left\{ \langle \lambda, p \rangle - W^*_{q;}(\lambda) \right\} \tag{4.101}$$

$$W^*_{q;}(\lambda) = \mu \sum_{j=1}^{n} q_j \ln \left(\frac{1}{q_j} \sum_{i=1}^{n} \exp\left(\frac{-C_{ij} + \lambda_i}{\mu} \right) \right). \tag{4.102}$$

Using this representation one can deduce the following theorem.

Theorem 4.8.1 ([84]). *For an arbitrary $q \in S_n(1)$ the entropic Wasserstein distance $W(\cdot, q) : S_n(1) \to \mathbb{R}$ is μ-strongly convex w.r.t. ℓ_2-norm and M-Lipschitz continuous w.r.t. ℓ_2-norm. Moreover, $M \le \sqrt{n} M_1$ where M_1 is Lipschitz constant of $W(\cdot, q)$ w.r.t. ℓ_1-norm and $M_1 = \widetilde{O}(\|C\|_1)$.*

We also want to notice that function $W^*_{q;}(\lambda)$ is only strictly convex and the minimal eigenvalue of its hessian $\gamma \stackrel{\text{def}}{=} \lambda_{\min}(\nabla^2 W^*_{q;}(\lambda^*))$ evaluated in the solution $\lambda^* \stackrel{\text{def}}{=} \operatorname{argmax}_{\lambda \in \mathbb{R}^n} \left\{ \langle \lambda, p \rangle - W^*_{q;}(\lambda) \right\}$ is very small and there exist only such bounds that are exponentially small in n.

We will also use another useful relation (see [84]):

$$\nabla W(p,q) = \lambda^*, \quad \langle \lambda^*, \mathbf{1} \rangle = 0 \tag{4.103}$$

where the gradient $\nabla W(p,q)$ is taken w.r.t. the first argument.

4.8.2. SA Approach

Assume that one can obtain and use fresh samples $q^1, q^2, \ldots$ in online regime. This approach is called Stochastic Approximation (SA). It implies that at each iteration one can draw a fresh sample q^k and compute the gradient w.r.t. p of function $W(p, q^k)$ which is μ-strongly convex and M-Lipschitz continuous with $M = \widetilde{O}(\sqrt{n}kCk_1)$. Optimal methods for this case are based on iterations of the following form

$$p^{k+1} = \mathrm{proj}_{S_n(1)}\left(p^k - \eta_k \nabla W(p^k, q^k)\right)$$

where $\mathrm{proj}_{S_n(1)}(x)$ is a projection of $x \in \mathbb{R}^n$ on $S_n(1)$ and the gradient $\nabla W(p^k, q^k)$ is taken w.r.t. the first argument. One can show that `restarted-SGD` (`R-SGD`) from [88] that using biased stochastic gradients (see also [52, 84, 89]) $\widetilde{\nabla} W(p, q)$ such that

$$k\widetilde{\nabla} W(p, q) - \nabla W(p, q)k_2 \le \delta \tag{4.104}$$

for some $\delta \ge 0$ and for all $p, q \in S_n(1)$ after N calls of this oracle produces such a point p^N that with probability at least $1 - \beta$ the following inequalities hold:

$$W(p^N) - W(p^*) = O\left(\frac{nkCk_1^2 \ln(N/\beta)}{\mu N} + \delta\right) \tag{4.105}$$

and, as a consequence of μ-strong convexity of $W(p, q)$ for all q,

$$kp^N - p^*k_2 = O\left(\sqrt{\frac{nkCk_1^2 \ln(N/\beta)}{\mu^2 N}} + \frac{\delta}{\mu}\right). \tag{4.106}$$

That is, to guarantee

$$kp^N - p^*k_2 \le \varepsilon \tag{4.107}$$

with probability at least $1 - \beta$, `R-SGD` requires

$$\widetilde{O}\left(\frac{nkCk_1^2}{\mu^2\varepsilon^2}\right) \quad \widetilde{\nabla} W(p, q) \text{ oracle calls} \tag{4.108}$$

under additional assumption that $\delta = O(\mu\varepsilon^2)$.

However, it is computationally hard problem to find $\nabla W(p, q)$ with high-accuracy, i.e. find $\widetilde{\nabla} W(p, q)$ satisfying (4.104) with $\delta = O(\mu\varepsilon^2)$. Taking into account the relation (4.103) we get that it is needed to solve the problem (4.101) with accuracy $\delta = O(\mu\varepsilon^2)$ in terms of the distance to the optimum. i.e. it is needed to find such $\widetilde{\lambda}$ that $k\widetilde{\lambda} - \lambda^*k_2 \le \delta$ and set $\widetilde{\nabla} W(p, q) = \widetilde{\lambda}$. Using variants of Sinkhorn algorithm [48, 90, 91] one can show

[84] that R-SGD finds point p^N such that (4.107) holds with probability at least $1 - \beta$ and it requires

$$\widetilde{O}\left(\frac{n^3 kCk_1^2}{\mu^2\varepsilon^2}\min\left\{\exp\left(\frac{kCk_1}{\mu}\right)\left(\frac{kCk_1}{\mu} + \ln\left(\frac{kCk_1}{\gamma\mu^2\varepsilon^4}\right)\right), \sqrt{\frac{n}{\gamma\mu^3\varepsilon^4}}\right\}\right) \tag{4.109}$$

arithmetical operations.

4.8.3. SAA Approach

Now let us assume that large enough collection of samples $q^1, \ldots, q^m$ is available. Our goal is to find such $p \in S_n(1)$ that $\|\hat{p} - p^*\|_2 \leq \varepsilon$ with high probability, i.e. ε-approximation of the population barycenter, via solving empirical barycenter problem (4.100). This approach is called Stochastic Average Approximation (SAA). Since $W(p, q^i)$ is μ-strongly convex and M-Lipschitz in p with $M = \widetilde{O}(\sqrt{n}kCk_1)$ for all $i = 1, \ldots, m$ we can conclude that with probability $\geq 1 - \beta$

$$W(\hat{p}^*) - W(p^*) \overset{(4.5)}{=} O\left(\frac{nkCk_1^2 \ln(m)\ln(m/\beta)}{\mu m} + \sqrt{\frac{nkCk_1^2 \ln(1/\beta)}{m}}\right) \tag{4.110}$$

where we use that the diameter of $S_n(1)$ is $O(1)$. Moreover, in [7] it was shown that one can guarantee that with probability $\geq 1 - \beta$

$$W(\hat{p}^*) - W(p^*) \overset{(4.5)}{=} O\left(\frac{nkCk_1^2}{\beta\mu m}\right). \tag{4.111}$$

Taking advantages of both inequalities we get that if

$$m = \widetilde{}\left(\min\left\{\max\left\{\frac{nkCk_1^2}{\mu^2\varepsilon^2}, \frac{nkCk_1^2}{\mu^2\varepsilon^4}\right\}, \frac{nkCk_1^2}{\beta\mu^2\varepsilon^2}\right\}\right) = \widetilde{}\left(n\min\left\{\frac{kCk_1^2}{\mu^2\varepsilon^4}, \frac{kCk_1^2}{\beta\mu^2\varepsilon^2}\right\}\right) \tag{4.112}$$

then with probability at least $1 - \frac{\beta}{2}$

$$\|\hat{p}^* - p^*\|_2 \leq \sqrt{\frac{2}{\mu}\left(W(\hat{p}^*) - W(p^*)\right)} \overset{(4.110);(4.111);(4.112)}{\leq} \frac{\varepsilon}{2}. \tag{4.113}$$

Assuming that we have such $\hat{p} \in S_n(1)$ that with probability at least $1 - \frac{\beta}{2}$ the inequality

$$\|\hat{p} - \hat{p}^*\|_2 \leq \frac{\varepsilon}{2} \tag{4.114}$$

holds, we apply the union bound and get that with probability $\geq 1 - \beta$

$$\|\hat{p} - p^*\|_2 \leq \|\hat{p} - \hat{p}^*\|_2 + \|\hat{p}^* - p^*\|_2 \leq \varepsilon. \tag{4.115}$$

It remains to describe the approach that finds such $\hat{p} \in S_n(1)$ that satisfies (4.115) with probability at least $1 - \beta$. Recall that in this subsection we consider the following problem

$$\hat{W}(p) = \frac{1}{m}\sum_{i=1}^{m} W(p, q^i) \to \min_{p \in S_n(1)}. \tag{4.116}$$

For each summand $W(p, q^i)$ in the sum above we have the explicit formula (4.102) for the dual function $W^*_{q^i;}(\lambda)$. Note that one can compute the gradient of $W^*_{q^i;}(\lambda)$ via $O(n^2)$ arithmetical operations. What is more, $W^*_{q^i;}(\lambda)$ has a finite-sum structure, so, one can sample j-th component of q^i with probability q^i_j and get stochastic gradient

$$\nabla W^*_{q^i;}(\lambda, j) = \mu \nabla \left(\ln \left(\frac{1}{q^i_j} \sum_{i=1}^{n} \exp\left(\frac{-C_{ij} + \lambda_i}{\mu} \right) \right) \right) \tag{4.117}$$

which requires $O(n)$ arithmetical operations to be computed.

We start with the simple situation. Assume that each measures q^i are stored on m separate machines that form some network with Laplacian matrix $\overline{W} \in \mathbb{R}^{m \times m}$. For this scenario we can apply the dual approach described in Section 4.6 and apply bounds from Tables 4.5 and 4.6. If for all $i = 1, \ldots, m$ the i-th node computes the full gradient of dual functions $W_{q^i;}$ at each iteration then in order to find such a point $\hat{p}$ that with probability at least $1 - \frac{1}{2}$

$$\hat{W}(\hat{p}) - \hat{W}(p^*) \leq \varepsilon, \tag{4.118}$$

where $W = \overline{W} \otimes I_n$, this approach requires $\widetilde{O}\left(\sqrt{\frac{nkCk_1^2}{\varepsilon}\chi(W)}\right)$ communication rounds and $\widetilde{O}\left(n^{2.5}\sqrt{\frac{kCk_1^2}{\varepsilon}\chi(W)}\right)$ arithmetical operations per node to find gradients $\nabla W^*_{q^i;}(\lambda)$. If instead of full gradients workers use stochastic gradients $\nabla W^*_{q^i;}(\lambda, j)$ defined in (4.117) and these stochastic gradients have light-tailed distribution, i.e. satisfy the condition (4.90) with parameter $\sigma > 0$, then to guarantee (4.118) with probability $\geq 1 - \frac{1}{2}$ the aforementioned approach needs the same number of communications rounds and $\widetilde{O}\left(n \max\left\{\sqrt{\frac{nkCk_1^2}{\varepsilon}\chi(W)}, \frac{m^2 nkCk_1^2}{\varepsilon^2}\chi(W)\right\}\right)$ arithmetical operations per node to find gradients $\nabla W^*_{q^i;}(\lambda, j)$. Using μ-strong convexity of $W(p, q^i)$ for all $i = 1, \ldots, m$ and taking $\varepsilon = \frac{\varepsilon^2}{8}$ we get that our approach finds such a point $\hat{p}$ that satisfies (4.114) with probability at least $1 - \frac{1}{2}$ using

$$\widetilde{O}\left(\frac{\sqrt{\overline{n}kCk_1}}{\mu\varepsilon}\sqrt{\chi(W)}\right) \quad \text{communication rounds} \tag{4.119}$$

and

$$\widetilde{O}\left(n^{2.5}\frac{kCk_1}{\mu\varepsilon}\sqrt{\chi(W)}\right) \tag{4.120}$$

arithmetical operations per node to find gradients in the deterministic case and

$$\widetilde{O}\left(n \max\left\{\frac{\sqrt{\overline{n}kCk_1}}{\mu\varepsilon}\sqrt{\chi(W)}, \frac{m\sigma^2 nkCk_1^2}{\mu^2\varepsilon^4}\chi(W)\right\}\right)$$

arithmetical operations per node to find stochastic gradients in the stochastic case. However, the state-of-the-art theory of learning states (see (4.112)) that m should so large that in the stochastic case the second term in the bound for arithmetical operations typically dominates the first term and the dimensional dependence reduction from $n^{2.5}$ in the deterministic case to $n^{1.5}$ in the stochastic case is typically negligible in comparison with how much $\frac{m\ \sqrt{\overline{n}kCk_1^2}}{2\,n^4}\chi(W)$ is larger than $\frac{kCk_1}{n}\sqrt{\chi(W)}$. That is, our theory says that it is better to use full gradients in the particular example considered in this section (see also Section 4.7). Therefore, further in the section we will assume that $\sigma^2 = 0$, i.e. workers use full gradients of dual functions $W^*_{q^i;}(\lambda)$.

However, bounds (4.119)-(4.120) were obtained under very restrictive at the first sight assumption that we have m workers and each worker stores only one measure which is unrealistic. One can relax this assumption in the following way. Assume that we have $\hat{l} < m$ machines connected in a network with Laplacian matrix $\hat{W}$ and j-th machine stores $\hat{m}_j \geq 1$ measures for $j = 1, \ldots, \hat{l}$ and $\sum_{j=1}^{\hat{l}} \hat{m}_j = m$. Next, for j-th machine we introduce $\hat{m}_j$ virtual workers also connected in some network that j-th machine can emulate along with communication between virtual workers and for every virtual worker we arrange one measure, e.g. it can be implemented as an array-like data structure with some formal rules for exchanging the data between cells that emulates communications. We also assume that inside the machine we can set the preferable network for the virtual nodes in such a way that each machine emulates communication between virtual nodes and computations inside them fast enough. Let us denote the Laplacian matrix of the obtained network of m virtual nodes as $\overline{W}$. Then, our approach finds such a point $\hat{p}$ that satisfies (4.114) with probability at least $1 - \frac{1}{2}$ using

$$\widetilde{O}\left(\underbrace{\left(\max_{j=1;\ldots;\hat{l}} T_{\mathrm{cm};j}\right)}_{T_{\mathrm{cm,max}}} \frac{\sqrt{\overline{n}kCk_1}}{\mu\varepsilon}\sqrt{\chi(W)}\right) \tag{4.121}$$

time to perform communications and

$$\widetilde{O}\left(\underbrace{\left(\max_{j=1,\dots,m} T_{\mathsf{cp};j}\right)}_{T_{\mathsf{cp},\max}} n^{2.5}\frac{kCk_1}{\mu\varepsilon}\sqrt{\chi(W)}\right) \tag{4.122}$$

time for arithmetical operations per machine to find gradients where $T_{\mathsf{cm};j}$ is time needed for j-th machine to emulate communication between corresponding virtual nodes at each iteration and $T_{\mathsf{cp};j}$ is time required by j-th machine to perform 1 arithmetical operation for all corresponding virtual nodes in the gradients computation process at each iteration. For example, if we have only one machine and network of virtual nodes forms a complete graph than $\chi(W) = 1$, but $T_{\mathsf{cm};\max}$ and $T_{\mathsf{cp};\max}$ can be large and to reduce the running time one should use more powerful machine. In contrast, if we have m machines connected in a star-graph than $T_{\mathsf{cm};\max}$ and $T_{\mathsf{cp};\max}$ will be much smaller, but $\chi(W)$ will be of order m which is large. Therefore, it is very important to choose balanced architecture of the network at least for virtual nodes per machine if it is possible. This question requires a separate thorough study and lies out of scope of this paper.

4.8.4. SA vs SAA comparison

Recall that in SA approach we assume that it is possible to sample new measures in online regime which means that the computational process is performed on one machine, whereas in SAA approach we assume that large enough collection of measures is distributed among the network of machines that form some computational network. In practice measures from $S_n(1)$ correspond to some images. As one can see from the complexity bounds, both SA and SAA approaches require large number of samples to learn the population barycenter defined in (4.99). If these samples are images, then they typically cannot be stored in RAM of one computer. Therefore, it is natural to use distributed systems to store the data.

Now let us compare complexity bounds for SA and SAA. We summarize them in Table 4.7. When the communication is fast enough and μ is small we typically have that SAA approach significantly outperforms SA approach in terms of the complexity as well even for communication architectures with big $\chi(W)$. Therefore, for balanced architecture one can expect that SAA approach will outperform SA even more.

To conclude, we state that population barycenter computation is a natural example

Approach	Complexity
SA	$\widetilde{O}\left(\frac{n^3 kCk_1^2}{2\varepsilon^2}\min\left\{\exp\left(\frac{kCk_1}{\varepsilon}\right)\left(\frac{kCk_1}{\varepsilon}+\ln\left(\frac{kCk_1}{2\varepsilon^4}\right)\right), \sqrt{\frac{n}{3\varepsilon^4}}\right\}\right)$ arithmetical operations
SA, the 2-d term is smaller	$\widetilde{O}\left(\frac{n^{3.5}kCk_1^2}{p\varepsilon^{3.5}\varepsilon^4}\right)$ arithmetical operations
SAA	$\widetilde{O}\left(T_{\text{cm;max}}\frac{\sqrt{n}kCk_1}{\varepsilon}\sqrt{\chi(W)}\right)$ time to perform communications, $\widetilde{O}\left(T_{\text{cp;max}}n^{2.5}\frac{kCk_1}{\varepsilon}\sqrt{\chi(W)}\right)$ time for arithmetical operations per machine, where $m = \widetilde{\Theta}\left(n\min\left\{\frac{kCk_1^2}{2\varepsilon^4}, \frac{kCk_1^2}{2\varepsilon^2}\right\}\right)$
SAA, $\chi(W) = \Theta(m)$, $T_{\text{cm;max}} = O(1)$, $T_{\text{cp;max}} = O(1)$, $\sqrt{p}\overline{\beta} \geq \varepsilon$	$\widetilde{O}\left(\frac{nkCk_1^2}{p\varepsilon^2\varepsilon^2}\right)$ communication rounds, $\widetilde{O}\left(\frac{n^3 kCk_1^2}{p\varepsilon^2\varepsilon^2}\right)$ arithmetical operations per machine

Table 4.7: Complexity bounds for SA and SAA approaches for computation of population barycenter defined in (4.99) with accuracy ε. The third row states the complexity bound for SA approach when the second term under the minimum in (4.109) is dominated by the first one, e.g. when μ is small enough. The last row corresponds to the case when $T_{\text{cm;max}} = O(1)$, $T_{\text{cp;max}} = O(1)$, $\sqrt{p}\overline{\beta} \geq \varepsilon$, e.g. $\beta = 0.01$ and $\varepsilon \leq 0.1$, and the communication network is star-like, which implies $\chi(W) = \Theta(m)$

when it is typically much more preferable to use distributed algorithms with dual oracle instead of SA approach in terms of memory and complexity bounds.

4.9. Missing Proofs, Technical Lemmas and Auxiliary Results

4.9.1. Basic Facts

In this section we enumerate for convenience basic facts that we use many times in our proofs.

Fenchel-Young inequality. For all $a, b \in \mathbb{R}^n$ and $\lambda > 0$

$$|\langle a, b \rangle| \leq \frac{\|a\|_2^2}{2\lambda} + \frac{\lambda \|b\|_2^2}{2}. \tag{4.123}$$

Squared norm of the sum. For all $a, b \in \mathbb{R}^n$

$$\|a + b\|_2^2 \leq 2\|a\|_2^2 + 2\|b\|_2^2. \tag{4.124}$$

4.9.2. Useful Facts about Duality

This section contains several useful results that we apply in our analysis.

Lemma 4.9.1 ([71]). *Let y^* be the solution of (4.22) with the smallest ℓ_2-norm $R_y \stackrel{def}{=} \|y^*\|_2$. Then*

$$R_y^2 \leq \frac{\|\nabla f(x^*)\|_2^2}{\lambda_{\min}^+(A^\top A)}. \tag{4.125}$$

Lemma 4.9.2. *Consider the function $f(x)$ defined on a closed convex set $Q \subset \mathbb{R}^n$ and linear operator A such that $\mathrm{Ker} A \neq \{0\}$ and its dual function $\psi(y)$ defined as $\psi(y) = \max_{x \in Q} \{\langle y, Ax \rangle - f(x)\}$. Then*

$$\psi(y^*) = -f(x^*) \geq \langle y^*, A\hat{x} \rangle - f(\hat{x}) \quad \forall \hat{x} \in Q. \tag{4.126}$$

Proof. We have

$$\psi(y^*) = \left\langle y^*, Ax(A^\top y^*) \right\rangle - f\left(x(A^\top y^*)\right).$$

From Demyanov–Danskin theorem [63] we have that $\nabla \psi(y) = Ax(A^\top y)$ which implies

$$0 = \nabla \psi(y^*) = Ax(A^\top y^*).$$

Using this we get

$$\begin{aligned} -f\left(x(A^\top y^*)\right) &= \psi(y^*) = \max_{Ax = 0; x \in Q} \left\{ \underbrace{\langle y^*, Ax \rangle}_{=0} - f(x) \right\} \\ &= -f(x^*). \end{aligned}$$

Finally,

$$\psi(y^*) = -f(x^*) = \max_{Ax = 0; x \in Q} \{\langle y^*, Ax \rangle - f(x)\} \geq \langle y^*, A\hat{x} \rangle - f(\hat{x}).$$

□

4.9.3. Auxiliary Results

In this section, we present the results from other papers that we rely on in our proofs.

Lemma 4.9.3 (Lemma 2 from [92]). *For random vector $\xi \in \mathsf{R}^n$ following statements are equivalent up to absolute constant di erence in σ.*

1. *Tails:* $\mathbf{P}\{\|\xi\|_2 \geq \gamma\} \leq 2\exp\left(-\frac{\gamma^2}{2\sigma^2}\right)$ *$\forall\gamma \geq 0$.*
2. *Moments:* $(\mathbf{E}[\xi^p])^{\frac{1}{p}} \leq \sigma\sqrt{p}$ *for any positive integer p.*
3. *Super-exponential moment:* $\mathbf{E}\left[\exp\left(\frac{\|\xi\|_2^2}{\sigma^2}\right)\right] \leq \exp(1)$.

Lemma 4.9.4 (Corollary 8 from [92]). *Let $\{\xi_k\}_{k=1}^N$ be a sequence of random vectors with values in R^n such that for $k = 1,\ldots,N$ and for all $\gamma \geq 0$*

$$\mathbf{E}[\xi_k \mid \xi_1,\ldots,\xi_{k-1}] = 0, \quad \mathbf{E}[\|\xi_k\|_2 \geq \gamma \mid \xi_1,\ldots,\xi_{k-1}] \leq \exp\left(-\frac{\gamma^2}{2\sigma_k^2}\right) \quad \text{almost surely,}$$

where σ_k^2 belongs to the filtration $\sigma(\xi_1,\ldots,\xi_{k-1})$ for all $k = 1,\ldots,N$. Let $S_N = \sum_{k=1}^N \xi_k$. Then there exists an absolute constant C_1 such that for any fixed $\beta > 0$ and $B > b > 0$ with probability at least $1-\beta$:

$$\text{either } \sum_{k=1}^N \sigma_k^2 \geq B \quad \text{or} \quad \|S_N\|_2 \leq C_1\sqrt{\max\left\{\sum_{k=1}^N \sigma_k^2, b\right\}\left(\ln\frac{2n}{\beta} + \ln\ln\frac{B}{b}\right)}.$$

Lemma 4.9.5 (corollary of Theorem 2.1, item (ii) from [93]). *Let $\{\xi_k\}_{k=1}^N$ be a sequence of random vectors with values in R^n such that*

$$\mathbf{E}[\xi_k \mid \xi_1,\ldots,\xi_{k-1}] = 0 \text{ almost surely,} \quad k = 1,\ldots,N$$

and let $S_N = \sum_{k=1}^N \xi_k$. Assume that the sequence $\{\xi_k\}_{k=1}^N$ satisfy "light-tail" assumption:

$$\mathbf{E}\left[\exp\left(\frac{\|\xi_k\|_2^2}{\sigma_k^2}\right) \mid \xi_1,\ldots,\xi_{k-1}\right] \leq \exp(1) \text{ almost surely,} \quad k = 1,\ldots,N,$$

where $\sigma_1,\ldots,\sigma_N$ are some positive numbers. Then for all $\gamma \geq 0$

$$\mathbf{P}\left\{\|S_N\|_2 \geq \left(\sqrt{2} + \sqrt{2\gamma}\right)\sqrt{\sum_{k=1}^N \sigma_k^2}\right\} \leq \exp\left(-\frac{\gamma^2}{3}\right). \tag{4.127}$$

4.9.4. Missing Proofs from Section 4.3

Proof of Lemma 4.3.1

Since x^* is a minimizer of $g(x)$ on $\mathbb{R}^n$, we have $\nabla g(x^*) = 0$ and [1]

$$\|\nabla g(\hat{x})\|_2 \leq 2L(g(\hat{x}) - g(x^*)).$$

Next, using this, Cauchy-Schwarz inequality and definition of $\hat{x}$ we get

$$\langle \nabla g(\hat{x}), \hat{x} - x \rangle \leq \|\nabla g(\hat{x})\|_2 \cdot \|\hat{x} - x\|_2 \leq \sqrt{2L(g(\hat{x}) - g(x^*))}\|\hat{x} - x\|_2 \leq \sqrt{2L\delta}\|\hat{x} - x\|_2,$$

that concludes the proof.

Proof of Lemma 4.3.2

First of all, we prove by induction that $x^{k+1}, x^k, z^k \in B_{\widetilde{R}_k}(x^*)$ for $k = 0, 1, \ldots$. For $k = 0$ this is true since $x^0 = z^0$, $\widetilde{R}_0 = R_0 = \|z^0 - x^*\|$ and $x^1 = (A_0 x^0 + \alpha_{k+1} z^0)/A_1 = z^0$, since $A_0 = \alpha_0 = 0$ and $A_1 = \alpha_1$. Next, assume that $x^{k+1}, x^k, z^k \in B_{\widetilde{R}_k}(x^*)$ for some $k \geq 0$. By definition of R_{k+1} and $\widetilde{R}_{k+1}$ we have $z^{k+1} \in B_{R_{k+1}}(x^*) \quad B_{\widetilde{R}_{k+1}}(x^*)$. Due to the assumption that $x^k \in B_{R_k}(x^*) \quad B_{R_{k+1}}(x^*) \quad B_{\widetilde{R}_{k+1}}(x^*)$ and convexity of the $B_{\widetilde{R}_{k+1}}(x^*)$ we get that $x^{k+1} \in B_{\widetilde{R}_{k+1}}(x^*)$ since it is a convex combination of x^k and z^{k+1}, i.e. $x^{k+1} = (A_k x^k + \alpha_{k+1} z^{k+1})/A_{k+1}$. Similarly, x^{k+2} lies in the ball $B_{\widetilde{R}_{k+1}}(x^*)$ since it is a convex combination of x^{k+1} and z^{k+1}, i.e. $x^{k+1} = (A_k x^{k+1} + \alpha_{k+1} z^{k+1})/A_{k+1}$. That is, we proved that $x^{k+1}, x^k, z^k \in B_{\widetilde{R}_k}(x^*)$ for all non-negative integers k.

Since $z^{k+1} = \operatorname{argmin}_{z \in \mathbb{R}^n} g_{k+1}(z)$ and $g_{k+1}(z)$ is 1-strongly convex and $(\alpha_{k+1}L_h + 1)$-smooth we can apply Lemma 4.3.1 and get

$$\langle \nabla g_{k+1}(z^{k+1}), z^{k+1} - x^* \rangle \leq \sqrt{2(\alpha_{k+1}L_h + 1)\delta\|z^k - \hat{z}^{k+1}\|_2^2} \cdot \|z^{k+1} - x^*\|_2. \tag{4.128}$$

From 1-strong convexity of $g_{k+1}(z)$ we have

$$\|z^{k+1} - \hat{z}^{k+1}\|_2^2 \leq 2(g_{k+1}(z^{k+1}) - g_{k+1}(\hat{z}^{k+1})) \leq 2\delta\|z^k - \hat{z}^{k+1}\|_2^2.$$

Together with triangle inequality it implies that

$$\|z^k - \hat{z}^{k+1}\|_2 \leq \|z^k - x^*\|_2 + \|x^* - z^{k+1}\|_2 + \|z^{k+1} - \hat{z}^{k+1}\|_2 \leq 2\widetilde{R}_{k+1} + \sqrt{2\delta}\|z^k - \hat{z}^{k+1}\|_2,$$

and, after rearranging the terms,

$$\|z^k - \hat{z}^{k+1}\|_2 \leq \frac{2}{1 - \sqrt{2\delta}}\widetilde{R}_{k+1}. \tag{4.129}$$

Applying inequality above and (4.225) for the r.h.s. of (4.128) we obtain

$$\langle z^{k+1} - z^k + \alpha_{k+1} \nabla f(x^{k+1}) + \alpha_{k+1} \nabla h(z^{k+1}), z^{k+1} - x^* \rangle \leq \hat{\delta} \sqrt{k+2} \widetilde{R}_{k+1}^2, \qquad (4.130)$$

where we used

$$2\sqrt{\frac{2(\alpha_{k+1} L_h + 1)\delta}{(1 - \sqrt{2\delta})^2}} \overset{(4.225)}{\leq} 2\sqrt{\frac{2\left((k+2)L_h + 2(k+2)L\right)\delta}{2(1 - \sqrt{2\delta})^2 L}} \leq 2\sqrt{\frac{(L_h + 2L)\,\delta}{(1 - \sqrt{2\delta})^2 L}}\sqrt{k+2}$$

and $\hat{\delta} \overset{\text{def}}{=} 2\sqrt{\frac{(L_h + 2L)}{(1-\sqrt{2})^2 L}}$. Using this we get

$$\begin{aligned}
\alpha_{k+1} \langle \nabla f(x^{k+1}), z^k - x^* \rangle &= \alpha_{k+1} \langle \nabla f(x^{k+1}), z^k - z^{k+1} \rangle + \alpha_{k+1} \langle \nabla f(x^{k+1}), z^{k+1} - x^* \rangle \\
&\overset{(4.130)}{\leq} \alpha_{k+1} \langle \nabla f(x^{k+1}), z^k - z^{k+1} \rangle + \langle z^{k+1} - z^k, x^* - z^{k+1} \rangle \\
&\quad + \alpha_{k+1} \langle \nabla h(z^{k+1}), x^* - z^{k+1} \rangle + \hat{\delta} \sqrt{k+2} \widetilde{R}_{k+1}^2.
\end{aligned}$$

One can check via direct calculations that

$$\langle a, b \rangle = \frac{1}{2}\|a + b\|_2^2 - \frac{1}{2}\|a\|_2^2 - \frac{1}{2}\|b\|_2^2, \quad \forall\, a, b \in \mathbb{R}^n.$$

From the convexity of h

$$\langle \nabla h(z^{k+1}), x^* - z^{k+1} \rangle \leq h(x^*) - h(z^{k+1}).$$

Combining previous three inequalities we obtain

$$\begin{aligned}
\alpha_{k+1} \langle \nabla f(x^{k+1}), z^k - x^* \rangle &\leq \alpha_{k+1} \langle \nabla f(x^{k+1}), z^k - z^{k+1} \rangle - \frac{1}{2}\|z^k - z^{k+1}\|_2^2 + \frac{1}{2}\|z^k - x^*\|_2^2 \\
&\quad - \frac{1}{2}\|z^{k+1} - x^*\|_2^2 + \alpha_{k+1}\left(h(x^*) - h(z^{k+1})\right) + \hat{\delta} \sqrt{k+2} \widetilde{R}_{k+1}^2.
\end{aligned}$$

By definition of x^{k+1} and x^{k+1}

$$\begin{aligned}
x^{k+1} &= \frac{A_k x^k + \alpha_{k+1} z^{k+1}}{A_{k+1}} = \frac{A_k x^k + \alpha_{k+1} z^k}{A_{k+1}} + \frac{\alpha_{k+1}}{A_{k+1}}\left(z^{k+1} - z^k\right) \\
&= x^{k+1} + \frac{\alpha_{k+1}}{A_{k+1}}\left(z^{k+1} - z^k\right).
\end{aligned}$$

Together with the previous inequality and $A_{k+1} = 2L\alpha_{k+1}^2$, it implies

$$\begin{aligned}
\alpha_{k+1}\langle f'(x^{k+1}), z^k - x^*\rangle &\leq A_{k+1}\langle f'(x^{k+1}), x^{k+1} - x^{k+1}\rangle \\
&\quad - \frac{A_{k+1}^2}{2\alpha_{k+1}^2}\|x^{k+1} - x^{k+1}\|_2^2 + \frac{1}{2}\|z^k - x^*\|_2^2 - \frac{1}{2}\|z^{k+1} - x^*\|_2^2 \\
&\quad + \alpha_{k+1}\left(h(x^*) - h(z^{k+1})\right) + \delta\sqrt{k+2}\widetilde{R}_{k+1}^2 \\
&\leq A_{k+1}\left(\langle f'(x^{k+1}), x^{k+1} - x^{k+1}\rangle - \frac{2L}{2}\|x^{k+1} - x^{k+1}\|_2^2\right) \\
&\quad + \frac{1}{2}\|z^k - x^*\|_2^2 - \frac{1}{2}\|z^{k+1} - x^*\|_2^2 \\
&\quad + \alpha_{k+1}\left(h(x^*) - h(z^{k+1})\right) + \delta\sqrt{k+2}\widetilde{R}_{k+1}^2 \\
&\leq A_{k+1}(f(x^{k+1}) - f(x^{k+1})) + \frac{1}{2}\|z^k - x^*\|_2^2 - \frac{1}{2}\|z^{k+1} - x^*\|_2^2 \\
&\quad + \alpha_{k+1}\left(h(x^*) - h(z^{k+1})\right) + \delta\sqrt{k+2}\widetilde{R}_{k+1}^2 \qquad (4.131)
\end{aligned}$$

From the convexity of f we get

$$\langle f'(x^{k+1}), x^k - x^{k+1}\rangle \leq f(x^k) - f(x^{k+1}). \qquad (4.132)$$

By definition of x^{k+1} we have

$$\alpha_{k+1}\left(x^{k+1} - z^k\right) = A_k\left(x^k - x^{k+1}\right). \qquad (4.133)$$

Putting all together, we get

$$\begin{aligned}
\alpha_{k+1}\langle f'(x^{k+1}), x^{k+1} - x^*\rangle &= \alpha_{k+1}\langle f'(x^{k+1}), x^{k+1} - z^k\rangle \\
&\quad + \alpha_{k+1}\langle f'(x^{k+1}), z^k - x^*\rangle \\
&\overset{(4.133)}{=} A_k\langle f'(x^{k+1}), x^k - x^{k+1}\rangle \\
&\quad + \alpha_{k+1}\langle f'(x^{k+1}), z^k - x^*\rangle \\
&\overset{(4.131);(4.132)}{\leq} A_k\left(f(x^k) - f(x^{k+1})\right) \\
&\quad + A_{k+1}\left(f(x^{k+1}) - f(x^{k+1})\right) \\
&\quad + \frac{1}{2}\|z^k - x^*\|_2^2 - \frac{1}{2}\|z^{k+1} - x^*\|_2^2 \\
&\quad + \alpha_{k+1}\left(h(x^*) - h(z^{k+1})\right) + \delta\sqrt{k+2}\widetilde{R}_{k+1}^2.
\end{aligned}$$

Rearranging the terms and using $A_{k+1} = A_k + \alpha_{k+1}$, we obtain

$$\begin{aligned}
A_{k+1}f(x^{k+1}) - A_k f(x^k) &\leq \alpha_{k+1}\left(f(x^{k+1}) + \langle f'(x^{k+1}), x^* - x^{k+1}\rangle\right) + \frac{1}{2}\|z^k - x^*\|_2^2 \\
&\quad - \frac{1}{2}\|z^{k+1} - x^*\|_2^2 + \alpha_{k+1}\left(h(x^*) - h(z^{k+1})\right) + \delta\sqrt{k+2}\widetilde{R}_{k+1}^2,
\end{aligned}$$

and after summing these inequalities for $k = 0, \ldots, N-1$ and applying convexity of f, i.e. inequality $\langle f'(x^{k+1}), x^* - x^{k+1} \rangle \leq f(x^*) - f(x^{k+1})$, we get

$$A_N f(x^N) \leq \frac{1}{2}R_0^2 - \frac{1}{2}R_N^2 + A_N f(x^*) + A_N h(x^*) - \sum_{k=0}^{N-1} \alpha_{k+1} h(z^{k+1}) + \hat{\delta} \sum_{k=0}^{N-1} \sqrt{k+2}\, \widetilde{R}_{k+1}^2,$$

where we used that $A_0 = 0$. Finally, convexity of h and definition of x^{k+1}, i.e. $x^{k+1} = (A_k x^k + \alpha_{k+1} z^{k+1}) / A_{k+1}$, implies

$$A_N h(x^N) \leq A_{N-1} h(x^{N-1}) + \alpha_N h(z^N).$$

Applying this inequality for $A_{N-1}h(x^{N-1}), A_{N-2}h(x^{N-2}), \ldots, A_1 h(x^1)$ in a sequence we get

$$A_N h(x^N) \leq A_0 h(x^0) + \sum_{k=0}^{N-1} \alpha_{k+1} h(z^{k+1}) = \sum_{k=0}^{N-1} \alpha_{k+1} h(z^{k+1}),$$

which implies

$$A_N \left(F(x^N) - F(x^*)\right) \leq \frac{1}{2}R_0^2 - \frac{1}{2}R_N^2 + \hat{\delta} \sum_{k=0}^{N-1} \sqrt{k+2}\, \widetilde{R}_{k+1}^2,$$

that finishes the proof.

Proof of Theorem 4.3.1

Lemma 4.3.2 implies that

$$A_l \left(F(x^l) - F(x^*)\right) \leq \frac{1}{2}R_0^2 - \frac{1}{2}R_l^2 + \hat{\delta} \sum_{k=0}^{l-1} \sqrt{k+2}\, \widetilde{R}_{k+1}^2 \tag{4.134}$$

for $l = 1, 2, \ldots, N$. Since $F(x^l) \geq F(x^*)$ for each l and $\hat{\delta} \leq \frac{C}{(N+1)^{3/2}}$ we get the recurrence

$$R_l^2 \leq R_0^2 + \frac{2C}{(N+1)^{3/2}} \sum_{k=0}^{l-1} (k+2)^{1/2} \widetilde{R}_{k+1}^2, \quad \forall l = 1, \ldots, N.$$

Note that the r.h.s. of the previous inequality is non-decreasing function of l. Let us define $\hat{l}$ as the largest integer such that $\hat{l} \leq l$ and $\widetilde{R}_{\hat{l}} = R_{\hat{l}}$. Then $R_{\hat{l}} = \widetilde{R}_{\hat{l}} = \widetilde{R}_{\hat{l}+1} = \ldots = \widetilde{R}_l$ and, as a consequence,

$$\widetilde{R}_l^2 \leq \widetilde{R}_0^2 + \frac{2C}{(N+1)^{3/2}} \sum_{k=0}^{l-1} (k+2)^{1/2} \widetilde{R}_{k+1}^2, \quad \forall l = 1, \ldots, N. \tag{4.135}$$

Using Lemma 4.9.12 we get that $\widetilde{R}_l \le 2R_0^2$ for all $l = 1, \dots, N$. We plug this inequality together with $\delta \le \frac{C}{(N+1)^{3/2}} \le \frac{1}{4(N+1)^{3/2}}$ and $R_N^2 \ge 0$ in (4.134) and get

$$\begin{aligned} A_N(F(x^N) - F(x^*)) &\le \frac{1}{2}R_0^2 + \frac{4R_0^2}{4(N+1)^{3=2}} \sum_{k=0}^{N-1} (k+2)^{1=2} \\ &\le \frac{3}{2}R_0^2, \end{aligned}$$

which concludes the proof.

Proof of Corollary 4.3.1

The first part of the corollary follows from (4.18) and Lemma 4.9.9. Relation (4.20) follows from the definition of $\hat{\delta}$ and $\hat{\delta} \le \frac{C}{(N+1)^{3/2}}$. Indeed, since $\hat{\delta} \stackrel{\text{def}}{=} 2\sqrt{\frac{(L_h+2L)}{(1-\overline{2})^2 L}}$ and $C \le \frac{1}{4}$ we get that

$$\delta \le \frac{C^2(1-\overline{2\delta})^2 L}{4(L_h+2L)(N+1)^3} \le \frac{L}{64(L_h+2L)N^3} \le \frac{1}{64}\frac{L}{(L_h+L)N^3}.$$

4.9.5. Missing Proofs from Section 4.4

Proof of Theorem 4.4.1

By definition of F

$$\begin{aligned} F(x^N) - \min_{x \in Q} F(x) &= f(x^N) + \frac{R_y^2}{\varepsilon}\|Ax^N\|_2^2 - \min_{x \in Q}\left\{f(x) + \frac{R_y^2}{\varepsilon}\|Ax\|_2^2\right\} \\ &\ge f(x^N) + \frac{R_y^2}{\varepsilon}\|Ax^N\|_2^2 - \min_{Ax=0; x \in Q}\left\{f(x) + \frac{R_y^2}{\varepsilon}\|Ax\|_2^2\right\} \\ &= f(x^N) - \min_{Ax=0; x \in Q} f(x) + \frac{R_y^2}{\varepsilon}\|Ax^N\|_2^2, \end{aligned}$$

which implies

$$f(x^N) - f(x^*) + \frac{R_y^2}{\varepsilon}\|Ax^N\|_2^2 \stackrel{(4.26)}{\le} \varepsilon, \tag{4.136}$$

where x^* is an arbitrary solution of (4.21). Taking inequality $\|Ax^N\|_2^2 \ge 0$ into account we get the first part of (4.27). From Cauchy-Schwarz inequality we obtain

$$-R_y\|Ax^N\|_2 \le -\|y^*\|_2 \cdot \|Ax^N\|_2 \le \langle y^*, Ax^N\rangle \stackrel{(4.126)}{\le} f(x^N) - f(x^*).$$

Together with (4.136) it gives us quadratic inequality on $R_y\|Ax^N\|_2$:

$$-R_y\|Ax^N\|_2 + \frac{R_y^2}{\varepsilon}\|Ax^N\|_2^2 \le \varepsilon.$$

Therefore, $R_y\|Ax^N\|_2$ should be less then the greatest root of the corresponding quadratic equation, i.e. $R_y\|Ax^N\|_2 \le \frac{1+\sqrt{5}}{2}\varepsilon < 2\varepsilon$.

Proof of Theorem 4.4.2

Note that $h(x)$ is convex and L_h-smooth in $\mathbb{R}^n$ with $L_h = 2R_y^2 \lambda_{\max}(A^\top A)/\varepsilon$ since $\nabla h(x) = 2R_y^2 A^\top Ax/\varepsilon$ and

$$\begin{aligned} \|\nabla h(x) - \nabla h(y)\|_2 &= \frac{2R_y^2}{\varepsilon}\|A^\top A(x-y)\|_2 \le \frac{2R_y^2}{\varepsilon}\|A^\top A\|_2 \cdot \|x-y\|_2 \\ &\le \frac{2R_y^2 \lambda_{\max}(A^\top A)}{\varepsilon}\|x-y\|_2 \end{aligned}$$

for all $x, y \in \mathbb{R}^n$. We can apply STM with inexact proximal step (STP_IPS) which is presented in Section 4.3 as Algorithm 2 to solve problem (4.25). Corollary 4.3.1 (see Section 4.3 in the Appendix; see also the text after the corollary) states that in order to get such x^N that satisfy (4.26) we should run STP_IPS for $N = O\left(\sqrt{LR^2/\varepsilon}\right)$ iterations with $\delta = O\left(\varepsilon^{3/2}/((L_h+L)^p \bar{L} R^3)\right)$, where $R = \|x^0 - x^*\|_2$, x^* is the closest to x^0 minimizer of F and δ is such that for all $k = 0, \ldots, N-1$ the auxiliary problem $g_{k+1}(z) \to \min_{z \in \mathbb{R}^n}$ for finding z^{k+1} is solved with accuracy $g_{k+1}(z^{k+1}) - g_{k+1}(\hat{z}^{k+1}) \le \delta\|z^k - \hat{z}^{k+1}\|_2^2$ where $g_{k+1}(z)$ is defined as (see also (4.16))

$$g_{k+1}(z) = \frac{1}{2}\|z^k - z\|_2^2 + \alpha_{k+1}\left(f(x^{k+1}) + \langle \nabla f(x^{k+1}), z - x^{k+1}\rangle + h(z)\right)$$

for $k = 0, 1, \ldots$ and $\hat{z}^{k+1} = \arg\min_{z \in \mathbb{R}^n} g_{k+1}(z)$. That is, if the auxiliary problem is solved accurate enough at each iteration, then number of iterations, i.e. number of calculations $\nabla f(x)$, corresponds to the optimal bound presented in Table 4.1.

However, in order to solve the auxiliary problem $\min_{z \in \mathbb{R}^n} g_{k+1}(z)$ one should run another optimization method as a subroutine, e.g. STM. Note that $\mathrm{Im}A = \mathrm{Im}A^\top = (\mathrm{Ker}A)^\perp$ and if the starting point for this problem is chosen as $z^k - \alpha_{k+1}\nabla f(x^{k+1})$ then the iterates of STM applied to solve problem $\min_{z \in \mathbb{R}^n} g_{k+1}(z)$ lie in $z^k - \alpha_{k+1}\nabla f(x^{k+1}) + (\mathrm{Ker}A)^\perp$ since $\nabla g_{k+1}(z) \in \mathrm{Im}(A)$ for all $z \in z^k - \alpha_{k+1}\nabla f(x^{k+1}) + (\mathrm{Ker}A)^\perp$ (one can prove it using simple induction, see Theorem 4.5.7 for the details of the proof of the similar result). Therefore, the auxiliary problem can be considered as a minimization of $(1 + 2\alpha_{k+1}R_y^2 \lambda^+_{\min}(A^\top A)/\varepsilon)$-strongly convex on $z^k - \alpha_{k+1}\nabla f(x^{k+1}) + (\mathrm{Ker}A)^\perp$ and $(1 + 2\alpha_{k+1}R_y^2 \lambda_{\max}(A^\top A)/\varepsilon)$-smooth on $\mathbb{R}^n$ function. Then, one can estimate the overall complexity of the auxiliary problem using the condition number of $g_{k+1}(z)$ on $z^k - \alpha_{k+1}\nabla f(x^{k+1}) + (\mathrm{Ker}A)^\perp$:

$$\frac{1 + 2\alpha_{k+1}R_y^2 \lambda_{\max}(A^\top A)/\varepsilon}{1 + 2\alpha_{k+1}R_y^2 \lambda^+_{\min}(A^\top A)/\varepsilon} \le \frac{\lambda_{\max}(A^\top A)}{\lambda^+_{\min}(A^\top A)} \stackrel{\text{def}}{=} \chi(A^\top A). \tag{4.137}$$

Assume that z^{k+1} is such that $g_{k+1}(z^{k+1}) - g_{k+1}(\hat{z}^{k+1}) \leq \tilde{\delta}\|z^k - \alpha_{k+1} r f(x^{k+1}) - \hat{z}^{k+1}\|_2$. Then

$$\begin{aligned}
\|z^k - \alpha_{k+1} r f(x^{k+1}) - \hat{z}^{k+1}\|_2 &\leq \|z^k - \hat{z}^{k+1}\|_2 + \alpha_{k+1}\|r f(x^{k+1})\|_2 \\
&\leq \|z^k - \hat{z}^{k+1}\|_2 + \alpha_{k+1}\|r f(x^{k+1}) - r f(x^*)\|_2 + \alpha_{k+1}\|r f(x^*)\|_2 \\
&\leq \|z^k - \hat{z}^{k+1}\|_2 + \alpha_{k+1}L\|x^{k+1} - x^*\|_2 + \alpha_{k+1}\|r f(x^*)\|_2 \\
&\overset{(4.225)}{\leq} \|z^k - \hat{z}^{k+1}\|_2 + \frac{k+2}{2}\widetilde{R}_{k+1} + \frac{k+2}{2L}\|r f(x^*)\|_2
\end{aligned}$$

and using the similar steps as in the proof of inequality (4.129) we get

$$\|z^k - \hat{z}^{k+1}\|_2 \leq \frac{\left(2 + \frac{(k+2)^{\frac{p}{2}}}{2}\right)\widetilde{R}_{k+1}}{1 - \sqrt{2\tilde{\delta}}} + \frac{(k+2)\|r f(x^*)\|_2}{2L\left(1 - \sqrt{2\tilde{\delta}}\right)}.$$

Combining previous two inequalities we conclude that

$$\begin{aligned}
\|z^k - \alpha_{k+1} r f(x^{k+1}) - \hat{z}^{k+1}\|_2 \leq &\left(\frac{2 + \frac{(k+2)^{\frac{p}{2}}}{2}}{1 - \sqrt{2\tilde{\delta}}} + \frac{k+2}{2}\right)\widetilde{R}_{k+1} \\
&+ \frac{k+2}{2L}\left(1 + \frac{1}{1 - \sqrt{2\tilde{\delta}}}\right)\|r f(x^*)\|_2.
\end{aligned}$$

It means that to achieve $g_{k+1}(z^{k+1}) - g_{k+1}(\hat{z}^{k+1}) \leq \delta\widetilde{R}_{k+1}^2$ with $\delta = O\left(\varepsilon^{3/2}/((L_h+L)^{p}\overline{L}R^3)\right)$ one can run STM to solve the auxiliary problem $g_{k+1}(z) \to \min_{z \in \mathbb{R}^n}$ for T iterations with the starting point $z^k - \alpha_{k+1} r f(x^{k+1})$ where

$$\begin{aligned}
T &= O\left(\sqrt{\chi(A^\top A)}\ln\left(\frac{L_{g_N}L^{3/2}\left(R_y^2\lambda_{\max}(A^\top A)/\varepsilon + L\right)R^3\left(R^2 + \|r f(x^*)\|_2^2/L^2\right)}{\varepsilon^{5/2}}\right)\right), \\
L_{g_N} &= 1 + \frac{2\alpha_{k+1}R_y^2\lambda_{\max}(A^\top A)}{\varepsilon} \overset{(4.19)+(4.225)}{=} O\left(\frac{R_y^2 R\lambda_{\max}(A^\top A)}{p\overline{L}\varepsilon^{3/2}}\right)
\end{aligned}$$

or, equivalently,

$$T = O\left(\sqrt{\chi(A^\top A)}\ln\left(\frac{\lambda_{\max}(A^\top A)L\left(R_y^2\lambda_{\max}(A^\top A)/\varepsilon + L\right)R_y^2R^4\left(R^2 + \|r f(x^*)\|_2^2/L^2\right)}{\varepsilon^4}\right)\right).$$

4.9.6. Missing Lemmas and Proofs from Section 4.5.1

Lemmas

The following lemma is rather technical and provides useful inequalities that show how biasedness of $r(y, \cdot)$ interacts with convexity and L-smoothness of ψ.

Lemma 4.9.6. *Assume that function $\psi(y)$ is convex and L -smooth on $\mathbb{R}^n$. Then for all $x, y \in \mathbb{R}^n$*

$$\psi(y) \geq \psi(x) + \left\langle \mathbf{E}\left[\nabla f(x, \xi^k)\right], y - x \right\rangle - \delta\|y - x\|_2, \tag{4.138}$$

$$\psi(y) \leq \psi(x) + \left\langle \mathbf{E}\left[\nabla f(x, \xi^k)\right], y - x \right\rangle + L\|y - x\|_2^2 + \frac{\delta^2}{2L}. \tag{4.139}$$

Proof. From the convexity of ψ we have

$$\begin{aligned}
\psi(x) - \psi(y) &\leq \langle \nabla \psi(x), x - y\rangle = \left\langle \mathbf{E}\left[\nabla f(x, \xi^k)\right], x - y \right\rangle + \left\langle \nabla \psi(x) - \mathbf{E}\left[\nabla f(x, \xi^k)\right], x - y \right\rangle \\
&\leq \left\langle \mathbf{E}\left[\nabla f(x, \xi^k)\right], x - y \right\rangle + \left\| \nabla \psi(x) - \mathbf{E}\left[\nabla f(x, \xi^k)\right] \right\|_2 \cdot \|x - y\|_2 \\
&\overset{(4.42)}{\leq} \left\langle \mathbf{E}\left[\nabla f(x, \xi^k)\right], x - y \right\rangle + \delta\|x - y\|_2,
\end{aligned}$$

which proves the inequality (4.138). Applying L-smoothness of $\psi(x)$ we get

$$\begin{aligned}
\psi(y) &\leq \psi(x) + \langle \nabla \psi(x), y - x\rangle + \frac{L}{2}\|y - x\|_2^2 \\
&= \psi(x) + \left\langle \mathbf{E}\left[\nabla f(x, \xi^k)\right], y - x \right\rangle + \left\langle \nabla \psi(x) - \mathbf{E}\left[\nabla f(x, \xi^k)\right], y - x \right\rangle + \frac{L}{2}\|y - x\|_2^2.
\end{aligned}$$

Due to Fenchel-Young inequality $\langle a, b\rangle \leq \frac{1}{2}\|a\|_2^2 + \frac{1}{2}\|b\|_2^2$, $a, b \in \mathbb{R}^n$, $\lambda > 0$,

$$\begin{aligned}
\left\langle \nabla \psi(x) - \mathbf{E}\left[\nabla f(x, \xi^k)\right], y - x \right\rangle &\leq \frac{1}{2L}\left\| \nabla \psi(x) - \mathbf{E}\left[\nabla f(x, \xi^k)\right] \right\|_2^2 + \frac{L}{2}\|y - x\|_2^2 \\
&\overset{(4.42)}{\leq} \frac{\delta^2}{2L} + \frac{L}{2}\|y - x\|_2^2.
\end{aligned}$$

Combining these two inequalities we get (4.139). □

Next, we will use the following notation: $\mathbf{E}_k[\cdot] = \mathbf{E}_{\xi^{k+1}}[\cdot]$ which denotes conditional mathematical expectation with respect to all randomness that comes from ξ^{k+1}.

Lemma 4.9.7 (see also Theorem 1 from [54]). *For each iteration of Algorithm 3 we have*

$$\begin{aligned}
A_N \psi(y^N) \leq\ & \frac{1}{2}\|z - z^0\|_2^2 - \frac{1}{2}\|z - z^N\|_2^2 \\
&+ \sum_{k=0}^{N-1} \alpha_{k+1}\left(\psi(y^{k+1}) + \langle \nabla f(y^{k+1}, \xi^{k+1}), z - y^{k+1}\rangle\right) \\
&+ \sum_{k=0}^{N-1} A_k \left\langle \nabla f(y^{k+1}, \xi^{k+1}) - \mathbf{E}_k\left[\nabla f(y^{k+1}, \xi^{k+1})\right], y^k - y^{k+1} \right\rangle \\
&+ \sum_{k=0}^{N-1} \frac{A_{k+1}}{2\tilde{L}} \left\| \mathbf{E}_k\left[\nabla f(y^{k+1}, \xi^{k+1})\right] - \nabla f(y^{k+1}, \xi^{k+1}) \right\|_2^2 \\
&+ \delta \sum_{k=0}^{N-1} A_k \|y^k - y^{k+1}\|_2 + \delta^2 \sum_{k=0}^{N-1} \frac{A_{k+1}}{\tilde{L}},
\end{aligned} \tag{4.140}$$

for arbitrary $z \in \mathbb{R}^n$.

Proof. The proof of this lemma follows a similar way as in the proof of Theorem 1 from [54]. We can rewrite the update rule for z^k in the equivalent way:

$$z^{k+1} = \operatorname*{argmin}_{z \in \mathbb{R}^n} \left\{ \alpha_{k+1} \langle \nabla f(\psi^{k+1}, \ ^{k+1}), z - \psi^{k+1} \rangle + \frac{1}{2} \|z - z^k\|_2^2 \right\}.$$

From the optimality condition we have that for all $z \in \mathbb{R}^n$

$$\langle z^{k+1} - z^k + \alpha_{k+1} \nabla f(\psi^{k+1}, \ ^{k+1}), z - z^{k+1} \rangle \geq 0. \tag{4.141}$$

Using this we get

$$\begin{aligned}
&\alpha_{k+1} \langle \nabla f(\psi^{k+1}, \ ^{k+1}), z^k - z \rangle \\
&\qquad = \alpha_{k+1} \langle \nabla f(\psi^{k+1}, \ ^{k+1}), z^k - z^{k+1} \rangle + \alpha_{k+1} \langle \nabla f(\psi^{k+1}, \ ^{k+1}), z^{k+1} - z \rangle \\
&\qquad \overset{(4.141)}{\leq} \alpha_{k+1} \langle \nabla f(\psi^{k+1}, \ ^{k+1}), z^k - z^{k+1} \rangle + \langle z^{k+1} - z^k, z - z^{k+1} \rangle.
\end{aligned}$$

One can check via direct calculations that

$$\langle a, b \rangle \leq \frac{1}{2} \|a + b\|_2^2 - \frac{1}{2} \|a\|_2^2 - \frac{1}{2} \|b\|_2^2, \quad \forall\, a, b \in \mathbb{R}^n.$$

Combining previous two inequalities we obtain

$$\begin{aligned}
\alpha_{k+1} \langle \nabla f(\psi^{k+1}, \ ^{k+1}), z^k - z \rangle \ \leq \ & \alpha_{k+1} \langle \nabla f(\psi^{k+1}, \ ^{k+1}), z^k - z^{k+1} \rangle - \frac{1}{2} \|z^k - z^{k+1}\|_2^2 \\
& + \frac{1}{2} \|z^k - z\|_2^2 - \frac{1}{2} \|z^{k+1} - z\|_2^2.
\end{aligned}$$

By definition of y^{k+1} and ψ^{k+1}

$$\begin{aligned}
y^{k+1} &= \frac{A_k y^k + \alpha_{k+1} z^{k+1}}{A_{k+1}} = \frac{A_k y^k + \alpha_{k+1} z^k}{A_{k+1}} + \frac{\alpha_{k+1}}{A_{k+1}} \left(z^{k+1} - z^k \right) \\
&= \psi^{k+1} + \frac{\alpha_{k+1}}{A_{k+1}} \left(z^{k+1} - z^k \right).
\end{aligned}$$

Together with previous inequality, it implies

$$\begin{aligned}
\alpha_{k+1}\langle \nabla f(\psi(y^{k+1}), \nabla\psi(y^{k+1})), z^k - z\rangle &\le A_{k+1}\langle \nabla f(\psi(y^{k+1}), \nabla\psi(y^{k+1})), y^{k+1} - y^{k+1}\rangle \\
&\quad - \frac{A_{k+1}^2}{2\alpha_{k+1}^2}\|y^{k+1} - y^{k+1}\|_2^2 + \frac{1}{2}\|z^k - z\|_2^2 - \frac{1}{2}\|z^{k+1} - z\|_2^2 \\
&\le A_{k+1}\Bigg(\langle \nabla f(\psi(y^{k+1}), \nabla\psi(y^{k+1})), y^{k+1} - y^{k+1}\rangle \\
&\qquad - \frac{2L}{2}\|y^{k+1} - y^{k+1}\|_2^2\Bigg) \\
&\quad + \frac{1}{2}\|z^k - z\|_2^2 - \frac{1}{2}\|z^{k+1} - z\|_2^2 \\
&= A_{k+1}\Bigg(\left\langle \mathbf{E}_k\left[\nabla f(\psi(y^{k+1}), \nabla\psi(y^{k+1}))\right], y^{k+1} - y^{k+1}\right\rangle \\
&\qquad - \frac{2L}{2}\|y^{k+1} - y^{k+1}\|_2^2\Bigg) \\
&\quad + A_{k+1}\left\langle \nabla f(\psi(y^{k+1}), \nabla\psi(y^{k+1})) - \mathbf{E}_k\left[\nabla f(\psi(y^{k+1}), \nabla\psi(y^{k+1}))\right], y^{k+1} - y^{k+1}\right\rangle \\
&\quad + \frac{1}{2}\|z^k - z\|_2^2 - \frac{1}{2}\|z^{k+1} - z\|_2^2.
\end{aligned}$$

From Fenchel-Young inequality $\langle a, b\rangle \le \frac{1}{2}\|a\|_2^2 + \frac{1}{2}\|b\|_2^2$, $a, b \in \mathbb{R}^n$, $\lambda > 0$, we have

$$\begin{aligned}
&\left\langle \nabla f(\psi(y^{k+1}), \nabla\psi(y^{k+1})) - \mathbf{E}_k\left[\nabla f(\psi(y^{k+1}), \nabla\psi(y^{k+1}))\right], y^{k+1} - y^{k+1}\right\rangle \\
&\qquad \le \frac{1}{2L}\left\|\nabla f(\psi(y^{k+1}), \nabla\psi(y^{k+1})) - \mathbf{E}_k\left[\nabla f(\psi(y^{k+1}), \nabla\psi(y^{k+1}))\right]\right\|_2^2 + \frac{L}{2}\|y^{k+1} - y^{k+1}\|_2^2.
\end{aligned}$$

Using this, we get

$$\begin{aligned}
\alpha_{k+1}\langle \nabla f(\psi(y^{k+1}), \nabla\psi(y^{k+1})), z^k - z\rangle &\le A_{k+1}\Bigg(\left\langle \mathbf{E}_k\left[\nabla f(\psi(y^{k+1}), \nabla\psi(y^{k+1}))\right], y^{k+1} - y^{k+1}\right\rangle \\
&\qquad - \frac{L}{2}\|y^{k+1} - y^{k+1}\|_2^2\Bigg) \\
&\quad + \frac{A_{k+1}}{2L}\left\|\nabla f(\psi(y^{k+1}), \nabla\psi(y^{k+1})) - \mathbf{E}_k\left[\nabla f(\psi(y^{k+1}), \nabla\psi(y^{k+1}))\right]\right\|_2^2 \\
&\quad + \frac{1}{2}\|z^k - z\|_2^2 - \frac{1}{2}\|z^{k+1} - z\|_2^2 \\
&\overset{(4.139)}{\le} A_{k+1}\left(\psi(y^{k+1}) - \psi(y^{k+1}) + \frac{\delta^2}{L}\right) \\
&\quad + \frac{1}{2}\|z^k - z\|_2^2 - \frac{1}{2}\|z^{k+1} - z\|_2^2 \qquad (4.142) \\
&\quad + \frac{A_{k+1}}{2L}\left\|\nabla f(\psi(y^{k+1}), \nabla\psi(y^{k+1})) - \mathbf{E}_k\left[\nabla f(\psi(y^{k+1}), \nabla\psi(y^{k+1}))\right]\right\|_2^2.
\end{aligned}$$

With Lemma 4.9.6 in hand, we have

$$\begin{aligned}
\langle \nabla f(\tilde{y}^{k+1}, \tilde{z}^{k+1}), y^k - \tilde{y}^{k+1} \rangle &= \left\langle \mathbf{E}_k \left[\nabla f(\tilde{y}^{k+1}, \tilde{z}^{k+1}) \right], y^k - \tilde{y}^{k+1} \right\rangle \\
&\quad + \left\langle \nabla f(\tilde{y}^{k+1}, \tilde{z}^{k+1}) - \mathbf{E}_k \left[\nabla f(\tilde{y}^{k+1}, \tilde{z}^{k+1}) \right], y^k - \tilde{y}^{k+1} \right\rangle \\
&\overset{(4.138)}{\leq} \psi(y^k) - \psi(\tilde{y}^{k+1}) + \delta \|y^k - \tilde{y}^{k+1}\|_2 \\
&\quad + \left\langle \nabla f(\tilde{y}^{k+1}, \tilde{z}^{k+1}) - \mathbf{E}_k \left[\nabla f(\tilde{y}^{k+1}, \tilde{z}^{k+1}) \right], y^k - \tilde{y}^{k+1} \right\rangle .
\end{aligned} \tag{4.143}$$

By definition of $\tilde{y}^{k+1}$ we have

$$\alpha_{k+1} \left(\tilde{y}^{k+1} - z^k \right) = A_k \left(y^k - \tilde{y}^{k+1} \right). \tag{4.144}$$

Putting all together, we get

$$\begin{aligned}
&\alpha_{k+1} \langle \nabla f(\tilde{y}^{k+1}, \tilde{z}^{k+1}), \tilde{y}^{k+1} - z \rangle \\
&\quad = \alpha_{k+1} \langle \nabla f(\tilde{y}^{k+1}, \tilde{z}^{k+1}), \tilde{y}^{k+1} - z^k \rangle + \alpha_{k+1} \langle \nabla f(\tilde{y}^{k+1}, \tilde{z}^{k+1}), z^k - z \rangle \\
&\quad \overset{(4.144)}{=} A_k \langle \nabla f(\tilde{y}^{k+1}, \tilde{z}^{k+1}), y^k - \tilde{y}^{k+1} \rangle + \alpha_{k+1} \langle \nabla f(\tilde{y}^{k+1}, \tilde{z}^{k+1}), z^k - z \rangle \\
&\quad \overset{(4.142);(4.143)}{\leq} A_k \left(\psi(y^k) - \psi(\tilde{y}^{k+1}) + \delta \|y^k - \tilde{y}^{k+1}\|_2 \right) \\
&\qquad + A_k \left\langle \nabla f(\tilde{y}^{k+1}, \tilde{z}^{k+1}) - \mathbf{E}_k \left[\nabla f(\tilde{y}^{k+1}, \tilde{z}^{k+1}) \right], y^k - \tilde{y}^{k+1} \right\rangle \\
&\qquad + A_{k+1} \left(\psi(\tilde{y}^{k+1}) - \psi(y^{k+1}) + \tfrac{\varepsilon}{2} \right) + \tfrac{1}{2} L \|z^k - z\|_2^2 - \tfrac{1}{2} L \|z^{k+1} - z\|_2^2 \\
&\qquad + \tfrac{A_{k+1}}{2\varepsilon} \left\| \nabla f(\tilde{y}^{k+1}, \tilde{z}^{k+1}) - \mathbf{E}_k \left[\nabla f(\tilde{y}^{k+1}, \tilde{z}^{k+1}) \right] \right\|_2^2 .
\end{aligned}$$

Rearranging the terms and using $A_{k+1} = A_k + \alpha_{k+1}$, we obtain

$$\begin{aligned}
A_{k+1} \psi(y^{k+1}) - A_k \psi(y^k) &\leq \alpha_{k+1} \left(\psi(\tilde{y}^{k+1}) + \langle \nabla f(\tilde{y}^{k+1}, \tilde{z}^{k+1}), z - \tilde{y}^{k+1} \rangle \right) + \frac{1}{2} L \|z^k - z\|_2^2 \\
&\quad - \frac{1}{2} L \|z^{k+1} - z\|_2^2 + A_k \delta \|y^k - \tilde{y}^{k+1}\|_2 + \frac{A_{k+1} \delta^2}{L} \\
&\quad + \frac{A_{k+1}}{2L} \left\| \nabla f(\tilde{y}^{k+1}, \tilde{z}^{k+1}) - \mathbf{E}_k \left[\nabla f(\tilde{y}^{k+1}, \tilde{z}^{k+1}) \right] \right\|_2^2 \\
&\quad + A_k \left\langle \nabla f(\tilde{y}^{k+1}, \tilde{z}^{k+1}) - \mathbf{E}_k \left[\nabla f(\tilde{y}^{k+1}, \tilde{z}^{k+1}) \right], y^k - \tilde{y}^{k+1} \right\rangle ,
\end{aligned}$$

and after summing these inequalities for $k = 0, \ldots, N-1$ we get

$$
\begin{aligned}
A_N \psi(y^N) \;\le\; & \frac{1}{2}\|z - z^0\|_2^2 - \frac{1}{2}\|z - z^N\|_2^2 + \sum_{k=0}^{N-1} \alpha_{k+1} \left(\psi(y^{k+1}) + \langle \nabla f(y^{k+1}, \zeta^{k+1}), z - y^{k+1} \rangle \right) \\
& + \sum_{k=0}^{N-1} A_k \left\langle \nabla f(y^{k+1}, \zeta^{k+1}) - \mathbf{E}_k \left[\nabla f(y^{k+1}, \zeta^{k+1}) \right], y^k - y^{k+1} \right\rangle \\
& + \sum_{k=0}^{N-1} \frac{A_{k+1}}{2L} \left\| \mathbf{E}_k \left[\nabla f(y^{k+1}, \zeta^{k+1}) \right] - \nabla f(y^{k+1}, \zeta^{k+1}) \right\|_2^2 \\
& + \delta \sum_{k=0}^{N-1} A_k \|y^k - y^{k+1}\|_2 + \delta^2 \sum_{k=0}^{N-1} \frac{A_{k+1}}{L},
\end{aligned}
$$

where we use that $A_0 = 0$. □

The following lemma plays the central role in our analysis and it serves as the key to prove that the iterates of SPDSTM lie in the ball of radius R_y up to some polylogarithmic factor of N.

Lemma 4.9.8 (see also Lemma 7 from [46]). *Let the sequences of non-negative numbers $\{\alpha_k\}_{k\ge 0}$, random non-negative variables $\{R_k\}_{k\ge 0}$ and random vectors $\{\eta^k\}_{k\ge 0}$, $\{a^k\}_{k\ge 0}$ satisfy inequality*

$$
\frac{1}{2} R_l^2 \le A + h\delta \sum_{k=0}^{l-1} \alpha_{k+1} \widetilde{R}_k + u \sum_{k=0}^{l-1} \alpha_{k+1} \langle \eta^k, a^k \rangle + c \sum_{k=0}^{l-1} \alpha_{k+1}^2 \|\eta^k\|_2^2, \tag{4.145}
$$

for all $l = 1, \ldots, N$, where h, δ, u and c are some non-negative constants. Assume that for each $k \ge 1$ vector a^k is a function of $\eta^0, \ldots, \eta^{k-1}$, a^0 is a deterministic vector, $u \ge 1$, sequence of random vectors $\{\eta^k\}_{k\ge 0}$ satisfy $\forall k \ge 0$

$$
\mathbf{E}\left[\eta^k \mid \eta^0, \ldots, \eta^{k-1}\right] = 0, \quad \mathbf{E}\left[\exp\left(\frac{\|\eta^k\|_2^2}{\sigma_k^2}\right) \mid \eta^0, \ldots, \eta^{k-1}\right] \le \exp(1), \tag{4.146}
$$

$\alpha_{k+1} \le \widetilde{\alpha}_{k+1} = D(k+2)$, $\sigma_k^2 \le \frac{C''}{\widetilde{\alpha}_{k+1} \ln\left(\frac{N}{\beta}\right)}$ for some $D, C > 0$, $\varepsilon > 0$, $\beta \in (0,1)$ and sequence of random variables $\{\widetilde{R}_k\}_{k\ge 0}$ is such that $\|a^k\|_2 \le d\widetilde{R}_k$ with some positive deterministic constant $d \ge 1$ and $\widetilde{R}_k = \max\{\widetilde{R}_{k-1}, R_k\}$ for all $k \ge 1$, $\widetilde{R}_0 = R_0$, $\widetilde{R}_k$ depends only on $\eta_0, \ldots, \eta^k$ and also assume that $\ln\left(\frac{N}{\beta}\right) \ge 3$. If additionally $\varepsilon \le \frac{HR_0^2}{N^2}$ and $\delta \le \frac{GR_0}{(N+1)^2}$, then with probability at least $1 - 2\beta$ the inequalities

$$
\widetilde{R}_l \le J R_0 \tag{4.147}
$$

and

$$
\begin{aligned}
u \sum_{k=0}^{l-1} \alpha_{k+1} \langle \eta^k, a^k \rangle + c \sum_{k=0}^{l-1} \alpha_{k+1}^2 \|\eta^k\|_2^2 \;\le\; & \Big(24cCDH + hGDJ \\
& + udC_1 \sqrt{CDHJ \ln(N)} \Big) R_0^2
\end{aligned} \tag{4.148}
$$

hold for all $l = 1, \dots, N$ *simultaneously, where* C_1 *is some positive constant,* $g(N) = \frac{\ln\left(\frac{N}{\beta}\right) + \ln\ln\left(\frac{B}{b}\right)}{\ln\left(\frac{N}{\beta}\right)}$,

$$B = 2d^2CDHR_0^2\left(2A + (1+ud)R_0^2 + 48CDHR_0^2(2c+ud) + h^2G^2R_0^2D\right)(2(1+ud))^N,$$

$b = \sigma_0^2\widetilde{\alpha}_1^2 d^2 \widetilde{R}_0^2$ *and*

$$J = \max\left\{1, udC_1\sqrt{CDHg(N)} + hGD + \sqrt{\left(udC_1\sqrt{CDHg(N)} + hGD\right)^2 + \frac{2A}{R_0^2} + 48cCDH}\right\}.$$

Proof. We start with applying Cauchy-Schwarz inequality to the second and the third terms in the right-hand side of (4.145):

$$\begin{aligned}\frac{1}{2}R_l^2 &\le A + h\delta\sum_{k=0}^{l-1}\alpha_{k+1}\widetilde{R}_k + ud\sum_{k=0}^{l-1}\alpha_{k+1}\|\eta^k\|_2\widetilde{R}_k + c\sum_{k=0}^{l-1}\alpha_{k+1}^2\|\eta^k\|_2^2,\\ &\le A + \frac{h^2\delta^2}{2}\sum_{k=0}^{l-1}\alpha_{k+1}^2 + \frac{ud+1}{2}\sum_{k=0}^{l-1}\widetilde{R}_k^2 + \left(c + \frac{ud}{2}\right)\sum_{k=0}^{l-1}\widetilde{\alpha}_{k+1}^2\|\eta^k\|_2^2. \quad (4.149)\end{aligned}$$

The idea of the proof is as following: estimate R_N^2 roughly, then apply Lemma 4.9.4 in order to estimate second term in the last row of (4.145) and after that use the obtained recurrence to estimate right-hand side of (4.145).

Using Lemma 4.9.5 we get that with probability at least $1 - \frac{\beta}{N}$

$$\begin{aligned}\|\eta^k\|_2 &\le \frac{\beta}{2}\left(1 + \sqrt{3\ln\frac{N}{\beta}}\right)\sigma_k \le \frac{\beta}{2}\left(1 + \sqrt{3\ln\frac{N}{\beta}}\right)\frac{\sqrt{C\varepsilon}}{\sqrt{\widetilde{\alpha}_{k+1}\ln\left(\frac{N}{\beta}\right)}}\\ &= \left(\frac{1}{\sqrt{\widetilde{\alpha}_{k+1}\ln\left(\frac{N}{\beta}\right)}} + \sqrt{\frac{3}{\widetilde{\alpha}_{k+1}}}\right)\sqrt{2C\varepsilon} \le 2\sqrt{\frac{3}{\widetilde{\alpha}_{k+1}}}\sqrt{2C\varepsilon}, \quad (4.150)\end{aligned}$$

where in the last inequality we use $\ln\frac{N}{\beta} \ge 3$. Using union bound and $\alpha_{k+1} \le \widetilde{\alpha}_{k+1} = D(k+2)$ we get that with probability $\ge 1 - \beta$ the inequality

$$\begin{aligned}\frac{1}{2}R_l^2 &\le A + \frac{h^2\delta^2D^2}{2}\sum_{k=0}^{l-1}(k+2)^2 + \frac{ud+1}{2}\sum_{k=0}^{l-1}\widetilde{R}_k^2 + 24C\varepsilon\left(c + \frac{ud}{2}\right)\sum_{k=0}^{l-1}\widetilde{\alpha}_{k+1}\\ &\le A + \frac{h^2\delta^2D^2}{2}l(l+1)^2 + \frac{ud+1}{2}\sum_{k=0}^{l-1}\widetilde{R}_k^2 + 24CD\varepsilon\left(c + \frac{ud}{2}\right)\sum_{k=0}^{l-1}(k+2)\\ &\le A + \frac{h^2\delta^2D^2}{2}l(l+1)^2 + \frac{ud+1}{2}\sum_{k=0}^{l-1}\widetilde{R}_k^2 + 12CD\varepsilon\left(c + \frac{ud}{2}\right)l(l+3)\end{aligned}$$

holds for all $l = 1, \ldots, N$ simultaneously. Note that the last row in the previous inequality is non-decreasing function of l. If we define $\hat{l}$ as the largest integer such that $\hat{l} \le l$ and $\widetilde{R}_{\hat{l}} = R_{\hat{l}}$, we will get that $R_{\hat{l}} = \widetilde{R}_{\hat{l}} = \widetilde{R}_{\hat{l}+1} = \ldots = \widetilde{R}_l$ and, as a consequence, with probability $\ge 1 - \beta$

$$\begin{aligned}\frac{1}{2}\widetilde{R}_l^2 &\le A + \frac{h^2\delta^2 D^2}{2}\hat{l}(\hat{l}+1)^2 + \frac{ud+1}{2}\sum_{k=0}^{\hat{l}-1}\widetilde{R}_k^2 + 12CD\varepsilon\left(c + \frac{ud}{2}\right)\hat{l}(\hat{l}+3)\\ &\le A + \frac{h^2\delta^2 D^2}{2}l(l+1)^2 + \frac{ud+1}{2}\sum_{k=0}^{l-1}\widetilde{R}_k^2 + 12CD\varepsilon\left(c + \frac{ud}{2}\right)l(l+3), \quad \forall l = 1, \ldots, N.\end{aligned}$$

Therefore, we have that with probability $\ge 1 - \beta$

$$\begin{aligned}\widetilde{R}_l^2 &\le 2A + (ud+1)\sum_{k=0}^{l-1}\widetilde{R}_k^2 + 12CD\varepsilon\,(2c+ud)\,l(l+3) + h^2\delta^2 D^2 l(l+1)^2\\ &\le 2A\underbrace{(2+ud)}_{\le 2(1+ud)} + \underbrace{(1+ud+(1+ud)^2)}_{\le 2(1+ud)^2}\sum_{k=0}^{l-2}\widetilde{R}_k^2\\ &\quad + 12CD\varepsilon(2c+ud)\underbrace{(l(l+3)+(1+ud)(l-1)(l+2))}_{\le 2(1+ud)l(l+3)}\\ &\quad + h^2\delta^2 D^2\underbrace{(l(l+1)^2+(1+ud)(l-1)l^2)}_{\le 2(1+ud)l(l+1)^2}\\ &\le 2(1+ud)\left(2A + (1+ud)\sum_{k=0}^{l-2}\widetilde{R}_k^2 + 12CD\varepsilon\,(2c+ud)\,l(l+3) + h^2\delta^2 D^2 l(l+1)^2\right),\end{aligned}$$

for all $l = 1, \ldots, N$. Unrolling the recurrence we get that with probability $\ge 1 - \beta$

$$\widetilde{R}_l^2 \le \left(2A + (1+ud)\widetilde{R}_0^2 + 12CD\varepsilon\,(2c+ud)\,l(l+3) + h^2\delta^2 D^2 l(l+1)^2\right)(2(1+ud))^l,$$

for all $l = 1, \ldots, N$. We emphasize that it is very rough estimate, but we show next that such a bound does not spoil the final result too much. It implies that with probability $\ge 1 - \beta$

$$\sum_{k=0}^{l-1}\widetilde{R}_k^2 \le l\left(2A + (1+ud)\widetilde{R}_0^2 + 12CD\varepsilon\,(2c+ud)\,l(l+3) + h^2\delta^2 D^2 l(l+1)^2\right)(2(1+ud))^l, \tag{4.151}$$

for all $l = 1, \ldots, N$. Next we apply delicate result from [92] which is presented in Section 4.9.3 as Lemma 4.9.4. We consider random variables $\xi^k = \widetilde{\alpha}_{k+1}\langle\eta^k, a^k\rangle$. Note that $\mathbf{E}\left[\xi^k \mid \xi^0, \ldots, \xi^{k-1}\right] = \widetilde{\alpha}_{k+1}\left\langle\mathbf{E}\left[\eta^k \mid \eta^0, \ldots, \eta^{k-1}\right], a^k\right\rangle = 0$ and

$$\begin{aligned}\mathbf{E}\left[\exp\left(\frac{(\xi^k)^2}{\sigma_k^2\widetilde{\alpha}_{k+1}^2 d^2\widetilde{R}_k^2}\right)\mid \xi^0, \ldots, \xi^{k-1}\right] &\le \mathbf{E}\left[\exp\left(\frac{\widetilde{\alpha}_{k+1}^2\|\eta^k\|_2^2 d^2\widetilde{R}_k^2}{\sigma_k^2\widetilde{\alpha}_{k+1}^2 d^2\widetilde{R}_k^2}\right)\mid \eta^0, \ldots, \eta^{k-1}\right]\\ &= \mathbf{E}\left[\exp\left(\frac{\|\eta^k\|_2^2}{\sigma_k^2}\right)\mid \eta^0, \ldots, \eta^{k-1}\right] \le \exp(1)\end{aligned}$$

due to Cauchy-Schwarz inequality and assumptions of the lemma. If we denote $\theta_k^2 = \sigma_k^2 \widetilde{\alpha}_{k+1}^2 d^2 \widetilde{R}_k^2$ and apply Lemma 4.9.4 with

$$B = 2d^2CDHR_0^2\left(2A + (1+ud)R_0^2 + 48CDHR_0^2(2c+ud) + h^2G^2R_0^2D^2\right)(2(1+ud))^N$$

and $b = \theta_0^2$, we get that for all $l = 1, \ldots, N$ with probability $\geq 1 - \frac{}{N}$

$$\text{either} \sum_{k=0}^{l-1} \theta_k^2 \geq B \text{ or } \left|\sum_{k=0}^{l-1} \xi^k\right| \leq C_1 \sqrt{\sum_{k=0}^{l-1} \theta_k^2 \left(\ln\left(\frac{N}{\beta}\right) + \ln\ln\left(\frac{B}{b}\right)\right)}$$

with some constant $C_1 > 0$ which does not depend on B or b. Using union bound we obtain that with probability $\geq 1 - \beta$

$$\text{either} \sum_{k=0}^{l-1} \theta_k^2 \geq B \text{ or } \left|\sum_{k=0}^{l-1} \xi^k\right| \leq C_1 \sqrt{\sum_{k=0}^{l-1} \theta_k^2 \left(\ln\left(\frac{N}{\beta}\right) + \ln\ln\left(\frac{B}{b}\right)\right)}$$

and it holds for all $l = 1, \ldots, N$ simultaneously. Note that with probability at least $1 - \beta$

$$\begin{aligned}
\sum_{k=0}^{l-1} \theta_k^2 &= d^2 \sum_{k=0}^{l-1} \sigma_k^2 \widetilde{\alpha}_{k+1}^2 \widetilde{R}_k^2 \leq d^2 \sum_{k=0}^{l-1} \frac{C\varepsilon}{\ln \frac{N}{}} \widetilde{\alpha}_{k+1} \widetilde{R}_k^2 \\
&\leq \frac{d^2CDHR_0^2}{N^2 \ln \frac{N}{}} \sum_{k=0}^{l-1} (k+2)\widetilde{R}_k^2 \leq \frac{d^2CDHR_0^2}{3N} \cdot \frac{N+1}{N} \sum_{k=0}^{l-1} \widetilde{R}_k^2 \\
&\overset{(4.151)}{\leq} \frac{d^2CDHR_0^2}{N} l(2(1+ud))^l \Bigg(2A + (1+ud)\widetilde{R}_0^2 + 12CD\varepsilon(2c+ud)\, l(l+3) \\
&\qquad\qquad + h^2\delta^2 D^2 l(l+1)^2 \Bigg) \\
&\leq d^2CDHR_0^2\left(2A + (1+ud)R_0^2 + 48CDHR_0^2(2c+ud) + h^2G^2R_0^2D^2\right)(2(1+ud))^N \\
&= \frac{B}{2}
\end{aligned}$$

for all $l = 1, \ldots, N$ simultaneously. Using union bound again we get that with probability $\geq 1 - 2\beta$ the inequality

$$\left|\sum_{k=0}^{l-1} \xi^k\right| \leq C_1 \sqrt{\sum_{k=0}^{l-1} \theta_k^2 \left(\ln\left(\frac{N}{\beta}\right) + \ln\ln\left(\frac{B}{b}\right)\right)} \tag{4.152}$$

holds for all $l = 1, \ldots, N$ simultaneously.

Note that we also proved that (4.150) is in the same event together with (4.152) and holds with probability $\geq 1 - 2\beta$. Putting all together in (4.145), we get that with

probability at least $1 - 2\beta$ the inequality

$$\begin{aligned}\frac{1}{2}\widetilde{R}_l^2 &\overset{(4.145)}{\leq} A + h\delta\sum_{k=0}^{l-1}\alpha_{k+1}\widetilde{R}_k + u\sum_{k=0}^{l-1}\alpha_{k+1}\langle h\eta^k, a^k i\rangle + c\sum_{k=0}^{l-1}\alpha_{k+1}^2 k\eta^k k_2^2 \\ &\overset{(4.152)}{\leq} A + h\delta\sum_{k=0}^{l-1}\alpha_{k+1}\widetilde{R}_k + uC_1\sqrt{\sum_{k=0}^{l-1}\theta_k^2\left(\ln\left(\frac{N}{\beta}\right) + \ln\ln\left(\frac{B}{b}\right)\right)} + 24cC\varepsilon\sum_{k=0}^{l-1}\widetilde{\alpha}_{k+1}\end{aligned}$$

holds for all $l = 1, \ldots, N$ simultaneously. For brevity, we introduce new notation: $g(N) = \frac{\ln\left(\frac{N}{\beta}\right) + \ln\ln\left(\frac{B}{b}\right)}{\ln\left(\frac{N}{\beta}\right)} \asymp 1$ (neglecting constant factor). Using our assumption $\sigma_k^2 \leq \frac{C''}{\widetilde{\alpha}_{k+1}\ln\left(\frac{N}{\beta}\right)}$ and definition $\theta_k^2 = \sigma_k^2\widetilde{\alpha}_{k+1}^2 d^2\widetilde{R}_k^2$ we obtain that with probability at least $1 - 2\beta$ the inequality

$$\begin{aligned}\frac{1}{2}\widetilde{R}_l^2 &\leq A + h\delta\sum_{k=0}^{l-1}\alpha_{k+1}\widetilde{R}_k + u\sum_{k=0}^{l-1}\alpha_{k+1}\langle h\eta^k, a^k i\rangle + c\sum_{k=0}^{l-1}\alpha_{k+1}^2 k\eta^k k_2^2 \\ &\leq A + \frac{hGDR_0}{(N+1)^2}\sum_{k=0}^{l-1}(k+2)\widetilde{R}_k + udC_1\sqrt{C\varepsilon g(N)}\sqrt{\sum_{k=0}^{l-1}\widetilde{\alpha}_{k+1}\widetilde{R}_k^2} + 24cC\varepsilon\sum_{k=0}^{l-1}\widetilde{\alpha}_{k+1} \\ &\leq A + \frac{hGDR_0}{(N+1)^2}\sum_{k=0}^{l-1}(k+2)\widetilde{R}_k + udC_1\sqrt{CD\varepsilon g(N)}\sqrt{\sum_{k=0}^{l-1}(k+2)\widetilde{R}_k^2} \\ &\quad + 24cCD\varepsilon\sum_{k=0}^{l-1}(k+2) \\ &\leq A + 24cCD\frac{HR_0^2}{N^2}\frac{l(l+1)}{2} + \frac{hGDR_0}{(N+1)^2}\sum_{k=0}^{l-1}(k+2)\widetilde{R}_k \\ &\quad + udC_1\sqrt{CD\frac{HR_0^2}{N^2}g(N)}\sqrt{\sum_{k=0}^{l-1}(k+2)\widetilde{R}_k^2} \\ &\leq \left(\frac{A}{R_0^2} + 24cCDH\right)R_0^2 + \frac{hGDR_0}{(N+1)^2}\sum_{k=0}^{l-1}(k+2)\widetilde{R}_k \\ &\quad + \frac{udC_1R_0}{N}\sqrt{CDHg(N)}\sqrt{\sum_{k=0}^{l-1}(k+2)\widetilde{R}_k^2} \qquad (4.153)\end{aligned}$$

holds for all $l = 1, \ldots, N$ simultaneously. Next we apply Lemma 4.9.11 with $A = \frac{A}{R_0^2} + 24cCDH$, $B = udC_1\sqrt{CDHg(N)}$, $D = hGD$, $r_k = \widetilde{R}_k$ and get that with probability at least $1 - 2\beta$ inequality

$$\widetilde{R}_l \leq JR_0$$

holds for all $l = 1, \ldots, N$ simultaneously with

$$J = \max\left\{1, udC_1\sqrt{CDHg(N)} + hGD + \sqrt{\left(udC_1\sqrt{CDHg(N)} + hGD\right)^2 + \frac{2A}{R_0^2} + 48cCDH}\right\}.$$

It implies that with probability at least $1 - 2\beta$ the inequality

$$\begin{aligned} A + h\delta\sum_{k=0}^{l-1}\alpha_{k+1}\widetilde{R}_k + u\sum_{k=0}^{l-1}\alpha_{k+1}\langle\eta^k, a^k\rangle + c\sum_{k=0}^{l-1}\alpha_{k+1}^2\|\eta^k\|_2^2 \\ \leq \left(\frac{A}{R_0^2} + 24cCDH\right)R_0^2 + \frac{hGDJR_0^2}{(N+1)^2}\sum_{k=0}^{l-1}(k+2) + \frac{udC_1R_0^2}{N}\sqrt{CDHg(N)}\sqrt{\sum_{k=0}^{l-1}(k+2)J} \\ \leq A + \left(24cCDH + hGDJ + udC_1\sqrt{CDHJg(N)}\frac{1}{N}\sqrt{\frac{l(l+1)}{2}}\right)R_0^2 \\ \leq A + \left(24cCDH + hGDJ + udC_1\sqrt{CDHJg(N)}\right)R_0^2 \end{aligned}$$

holds for all $l = 1, \ldots, N$ simultaneously. □

Proof of Theorem 4.5.1

For the convenience we put here the extended statement of the theorem.

Theorem 4.9.1. *Assume that f is μ-strongly convex and $\|\nabla f(x^*)\|_2 = M_f$. Let $\varepsilon > 0$ be a desired accuracy. Next, assume that f is L_f-Lipschitz continuous on the ball $B_{R_f}(0)$ with*

$$R_f = \widetilde{\ } \left(\max\left\{\frac{R_y}{A_N\sqrt{\lambda_{\max}(A^\top A)}}, \frac{\sqrt{\lambda_{\max}(A^\top A)}R_y}{\mu}, R_x\right\}\right),$$

where R_y is such that $\|y^*\|_2 \leq R_y$, y^* is the solution of the dual problem (4.22), and $R_x = \|x(A^\top y^*)\|_2$. Assume that at iteration k of Algorithm 3 batch size is chosen according to the formula $r_k \geq \max\left\{1, \frac{\widetilde{\psi}_k \ln(N/\beta)}{\hat{C}}\right\}$, where $\widetilde{\alpha}_k = \frac{k+1}{2L}$, $0 < \varepsilon \leq \frac{HLR_0^2}{N^2}$, $0 \leq \delta \leq \frac{GLR_0}{(N+1)^2}$ and $N \geq 1$ for some numeric constant $H > 0$, $G > 0$ and $\hat{C} > 0$. Then with probability $\geq 1 - 4\beta$

$$\begin{aligned} \psi(y^N) + f(x^N) + 2R_y\|Ax^N\|_2 \leq \frac{R_y^2}{A_N}\Bigg(8\sqrt{HC_2} + 2 + 12CH + \frac{G(6J+4)}{2} \\ + \frac{L_f\left(\sqrt{96C_2H} + G\right)}{2R_y\sqrt{\lambda_{\max}(A^\top A)}} + \frac{G^2}{2(N+1)} \\ + C_1\sqrt{\frac{CHJg(N)}{2}} + \sqrt{96C_2H} + G\Bigg), \end{aligned} \quad (4.154)$$

where $\beta \in (0, 1/4)$ *is such that* $\frac{1+\sqrt{\ln\frac{1}{\beta}}}{\sqrt{\ln\frac{N}{\beta}}} \le 2$, C_2, C, C_1 *are some positive numeric constants,* $g(N) = \frac{\ln\left(\frac{N}{\beta}\right)+\ln\ln\left(\frac{B}{b}\right)}{\ln\left(\frac{N}{\beta}\right)}$,

$$B = CHR_0^2\left(2A + 2R_0^2 + 72CHR_0^2 + \frac{9G^2\tilde{L}R_0^2}{2}\right)4^N,$$

$b = \sigma_0^2\widetilde{\alpha}_1^2R_0^2$ *and*

$$J = \max\left\{1, C_1\sqrt{\frac{CHg(N)}{2}} + \frac{3G}{2} + \sqrt{\left(C_1\sqrt{\frac{CHg(N)}{2}} + \frac{3G}{2}\right)^2 + \frac{2A}{R_0^2} + 24CH}\right\}.$$

This means that after $N = \widetilde{O}\left(\sqrt{\frac{M_f}{\mu\varepsilon}\chi(A^\top A)}\right)$ *iterations where* $\chi(A^\top A) = \frac{\max(A^\top A)}{\min_+(A^\top A)}$, *the outputs* x^N *and* y^N *of Algorithm 3 satisfy the following condition*

$$f(x^N) - f(x^*) \le f(x^N) + \psi(y^N) \le \varepsilon, \quad \|Ax^N\|_2 \le \frac{\varepsilon}{R_y} \tag{4.155}$$

with probability at least $1 - 4\beta$. *What is more, to guarantee* (4.155) *with probability at least* $1 - 4\beta$ *Algorithm 3 requires*

$$\widetilde{O}\left(\max\left\{\frac{\sigma_x^2M_f^2}{\varepsilon^2}\chi(A^\top A)\ln\left(\frac{1}{\beta}\sqrt{\frac{M_f}{\mu\varepsilon}\chi(A^\top A)}\right), \sqrt{\frac{M_f}{\mu\varepsilon}\chi(A^\top A)}\right\}\right) \tag{4.156}$$

calls of the biased stochastic oracle $\nabla\psi(y, \xi)$, *i.e.* $x(y, \xi)$.

Proof. Lemma 4.9.7 states that

$$\begin{aligned} A_N\psi(y^N) \le\ & \frac{1}{2}\|u - z^0\|_2^2 - \frac{1}{2}\|u - z^N\|_2^2 \\ & + \sum_{k=0}^{N-1}\alpha_{k+1}\left(\psi(u^{k+1}) + \langle\nabla\psi(u^{k+1}, \xi^{k+1}), u - u^{k+1}\rangle\right) \\ & + \sum_{k=0}^{N-1}A_k\left\langle\nabla\psi(u^{k+1}, \xi^{k+1}) - \mathbf{E}_k\left[\nabla\psi(u^{k+1}, \xi^{k+1})\right], y^k - u^{k+1}\right\rangle \\ & + \sum_{k=0}^{N-1}\frac{A_{k+1}}{2\tilde{L}}\left\|\mathbf{E}_k\left[\nabla\psi(u^{k+1}, \xi^{k+1})\right] - \nabla\psi(u^{k+1}, \xi^{k+1})\right\|_2^2 \\ & + \delta\sum_{k=0}^{N-1}A_k\|y^k - u^{k+1}\|_2 + \delta^2\sum_{k=0}^{N-1}\frac{A_{k+1}}{\tilde{L}}, \end{aligned} \tag{4.157}$$

for arbitrary u. By definition of u^{k+1} we have

$$\alpha_{k+1}\left(u^{k+1} - z^k\right) = A_k\left(y^k - u^{k+1}\right). \tag{4.158}$$

Using this, we add and subtract $\sum_{k=0}^{N-1}\alpha_{k+1}\left\langle \mathbf{E}_k\left[\nabla \quad (y^{k+1},\ \ ^{k+1})\right], y^* - y^{k+1}\right\rangle$ in (4.157), and obtain the following inequality by choosing $y = y^*$ — the minimizer of $\psi(y)$:

$$\begin{aligned} A_N\psi(y^N) \quad \le \quad & \frac{1}{2}\|y^* - z^0\|_2^2 - \frac{1}{2}\|y^* - z^N\|_2^2 \\ & + \sum_{k=0}^{N-1}\alpha_{k+1}\left(\psi(y^{k+1}) + \left\langle \mathbf{E}_k\left[\nabla \quad (y^{k+1},\ \ ^{k+1})\right], y^* - y^{k+1}\right\rangle\right) \\ & + \sum_{k=0}^{N-1}\alpha_{k+1}\left\langle \nabla \quad (y^{k+1},\ \ ^{k+1}) - \mathbf{E}_k\left[\nabla \quad (y^{k+1},\ \ ^{k+1})\right], \mathbf{a}^k\right\rangle \\ & + \sum_{k=0}^{N-1}\alpha_{k+1}^2\left\| \nabla \quad (y^{k+1},\ \ ^{k+1}) - \mathbf{E}_k\left[\nabla \quad (y^{k+1},\ \ ^{k+1})\right]\right\|_2^2 \\ & + \delta\sum_{k=0}^{N-1}\alpha_{k+1}\|y^{k+1} - z^k\|_2 + \delta^2\sum_{k=0}^{N-1}\frac{A_{k+1}}{L}, \end{aligned} \tag{4.159}$$

where $\mathbf{a}^k = y^* - z^k$. From (4.138) we have

$$\begin{aligned} & \sum_{k=0}^{N-1}\alpha_{k+1}\left(\psi(y^{k+1}) + \left\langle \mathbf{E}_k\left[\nabla \quad (y^{k+1},\ \ ^{k+1})\right], y^* - y^{k+1}\right\rangle\right) \\ & \overset{(4.138)}{\le} \sum_{k=0}^{N-1}\alpha_{k+1}\left(\psi(y^{k+1}) + \psi(y^*) - \psi(y^{k+1}) + \delta\|y^{k+1} - y^*\|_2\right) \\ & = \sum_{k=0}^{N-1}\alpha_{k+1}\left(\psi(y^*) + \delta\|y^{k+1} - y^*\|_2\right) \\ & = A_N\psi(y^*) + \delta\sum_{k=0}^{N-1}\alpha_{k+1}\|y^{k+1} - y^*\|_2 \\ & \le A_N\psi(y^N) + \delta\sum_{k=0}^{N-1}\alpha_{k+1}\|y^{k+1} - y^*\|_2 \end{aligned}$$

From this and (4.159) we get

$$\begin{aligned} \frac{1}{2}\|y^* - z^N\|_2^2 \quad \overset{(4.159)}{\le} \quad & \frac{1}{2}\|y^* - z^0\|_2^2 + \delta^2\sum_{k=0}^{N-1}\frac{A_{k+1}}{L} \\ & + \delta\sum_{k=0}^{N-1}\alpha_{k+1}\left(\|y^{k+1} - z^k\|_2 + \|y^{k+1} - y^*\|_2\right) \\ & + \sum_{k=0}^{N-1}\alpha_{k+1}\left\langle \nabla \quad (y^{k+1},\ \ ^{k+1}) - \mathbf{E}_k\left[\nabla \quad (y^{k+1},\ \ ^{k+1})\right], \mathbf{a}^k\right\rangle \\ & + \sum_{k=0}^{N-1}\alpha_{k+1}^2\left\| \nabla \quad (y^{k+1},\ \ ^{k+1}) - \mathbf{E}_k\left[\nabla \quad (y^{k+1},\ \ ^{k+1})\right]\right\|_2^2 \end{aligned} \tag{4.160}$$

Next, we introduce the sequences $\{R_k\}_{k\ge 0}$ and $\{\widetilde{R}_k\}_{k\ge 0}$ as

$$R_k = \|z_k - y^*\|_2 \quad \text{and} \quad \widetilde{R}_k = \max\left\{\widetilde{R}_{k-1}, R_k\right\}, \widetilde{R}_0 = R_0$$

Since in Algorithm 3 we choose $z^0 = 0$, then $R_0 = R_y$. One can obtain by induction that $\forall k \geq 0$ we have $w^{k+1}, y^k, z^k \in B_{\widetilde{R}_k}(w^*)$, where $B_{\widetilde{R}_k}(w^*)$ is Euclidean ball with radius $\widetilde{R}_k$ at centre w^*. Indeed, since from lines 2 and 5 of Algorithm 3 y_{k+1} is a convex combination of $z_{k+1} \in B_{R_{k+1}}(w^*) \quad B_{\widetilde{R}_{k+1}}(w^*)$ and $y^k \in B_{\widetilde{R}_k}(w^*) \quad B_{\widetilde{R}_{k+1}}(w^*)$, where we use the fact that a ball is a convex set, we get $y^{k+1} \in B_{\widetilde{R}_{k+1}}(w^*)$. Analogously, since from lines 2 and 3 of Algorithm 3 w^{k+1} is a convex combination of y^k and z^k we have $w^{k+1} \in B_{\widetilde{R}_k}(w^*)$. It implies that

$$\|w^{k+1} - z^k\|_2 + \|w^{k+1} - w^*\|_2 \leq 2\widetilde{R}_k + \widetilde{R}_k = 3\widetilde{R}_k.$$

Using new notation we can rewrite (4.160) as

$$\begin{aligned}\frac{1}{2}R_N^2 \quad \leq \quad & \frac{1}{2}R_0^2 + \delta^2 \sum_{k=0}^{N-1} \frac{A_{k+1}}{L} + 3\delta \sum_{k=0}^{N-1} \alpha_{k+1}\widetilde{R}_k \\ & + \sum_{k=0}^{N-1} \alpha_{k+1} \left\langle \nabla f(w^{k+1}, \xi^{k+1}) - \mathbf{E}_k\left[\nabla f(w^{k+1}, \xi^{k+1})\right], \mathbf{a}^k \right\rangle \\ & + \sum_{k=0}^{N-1} \alpha_{k+1}^2 \left\| \nabla f(w^{k+1}, \xi^{k+1}) - \mathbf{E}_k\left[\nabla f(w^{k+1}, \xi^{k+1})\right] \right\|_2^2, \end{aligned} \tag{4.161}$$

where $\|\mathbf{a}^k\|_2 = \|w^* - z^k\|_2 \leq \widetilde{R}_k$. Note that (4.161) holds for all $N \geq 1$.

Let us denote $\eta^k = \nabla f(w^{k+1}, \xi^{k+1}) - \mathbf{E}_k\left[\nabla f(w^{k+1}, \xi^{k+1})\right]$. Theorem 2.1 from [93] (see Lemma 4.9.5 in the Section 4.9.3) says that

$$\mathbf{P}\left\{\|\eta^k\|_2 \geq \left(\sqrt{2} + \sqrt{2\gamma}\right)\sqrt{\frac{\sigma^2}{r_{k+1}}} \mid \eta^0, \ldots, \eta^{k-1}\right\} \leq \exp\left(-\frac{\gamma^2}{3}\right).$$

Using this and Lemma 2 from [92] (see Lemma 4.9.3 in the Section 4.9.3) we get that

$$\mathbf{E}\left[\exp\left(\frac{\|\eta^k\|_2^2}{\sigma_k^2}\right) \mid \eta^0, \ldots, \eta^{k-1}\right] \leq \exp(1),$$

where $\sigma_k^2 \leq \frac{\widetilde{C}^2 \psi^2}{r_{k+1}} \leq \frac{C''}{\widetilde{\gamma}_{k+1}\ln(\frac{N}{\delta})}$, $\widetilde{C}$ and $C = \widetilde{C} \cdot \hat{C}$ are some positive constants. From (4.225) we have that $\alpha_{k+1} \leq \widetilde{\alpha}_{k+1} = \frac{k+2}{2L}$. Moreover, $\mathbf{a}^k$ depends only on $\eta^0, \ldots, \eta^{k-1}$. Putting all together in (4.161) and changing the indices we get that for all $l = 1, \ldots, N$

$$\frac{1}{2}R_l^2 \leq \frac{1}{2}R_0^2 + \delta^2 \sum_{k=0}^{N-1} \frac{A_{k+1}}{L} + 3\delta \sum_{k=0}^{l-1} \alpha_{k+1}\widetilde{R}_k + \sum_{k=0}^{l-1} \alpha_{k+1}\langle \eta^k, \mathbf{a}^k \rangle + \sum_{k=0}^{l-1} \alpha_{k+1}^2 \|\eta^k\|_2^2.$$

Next we apply Lemma 4.9.8 with the constants $A = \frac{1}{2}R_0^2 + \delta^2 \sum\limits_{k=0}^{N-1} \frac{A_{k+1}}{L}, h = 3, u = 1, c = 1, D = \frac{1}{2L}, d = 1, \varepsilon \leq \frac{HLR_0^2}{N^2}$ and $\delta \leq \frac{GLR_0}{(N+1)^3}$, and get that with probability at least $1 - 2\beta$ the inequalities

$$\widetilde{R}_l \leq JR_0 \tag{4.162}$$

and

$$\sum_{k=0}^{l-1} \alpha_{k+1} \hbar \eta^k, \mathbf{a}^k i + \sum_{k=0}^{l-1} \alpha_{k+1}^2 k \eta^k k_2^2 \le \left(12CH + \frac{3GJ}{2} + C_1 \sqrt{\frac{CHJg(N)}{2}} \right) R_0^2 \quad (4.163)$$

hold for all $l = 1, \dots, N$ simultaneously, where C_1 is some positive constant, $g(N) = \frac{\ln\left(\frac{N}{\beta}\right) + \ln\ln\left(\frac{B}{b}\right)}{\ln\left(\frac{N}{\beta}\right)}$, $B = CHR_0^2 \left(2A + 2R_0^2 + 72CHR_0^2 + \frac{9G^2 L R_0^2}{2} \right) 4^N$, $b = \sigma_0^2 \widetilde{\alpha}_1^2 R_0^2$ and

$$J = \max \left\{ 1, C_1 \sqrt{\frac{CHg(N)}{2}} + \frac{3G}{2} + \sqrt{\left(C_1 \sqrt{\frac{CHg(N)}{2}} + \frac{3G}{2} \right)^2 + \frac{2A}{R_0^2} + 24CH} \right\}.$$

To estimate the duality gap we need again refer to (4.157). Since y is chosen arbitrary we can take the minimum in y over the set $B_{2R_y}(0) = \{y : \|y\|_2 \le 2R_y\}$:

$$\begin{aligned}
A_N \psi(y^N) \le & \min_{y \in B_{2R_y}(0)} \left\{ \frac{1}{2}\|y - z^0\|_2^2 \right. \\
& \left. + \sum_{k=0}^{N-1} \alpha_{k+1} \left(\psi(y^{k+1}) + \left\langle \nabla \quad (y^{k+1}, \quad^{k+1}), y - y^{k+1} \right\rangle \right) \right\} \\
& + \sum_{k=0}^{N-1} A_k \left\langle \nabla \quad (y^{k+1}, \quad^{k+1}) - \mathbf{E}_k \left[\nabla \quad (y^{k+1}, \quad^{k+1}) \right], y^k - y^{k+1} \right\rangle \\
& + \sum_{k=0}^{N-1} \frac{A_{k+1}}{2L} \left\| \mathbf{E}_k \left[\nabla \quad (y^{k+1}, \quad^{k+1}) \right] - \nabla \quad (y^{k+1}, \quad^{k+1}) \right\|_2^2 \\
& + \delta \sum_{k=0}^{N-1} A_k \|y^k - y^{k+1}\|_2 + \delta^2 \sum_{k=0}^{N-1} \frac{A_{k+1}}{L} \\
\le & \; 2R_y^2 + \min_{y \in B_{2R_y}(0)} \sum_{k=0}^{N-1} \alpha_{k+1} \left(\psi(y^{k+1}) + \left\langle \nabla \quad (y^{k+1}, \quad^{k+1}), y - y^{k+1} \right\rangle \right) \\
& + \sum_{k=0}^{N-1} A_k \left\langle \nabla \quad (y^{k+1}, \quad^{k+1}) - \mathbf{E}_k \left[\nabla \quad (y^{k+1}, \quad^{k+1}) \right], y^k - y^{k+1} \right\rangle \\
& + \sum_{k=0}^{N-1} \frac{A_{k+1}}{2L} \left\| \mathbf{E}_k \left[\nabla \quad (y^{k+1}, \quad^{k+1}) \right] - \nabla \quad (y^{k+1}, \quad^{k+1}) \right\|_2^2 \\
& + \delta \sum_{k=0}^{N-1} A_k \|y^k - y^{k+1}\|_2 + \delta^2 \sum_{k=0}^{N-1} \frac{A_{k+1}}{L}, \qquad (4.164)
\end{aligned}$$

where we also used $\frac{1}{2}\|y - z^N\|_2^2 \ge 0$ and $z^0 = 0$. By adding and subtracting

$\sum_{k=0}^{N-1} \alpha_{k+1} \left\langle \mathbf{E}_k \left[\tilde{r} \quad (y^{k+1}, \quad ^{k+1}) \right], y - y^{k+1} \right\rangle$ under the minimum in (4.164) we obtain

$$\begin{aligned}
&\min_{y \in B_{2R_y}(0)} \sum_{k=0}^{N-1} \alpha_{k+1} \left(\psi(y^{k+1}) + \left\langle \tilde{r} \quad (y^{k+1}, \quad ^{k+1}), y - y^{k+1} \right\rangle \right) \\
&\qquad \leq \min_{y \in B_{2R_y}(0)} \sum_{k=0}^{N-1} \alpha_{k+1} \left(\psi(y^{k+1}) + \left\langle \mathbf{E}_k \left[\tilde{r} \quad (y^{k+1}, \quad ^{k+1}) \right], y - y^{k+1} \right\rangle \right) \\
&\qquad\quad + \max_{y \in B_{2R_y}(0)} \sum_{k=0}^{N-1} \alpha_{k+1} \left\langle \tilde{r} \quad (y^{k+1}, \quad ^{k+1}) - \mathbf{E}_k \left[\tilde{r} \quad (y^{k+1}, \quad ^{k+1}) \right], y \right\rangle \\
&\qquad\quad + \sum_{k=0}^{N-1} \alpha_{k+1} \left\langle \tilde{r} \quad (y^{k+1}, \quad ^{k+1}) - \mathbf{E}_k \left[\tilde{r} \quad (y^{k+1}, \quad ^{k+1}) \right], -y^{k+1} \right\rangle .
\end{aligned}$$

Since $-y^* \in B_{2R_y}(0)$ we can bound the last term in the previous inequality as follows

$$\begin{aligned}
&\sum_{k=0}^{N-1} \alpha_{k+1} \left\langle \tilde{r} \quad (y^{k+1}, \quad ^{k+1}) - \mathbf{E}_k \left[\tilde{r} \quad (y^{k+1}, \quad ^{k+1}) \right], -y^{k+1} \right\rangle \\
&\qquad = \sum_{k=0}^{N-1} \alpha_{k+1} \left\langle \tilde{r} \quad (y^{k+1}, \quad ^{k+1}) - \mathbf{E}_k \left[\tilde{r} \quad (y^{k+1}, \quad ^{k+1}) \right], y^* - y^{k+1} \right\rangle \\
&\qquad\quad + \sum_{k=0}^{N-1} \alpha_{k+1} \left\langle \tilde{r} \quad (y^{k+1}, \quad ^{k+1}) - \mathbf{E}_k \left[\tilde{r} \quad (y^{k+1}, \quad ^{k+1}) \right], -y^* \right\rangle \\
&\qquad \leq \sum_{k=0}^{N-1} \alpha_{k+1} \left\langle \tilde{r} \quad (y^{k+1}, \quad ^{k+1}) - \mathbf{E}_k \left[\tilde{r} \quad (y^{k+1}, \quad ^{k+1}) \right], y^* - y^{k+1} \right\rangle \\
&\qquad\quad + \max_{y \in B_{2R_y}(0)} \sum_{k=0}^{N-1} \alpha_{k+1} \left\langle \tilde{r} \quad (y^{k+1}, \quad ^{k+1}) - \mathbf{E}_k \left[\tilde{r} \quad (y^{k+1}, \quad ^{k+1}) \right], y \right\rangle .
\end{aligned}$$

Putting all together in (4.164) and using (4.158) and line 2 from Algorithm 3 we get

$$\begin{aligned}
A_N \psi(y^N) \quad \leq \quad & 2R_y^2 + \min_{y \in B_{2R_y}(0)} \sum_{k=0}^{N-1} \alpha_{k+1} \left(\psi(y^{k+1}) + \left\langle \mathbf{E}_k \left[\tilde{r} \quad (y^{k+1}, \quad ^{k+1}) \right], y - y^{k+1} \right\rangle \right) \\
& + 2 \max_{y \in B_{2R_y}(0)} \sum_{k=0}^{N-1} \alpha_{k+1} \left\langle \tilde{r} \quad (y^{k+1}, \quad ^{k+1}) - \mathbf{E}_k \left[\tilde{r} \quad (y^{k+1}, \quad ^{k+1}) \right], y \right\rangle \\
& + \sum_{k=0}^{N-1} \alpha_{k+1} \left\langle \tilde{r} \quad (y^{k+1}, \quad ^{k+1}) - \mathbf{E}_k \left[\tilde{r} \quad (y^{k+1}, \quad ^{k+1}) \right], \mathbf{a}^k \right\rangle \\
& + \sum_{k=0}^{N-1} \alpha_{k+1}^2 \left\| \tilde{r} \quad (y^{k+1}, \quad ^{k+1}) - \mathbf{E}_k \left[\tilde{r} \quad (y^{k+1}, \quad ^{k+1}) \right] \right\|_2^2 \\
& + \delta \sum_{k=0}^{N-1} \alpha_{k+1} \| y^{k+1} - z^k \|_2 + \delta^2 \sum_{k=0}^{N-1} \frac{A_{k+1}}{L}, \qquad (4.165)
\end{aligned}$$

where $\mathbf{a}^k = y^* - z^k$. From (4.162) and (4.163) we have that with probability at least $1 - 2\beta$

the following inequality holds:

$$A_N\psi(y^N) \leq \min_{y\in B_{2R_y}(0)} \sum_{k=0}^{N-1} \alpha_{k+1}\left(\psi(y^{k+1}) + \left\langle \mathbf{E}_k\left[\nabla f(y^{k+1}, \xi^{k+1})\right], y - y^{k+1}\right\rangle\right)$$
$$+2\max_{y\in B_{2R_y}(0)} \sum_{k=0}^{N-1} \alpha_{k+1}\left\langle \nabla f(y^{k+1}, \xi^{k+1}) - \mathbf{E}_k\left[\nabla f(y^{k+1}, \xi^{k+1})\right], y\right\rangle$$
$$+2R_y^2 + \left(12CH + \frac{5GJ}{2} + \frac{G^2}{2(N+1)} + C_1\sqrt{\frac{CHJg(N)}{2}}\right) R_0^2, \quad (4.166)$$

where we used that $A_{k+1} \leq \frac{(k+2)^2}{2L}$ due to $\alpha_{k+1} \leq \frac{k+2}{2L}$ and

$$\delta\sum_{k=0}^{N-1}\alpha_{k+1}\|y^{k+1} - z^k\|_2 \leq 2\delta JR_0\sum_{k=0}^{N-1}\alpha_{k+1} \leq \frac{2GLR_0^2J}{(N+1)^2}\frac{1}{2L}\sum_{k=0}^{N-1}(k+2) \leq GJR_0^2,$$
$$\delta^2\sum_{k=0}^{N-1}\frac{A_{k+1}}{L} \leq \frac{G^2L^2R_0^2}{(N+1)^4}\sum_{k=0}^{N-1}\frac{(k+2)^2}{2L^2} \leq \frac{G^2R_0^2}{2(N+1)}$$

By the definition of the norm we get

$$\max_{y\in B_{2R_y}(0)} \sum_{k=0}^{N-1} \alpha_{k+1}\left\langle \nabla f(y^{k+1}, \xi^{k+1}) - \mathbf{E}_k\left[\nabla f(y^{k+1}, \xi^{k+1})\right], y\right\rangle$$
$$\leq 2R_y\left\|\sum_{k=0}^{N-1}\alpha_{k+1}\left(\nabla f(y^{k+1}, \xi^{k+1}) - \mathbf{E}_k\left[\nabla f(y^{k+1}, \xi^{k+1})\right]\right)\right\|_2 \quad (4.167)$$

Next we apply Lemma 4.9.5 to the right-hand side of the previous inequality and get

$$\mathbf{P}\left\{\left\|\sum_{k=0}^{N-1}\alpha_{k+1}\left(\nabla f(y^{k+1}, \xi^{k+1}) - \mathbf{E}_k\left[\nabla f(y^{k+1}, \xi^{k+1})\right]\right)\right\|_2 \right.$$
$$\left. \geq \left(\sqrt{2} + \sqrt{2}\gamma\right)\sqrt{\sum_{k=0}^{N-1}\alpha_{k+1}^2\frac{\sigma^2}{r_{k+1}}}\right\} \leq \exp\left(-\frac{\gamma^2}{3}\right).$$

Since $N^2 \leq \frac{HLR_0^2}{\sigma^2}$ and $r_k = \left(\max\left\{1, \frac{\sigma^2 k \ln(N^3)}{\sigma^2}\right\}\right)$ one can choose such $C_2 > 0$ that $\frac{\sigma^2}{r_k} \leq \frac{C_2\sigma^2}{k\ln\left(\frac{N}{\beta}\right)} \leq \frac{HLC_2R_0^2}{kN^2\ln\left(\frac{N}{\beta}\right)}$. Moreover, let us choose γ such that $\exp\left(-\frac{\gamma^2}{3}\right) = \beta \Rightarrow \gamma = \sqrt{3\ln\frac{1}{\beta}}$. From this we get that with probability at least $1-\beta$

$$\left\|\sum_{k=0}^{N-1}\alpha_{k+1}\left(\nabla f(y^{k+1}, \xi^{k+1}) - \mathbf{E}_k\left[\nabla f(y^{k+1}, \xi^{k+1})\right]\right)\right\|_2$$
$$\leq \sqrt{2}\left(1 + \sqrt{\ln\frac{1}{\beta}}\right)R_y\sqrt{\frac{HLC_2}{\ln\left(\frac{N}{\beta}\right)}}\sqrt{\sum_{k=0}^{N-1}\frac{k+1}{N^2}}$$
$$\overset{(4.225)}{\leq} 2\sqrt{2}R_y\sqrt{HLC_2}\sqrt{\sum_{k=0}^{N-1}\frac{k+2}{2LN^2}} = 2R_y\sqrt{2HC_2}\sqrt{\frac{N(N+3)}{N^2}} \leq 4R_y\sqrt{HC_2} \quad (4.168)$$

In the above inequality we used the fact that $R_y = R_0$. Putting all together and using union bound we get that with probability at least $1 - 3\beta$

$$
\begin{aligned}
A_N\psi(y^N) &\overset{(4.166)+(4.167)+(4.168)}{\leq} \min_{y \in B_{2R_y}(0)} \sum_{k=0}^{N-1} \alpha_{k+1} \left(\psi(y^{k+1}) + \left\langle \mathbf{E}_k \left[\nabla f(y^{k+1}, \xi^{k+1}) \right], y - y^{k+1} \right\rangle \right) \\
&\quad + \left(8\sqrt{HC_2} + 2 + 12CH + \frac{5GJ}{2} + \frac{G^2}{2(N+1)^3} \right. \\
&\quad \left. + C_1 \sqrt{\frac{CHJg(N)}{2}} \right) R_y^2 \\
&\leq \min_{y \in B_{2R_y}(0)} \sum_{k=0}^{N-1} \alpha_{k+1} \left(\psi(y^{k+1}) + \left\langle \nabla \psi(y^{k+1}), y - y^{k+1} \right\rangle \right) \\
&\quad + \max_{y \in B_{2R_y}(0)} \sum_{k=0}^{N-1} \alpha_{k+1} \left\langle \mathbf{E}_k \left[\nabla f(y^{k+1}, \xi^{k+1}) \right] - \nabla \psi(y^{k+1}), y - y^{k+1} \right\rangle \\
&\quad + \left(8\sqrt{HC_2} + 2 + 12CH + \frac{5GJ}{2} + \frac{G^2}{2(N+1)} + C_1 \sqrt{\frac{CHJg(N)}{2}} \right) R_y^2 \qquad (4.169)
\end{aligned}
$$

First of all, we notice that in the same probabilistic event we have $\|y^{k+1} - y^*\|_2 \leq \widetilde{R}_k \overset{(4.162)}{\leq} JR_0$. Therefore, in the same probabilistic event we get that $\|y^{k+1} - y\|_2 \leq \|y^{k+1} - y^*\|_2 + \|y^* - y\|_2 \leq (J+4)R_y$ for all $y \in B_{2R_y}(0)$, where we used $R_0 = R_y$. It implies that in the same probabilistic event we have

$$
\begin{aligned}
&\max_{y \in B_{2R_y}(0)} \sum_{k=0}^{N-1} \alpha_{k+1} \left\langle \mathbf{E}_k \left[\nabla f(y^{k+1}, \xi^{k+1}) \right] - \nabla \psi(y^{k+1}), y - y^{k+1} \right\rangle \\
&\qquad \leq \max_{y \in B_{2R_y}(0)} \sum_{k=0}^{N-1} \alpha_{k+1} \left\| \mathbf{E}_k \left[\nabla f(y^{k+1}, \xi^{k+1}) \right] - \nabla \psi(y^{k+1}) \right\|_2 \cdot \left\| y - y^{k+1} \right\|_2 \\
&\qquad \overset{(4.39)}{\leq} \sum_{k=0}^{N-1} \alpha_{k+1} \delta (J+4) R_y \leq \sum_{k=0}^{N-1} \frac{k+2}{2L} \frac{GLR_0}{(N+1)^2} (J+4) R_y \leq \frac{G(J+4)R_y^2}{2}.
\end{aligned}
$$

Secondly, using the same trick as in the proof of Theorem 1 from [94] we get that for arbitrary point y

$$
\psi(y) - \langle \nabla \psi(y), y \rangle \overset{(4.24)+(4.35)}{=} \langle y, Ax(A^\top y) \rangle - f\left(x(A^\top y)\right) - \langle Ax(A^\top y), y \rangle = -f(x(A^\top y)).
$$

Using these relations in (4.169) we obtain that with probability at least $1 - 3\beta$

$$
\begin{aligned}
A_N\psi(y^N) \quad &\leq \quad - \sum_{k=0}^{N-1} \alpha_{k+1} f(x(A^\top y^{k+1})) + \min_{y \in B_{2R_y}(0)} \sum_{k=0}^{N-1} \alpha_{k+1} \langle \nabla \psi(y^{k+1}), y \rangle \\
&\quad + \left(8\sqrt{HC_2} + 2 + 12CH + \frac{G(6J+4)}{2} + \frac{G^2}{2(N+1)} + \right. \\
&\quad \left. + C_1 \sqrt{\frac{CHJg(N)}{2}} \right) R_y^2. \qquad (4.170)
\end{aligned}
$$

To bound the first term in (4.170) we apply convexity of f and introduce the virtual primal iterate $\hat{x}^N = \frac{1}{A_N}\sum_{k=0}^{N-1}\alpha_{k+1}x(A^\top y^{k+1})$:

$$-\sum_{k=0}^{N-1}\alpha_{k+1}f(x(A^\top y^{k+1})) = -A_N\sum_{k=0}^{N-1}\frac{\alpha_{k+1}}{A_N}f(x(A^\top y^{k+1})) \le -A_N f(\hat{x}^N).$$

In order to bound the second term in the right-hand side of the previous inequality we use the definition of the norm we have

$$\begin{aligned}\min_{y\in 2B_{2R_y}(0)}\sum_{k=0}^{N-1}\alpha_{k+1}\langle r\psi(y^{k+1}), y\rangle &= \min_{y\in 2B_{2R_y}(0)}\left\langle\sum_{k=0}^{N-1}\alpha_{k+1}\, r\psi(y^{k+1}), y\right\rangle\\ &= -2R_y\left\|\sum_{k=0}^{N-1}\alpha_{k+1}\, r\psi(y^{k+1})\right\|_2\\ &= -2R_yA_N\|A\hat{x}^N\|_2,\end{aligned}$$

where we used equality (4.35). Putting all together we obtain that with probability at least $1-3\beta$

$$\begin{aligned}\psi(y^N) + f(\hat{x}^N) + 2R_y\|A\hat{x}^N\|_2 \le \frac{R_y^2}{A_N}\Bigg(&8\sqrt{HC_2} + 2 + 12CH + \frac{G(6J+4)}{2}\\ &+\frac{G^2}{2(N+1)} + C_1\sqrt{\frac{CHJg(N)}{2}}\Bigg)\end{aligned} \quad (4.171)$$

Lemma 4.9.5 implies that for all $\gamma > 0$

$$\mathbf{P}\left\{\left\|\sum_{k=0}^{N-1}\alpha_{k+1}\left(x(A^\top y^{k+1}, \xi^{k+1}) - \mathbf{E}\left[x(A^\top y^{k+1}, \xi^{k+1}) \mid y^{k+1}\right]\right)\right\|_2 \ge (\sqrt{2}+\sqrt{2}\gamma)\sqrt{\sum_{k=0}^{N-1}\frac{\alpha_{k+1}^2\sigma_x^2}{r_{k+1}}}\right\} \le \exp\left(-\frac{\gamma^2}{3}\right).$$

Using this inequality with $\gamma = \sqrt{3\ln\frac{1}{\beta}}$ and $r_k \ge \frac{\alpha_k^2\sigma_\psi^2\ln\frac{N}{\beta}}{C_2\varepsilon}$ we get that with probability at

least $1-\beta$

$$\begin{aligned}
\|x^N - \hat{x}^N\|_2 &= \frac{1}{A_N}\left\|\sum_{k=0}^{N-1}\alpha_{k+1}\left(x(A^\top y^{k+1},\ \xi^{k+1}) - x(A^\top y^{k+1})\right)\right\|_2 \\
&\le \frac{1}{A_N}\left\|\sum_{k=0}^{N-1}\alpha_{k+1}\left(x(A^\top y^{k+1},\ \xi^{k+1}) - \mathbf{E}\left[x(A^\top y^{k+1},\ \xi^{k+1}) \mid y^{k+1}\right]\right)\right\|_2 \\
&\quad + \frac{1}{A_N}\left\|\sum_{k=0}^{N-1}\alpha_{k+1}\left(\mathbf{E}\left[x(A^\top y^{k+1},\ \xi^{k+1}) \mid y^{k+1}\right] - x(A^\top y^{k+1})\right)\right\|_2 \\
&\le \frac{\sqrt{2}}{A_N}\left(1+\sqrt{3\ln\frac{1}{\beta}}\right)\sqrt{\sum_{k=0}^{N-1}\frac{\alpha_{k+1}^2\sigma_x^2}{r_{k+1}^2}} \\
&\quad + \frac{1}{A_N}\sum_{k=0}^{N-1}\alpha_{k+1}\left\|\mathbf{E}\left[x(A^\top y^{k+1},\ \xi^{k+1}) \mid y^{k+1}\right] - x(A^\top y^{k+1})\right\|_2 \\
&\overset{(4.37)}{\le} \frac{2}{A_N}\sqrt{6\ln\frac{1}{\beta}}\frac{1}{\sqrt{\ln N}}\sqrt{\sum_{k=0}^{N-1}\frac{C_2\alpha_{k+1}\varepsilon}{\lambda_{\max}(A^\top A)}} + \frac{1}{A_N}\sum_{k=0}^{N-1}\alpha_{k+1}\delta_y \\
&\le \frac{2}{A_N}\sqrt{\frac{6C_2}{\lambda_{\max}(A^\top A)}}\sqrt{\sum_{k=0}^{N-1}\frac{(k+2)H\tilde{L}R_y^2}{2\tilde{L}N^2}} \\
&\quad + \frac{1}{A_N}\sum_{k=0}^{N-1}\frac{k+2}{2\tilde{L}}\cdot\frac{G\tilde{L}R_y}{(N+1)^2\sqrt{\lambda_{\max}(A^\top A)}} \\
&\le \frac{2R_y}{A_N}\left(\sqrt{\frac{6C_2H}{\lambda_{\max}(A^\top A)}} + \frac{G}{4\sqrt{\lambda_{\max}(A^\top A)}}\right). \qquad (4.172)
\end{aligned}$$

It implies that with probability at least $1-\beta$

$$\begin{aligned}
\|Ax^N - A\hat{x}^N\|_2 &\le \|A\|_2\cdot\|x^N - \hat{x}^N\|_2 \\
&\overset{(4.172)}{\le} \sqrt{\lambda_{\max}(A^\top A)}\frac{2R_y}{A_N}\left(\sqrt{\frac{6C_2H}{\lambda_{\max}(A^\top A)}} + \frac{G}{4\sqrt{\lambda_{\max}(A^\top A)}}\right) \\
&= \frac{R_y}{2A_N}\left(\sqrt{96C_2H} + G\right) \qquad (4.173)
\end{aligned}$$

and due to triangle inequality with probability $\ge 1-\beta$

$$\begin{aligned}
2R_y\|A\hat{x}^N\|_2 &\ge 2R_y\|Ax^N\|_2 - 2R_yA_N\|A\hat{x}^N - Ax^N\|_2 \\
&\overset{(4.173)}{\ge} 2R_y\|Ax^N\|_2 - \frac{R_y^2\left(\sqrt{96C_2H} + G\right)}{A_N}. \qquad (4.174)
\end{aligned}$$

The next step is in applying Lipschitz continuity of f on $B_{R_f}(0)$. Recall that

$$x(y) \overset{\text{def}}{=} \operatorname*{argmax}_{x\in\mathbb{R}^n}\{\langle y, x\rangle - f(x)\}$$

and due to Demyanov-Danskin theorem $x(y) = r\varphi(y)$. Together with L'-smoothness of φ it implies that

$$\begin{aligned} \|x(A^\top \hat{y}^{k+1})\|_2 &= \|r\varphi(A^\top \hat{y}^{k+1})\|_2 \le \|r\varphi(A^\top \hat{y}^{k+1}) - r\varphi(A^\top y^*)\|_2 + \|r\varphi(A^\top y^*)\|_2 \\ &\le L'\|A^\top \hat{y}^{k+1} - A^\top y^*\|_2 + \|x(A^\top y^*)\|_2 \\ &\le \frac{\sqrt{\lambda_{\max}(A^\top A)}}{\mu}\|\hat{y}^{k+1} - y^*\|_2 + R_x. \end{aligned}$$

From this and (4.162) we get that with probability at least $1 - 2\beta$ the inequality

$$\|x(A^\top \hat{y}^{k+1})\|_2 \overset{(4.162)}{\le} \left(\frac{\sqrt{\lambda_{\max}(A^\top A)}J}{\mu} + \frac{R_x}{R_y}\right) R_y \tag{4.175}$$

holds for all $k = 0, 1, 2, \ldots, N-1$ simultaneously since $\hat{y}^{k+1} \in B_{R_k}(y^*) \subseteq B_{\widetilde{R}_{k+1}}(y^*)$. Using the convexity of the norm we get that with probability at least $1 - 2\beta$

$$\|\hat{x}^N\|_2 \le \frac{1}{A_N}\sum_{k=0}^{N-1}\alpha_{k+1}\|x(A^\top \hat{y}^{k+1})\|_2 \overset{(4.175)}{\le} \left(\frac{\sqrt{\lambda_{\max}(A^\top A)}J}{\mu} + \frac{R_x}{R_y}\right) R_y. \tag{4.176}$$

We notice that the last inequality lies in the same probability event when (4.162) holds.

Consider the probability event $E = \{$inequalities (4.171)$-$(4.176) hold simultaneously$\}$. Using union bound we get that $\mathbf{P}\{E\} \ge 1 - 4\beta$. Combining (4.172) and (4.176) we get that inequality

$$\begin{aligned} \|x^N\|_2 &\le \|x^N - \hat{x}^N\|_2 + \|\hat{x}^N\|_2 \\ &\le \left(\frac{\left(\sqrt{96C_2H} + G\right)}{2A_N\sqrt{\lambda_{\max}(A^\top A)}} + \frac{\sqrt{\lambda_{\max}(A^\top A)}J}{\mu} + \frac{R_x}{R_y}\right) R_y \end{aligned} \tag{4.177}$$

lies in the event E. From this we can obtain a lower bound for R_f:

$$R_f \ge \left(\frac{\left(\sqrt{96C_2H} + G\right)}{2A_N\sqrt{\lambda_{\max}(A^\top A)}} + \frac{\sqrt{\lambda_{\max}(A^\top A)}J}{\mu} + \frac{R_x}{R_y}\right) R_y.$$

Then we get that the fact that points x^N and $\hat{x}^N$ lie in $B_{R_f}(0)$ is a consequence of E. Therefore, we can apply Lipschitz-continuity of f for the points x^N and $\hat{x}^N$ and get that inequalities

$$|f(\hat{x}^N) - f(x^N)| \le L_f\|\hat{x}^N - x^N\|_2 \overset{(4.172)}{\le} \frac{L_f R_y\left(\sqrt{96C_2H} + G\right)}{2A_N\sqrt{\lambda_{\max}(A^\top A)}} \tag{4.178}$$

and

$$f(\hat{x}^N) = f(x^N) + \left(f(\hat{x}^N) - f(x^N)\right) \overset{(4.178)}{\ge} f(x^N) - \frac{L_f R_y\left(\sqrt{96C_2H} + G\right)}{2A_N\sqrt{\lambda_{\max}(A^\top A)}} \tag{4.179}$$

also lie in the event E. It remains to use inequalities (4.174) and (4.179) to bound first and second terms in the right hand side of inequality (4.171) and obtain that with probability at least $1 - 4\beta$

$$\begin{aligned}\psi(y^N) + f(x^N) + 2R_y\|Ax^N\|_2 \leq \frac{R_y^2}{A_N}\Bigg(&8\sqrt{HC_2} + 2 + 12CH + \frac{G(6J+4)}{2} \\ &+ \frac{L_f\left(\sqrt{96C_2H} + G\right)}{2R_y\sqrt{\lambda_{\max}(A^\top A)}} + \frac{G^2}{2(N+1)} \\ &+ C_1\sqrt{\frac{CHJg(N)}{2}} + \sqrt{96C_2H} + G\Bigg) \end{aligned} \quad (4.180)$$

Using that A_N grows as $\left(\frac{N^2}{L}\right)$ [1], $L \leq \frac{2\lambda_{\max}(A^\top A)}{}$ and $R_y \leq \frac{\|x^*\|_2^2}{\lambda_{\min}^+(A^\top A)}$ (see Section V-D from [46] for the details), we obtain that the choice of N in the theorem statement guarantees that the r.h.s. of the last inequality is no greater than ε. By weak duality $-f(x^*) \leq \psi(y^*)$ and we have with probability at least $1 - 4\beta$

$$f(x^N) - f(x^*) \leq f(x^N) + \psi(y^*) \leq f(x^N) + \psi(y^N) \leq \varepsilon. \quad (4.181)$$

Since y^* is the solution of the dual problem, we have, for any x, $f(x^*) \leq f(x) - \langle y^*, Ax\rangle$. Then using assumption $\|y^*\|_2 \leq R_y$, Cauchy-Schawrz inequality $\langle y, Ax\rangle \geq -\|y^*\|_2 \cdot \|Ax\|_2 \geq -R_y\|Ax\|_2$ and choosing $x = x^N$, we get

$$f(x^N) \geq f(x^*) - R_y\|Ax^N\|_2 \quad (4.182)$$

Using this and weak duality $-f(x^*) \leq \psi(y^*)$, we obtain

$$\psi(y^N) + f(x^N) \geq \psi(y^*) + f(x^N) \geq -f(x^*) + f(x^N) \geq -R_y\|Ax^N\|_2,$$

which implies that inequality

$$\|Ax^N\|_2 \overset{(4.180)+(4.181)}{\leq} \frac{\varepsilon}{R_y} \quad (4.183)$$

holds together with (4.181) with probability at least $1 - 4\beta$. The total number of stochastic gradient oracle calls is $\sum_{k=1}^{N} r_k$, which gives the bound in the problem statement since $\sum_{k=1}^{N} \alpha_{k+1} = A_N$. □

4.9.7. Missing Proofs from Section 4.5.2

Proof of Theorem 4.5.5

For simplicity we analyse only the first restart since the analysis of the later restarts is the same. We apply Theorem 4.5.3 with $N = N$ such that

$$\frac{CL^2 \ln^4 N}{\mu^2 N^4} \leq \frac{1}{32}$$

and batch-size

$$r_1 = \max\left\{1, \frac{64C\sigma^2 \ln^6 N}{N k r \quad (y^0, \quad ^0, \hat{\quad}_1) k_2^2}\right\}$$

together with simple inequality $k r \psi(y^0) k_2 \geq \mu \ k y^0 - y^* k_2$ and get for all $p = 1, \ldots, p_1$

$$\begin{aligned} \mathbf{E}\left[k r \psi(y^{1;p}) k_2^2 \mid y^0, r_1, \hat{\quad}_1\right] &\leq \frac{k r \psi(y^0) k_2^2}{32} + \frac{k r \quad (y^0, \quad ^0, \hat{\quad}_1) k_2^2}{64} \\ &\overset{(4.124)}{\leq} \frac{k r \psi(y^0) k_2^2}{16} + \frac{k r \quad (y^0, \quad ^0, \hat{\quad}_1) - \quad r \psi(y^0) k_2^2}{32} \end{aligned} \tag{4.184}$$

By Markov's inequality we have for each $p = 1, \ldots, p_1$ that for fixed $r \quad (y^0, \quad ^0, \hat{\quad}_1)$ with probability at most $1/2$

$$k r \psi(y^{1;p}) k_2^2 \geq \frac{k r \psi(y^0) k_2^2}{8} + \frac{k r \quad (y^0, \quad ^0, \hat{\quad}_1) - \quad r \psi(y^0) k_2^2}{16}.$$

Then, with probability at least $1 - 1/2^{p_1} \geq 1 - \quad /l$

$$k r \psi(y^{1;\hat{p}_1}) k_2^2 \leq \frac{k r \psi(y^0) k_2^2}{8} + \frac{k r \quad (y^0, \quad ^0, \hat{\quad}_1) - \quad r \psi(y^0) k_2^2}{16}, \tag{4.185}$$

where $\hat{p}_1$ is such that $k r \psi(y^{1;\hat{p}_1}) k_2^2 = \min_{p=1;\ldots;p_1} k r \psi(y^{1;p}) k_2^2$. From Lemma 4.9.5 we have for all $p = 1, \ldots, p_1$

$$\mathbf{P}\left\{\left\| r \quad (y^{1;p}, \quad ^{1;p}, r_1) - \quad r \psi(y^{1;p})\right\|_2 \geq \left(\sqrt{p} + \sqrt{2\gamma}\right)\sqrt{\frac{\sigma^2}{r_1}} \mid y^{1;p}\right\} \leq \exp\left(-\frac{\gamma^2}{3}\right).$$

Since $r_1 = \max\left\{1, \frac{128 \ \frac{2}{\psi}\left(1+\sqrt{3\ln\frac{lp_1}{\beta}}\right)^2 R_y^2}{\varepsilon^2}\right\}$ we can take $\gamma = \sqrt{3 \ln \frac{lp_1}{}}$ in the previous inequality and get that for all $p = 1, \ldots, p_1$ and fixed points $y^{1;p}$ with probability at least $1 - /(lp_1)$

$$\left\| r \quad (y^{1;p}, \quad ^{1;p}, r_1) - \quad r \psi(y^{1;p})\right\|_2^2 \leq \frac{\varepsilon^2}{64 R_y^2}.$$

Using union bound we get that with probability at least $1 - /l$ inequality

$$\left\| r \quad (y^{1;p}, \quad ^{1;p}, r_1) - \quad r \psi(y^{1;p})\right\|_2^2 \leq \frac{\varepsilon^2}{64 R_y^2}. \tag{4.186}$$

holds for all $p = 1, \dots, p_1$ simultaneously with fixed points $y^{1;p}$. Using union bound again we get that with probability at least $1 - 2\,/l$ for fixed $r \quad (y^0, \ ^0, \hat{r}_1)$

$$\begin{aligned}
k r\,\psi(y^{1;p(1)})k_2^2 &\overset{(4.124)}{\leq} 2\left\| r \quad (y^{1;p(1)}, \ ^{1;p(1)}, r_1)\right\|_2^2 \\
&\quad + 2\left\| r \quad (y^{1;p(1)}, \ ^{1;p(1)}, r_1) - r\,\psi(y^{1;p(1)})\right\|_2^2 \\
&\overset{(4.186)}{\leq} 2\left\| r \quad (y^{1;\hat{p}_1}, \ ^{1;\hat{p}_1}, r_1)\right\|_2^2 + \frac{\varepsilon^2}{32R_y^2} \\
&\overset{(4.124)}{\leq} 4k r\,\psi(y^{1;\hat{p}_1})k_2^2 + 4\left\| r \quad (y^{1;\hat{p}_1}, \ ^{1;\hat{p}_1}, r_1) - r\,\psi(y^{1;\hat{p}_1})\right\|_2^2 + \frac{\varepsilon^2}{32R_y^2} \\
&\overset{(4.185)+(4.186)}{\leq} \frac{k r\,\psi(y^0)k_2^2}{2} + \frac{k r \quad (y^0, \ ^0, \hat{r}_1) - r\,\psi(y^0)k_2^2}{4} + \frac{\varepsilon^2}{8R_y^2}. \qquad (4.187)
\end{aligned}$$

Using Lemma 4.9.5 with $\gamma = \sqrt{3\ln\frac{l}{}}$ and $\hat{r}_1 = \max\left\{1, \frac{4\,\frac{2}{\psi}\left(1+\sqrt{3\ln\frac{l}{\beta}}\right)^2 R_y^2}{\text{"}^2}\right\}$ we get that with probability at least $1 - \ /l$

$$k r \quad (y^0, \ ^0, \hat{r}_1) - r\,\psi(y^0)k_2^2 \leq \frac{\varepsilon^2}{2R_y^2}. \qquad (4.188)$$

Applying union bound again we get that with probability at least $1 - 3\,/l$ the following inequality holds:

$$k r\,\psi(y^{1;p(1)})k_2^2 \overset{(4.187)+(4.188)}{\leq} \frac{k r\,\psi(y^0)k_2^2}{2} + \frac{\varepsilon^2}{4R_y^2}.$$

Similarly, for all $k = 1, \dots, l$ with probability at least $1 - 3\,/l$

$$k r\,\psi(y^{k;p(k)})k_2^2 \leq \frac{k r\,\psi(y^{k-1;p(k-1)})k_2^2}{2} + \frac{\varepsilon^2}{4R_y^2}.$$

Using union bound we get that with probability at least $1 - 3\beta$ the inequality

$$k r\,\psi(y^{k;p(k)})k_2^2 \leq \frac{k r\,\psi(y^{k-1;p(k-1)})k_2^2}{2} + \frac{\varepsilon^2}{4R_y^2} \qquad (4.189)$$

holds for all $k = 1, \dots, l$ simultaneously. Finally, unrolling the recurrence an using our choice of $l = \max\{1, \log_2(2R_y^2 k r \quad (y^0)k_2^2/\text{"}^2)\}$ we obtain that with probability at least $1 - 3\beta$

$$\begin{aligned}
k r\,\psi(y^{l;p(l)})k_2^2 &\overset{(4.189)}{\leq} \frac{k r\,\psi(y^0)k_2^2}{2^l} + \frac{\varepsilon^2}{4R_y^2}\sum_{k=0}^{l-1} 2^{-k} \\
&\leq \frac{\varepsilon^2}{2R_y^2} + \frac{\varepsilon^2}{4R_y^2}\sum_{k=0}^{1} 2^{-k} \\
&= \frac{\varepsilon^2}{2R_y^2} + \frac{\varepsilon^2}{4R_y^2}\cdot 2 = \frac{\varepsilon^2}{R_y^2},
\end{aligned}$$

which concludes the proof. To get (4.55) we need to estimate $\sum_{k=1}^{l}(\hat{r}_k + Np_k r_k + p_k r_k)$ using our choice of parameters stated in (4.53).

Proof of Corollary 4.5.3

Theorem 4.5.5, Corollary 4.5.2 and inequality $\varepsilon \le \mu\ R_y^2$ imply that with probability at least $1 - 3\beta$

$$\|r\,\psi(y^{l;p(l)})\|_2 \le \frac{\varepsilon}{R_y}, \quad \|y^{l;p(l)}\|_2 \le \|y^{l;p(l)} - y^*\|_2 + \|y^*\|_2 \overset{(4.56)}{\le} 2R_y. \tag{4.190}$$

Applying Theorem 4.5.2 we get that with probability $1 - 3\beta$ we also have

$$f(\hat{x}^l) - f(x^*) \le 2\varepsilon, \quad \|A\hat{x}^l\|_2 \le \frac{\varepsilon}{R_y}, \tag{4.191}$$

where $\hat{x}^l \overset{\text{def}}{=} x(A^\top y^{l;p(l)})$. Next, we show that points $\hat{x}^{l;p} = x(A^\top y^{l;p})$ and $x^{l;p} \overset{\text{def}}{=} x(A^\top y^{l;p}, {}^{l;}, r_l)$ are *close* to each other with high probability for all $p = 1, \ldots, p_l$ and both lie in $B_{R_f}(0)$ with high probability. Lemma 4.9.5 states that

$$\mathbf{P}\left\{\left\|\hat{x}^{l;p} - x^{l;p}\right\|_2 \ge (\sqrt{2} + \sqrt{2\gamma})\sqrt{\frac{\sigma_x^2}{r_l}}\, j\, y^{l;p(l)}\right\} \le \exp\left(-\frac{\gamma^2}{3}\right).$$

Taking $\gamma = \sqrt{3 \ln \frac{p_l}{}}$ and using $r_l = \max\left\{1, \frac{128\ \frac{2}{\psi}\left(1+\sqrt{3\ln\frac{lp_l}{\beta}}\right)R_y^2}{\text{“2}}\right\}$ we get that for all $p = 1, \ldots, p_l$ with probability at least $1 - \ /p_l$

$$\|\hat{x}^{l;p} - x^{l;p}\|_2 \le \frac{\varepsilon}{8R_y} \cdot \sqrt{\frac{\sigma_x^2}{\sigma^2}} = \frac{\varepsilon}{8R_y\sqrt{\lambda_{\max}(A^\top A)}},$$

where we use $\sigma\ = \sqrt{\lambda_{\max}(A^\top A)}\sigma_x$. Using union bound we get that with probability at least $1 - \beta$ the inequality

$$\|\hat{x}^{l;p} - x^{l;p}\|_2 \le \frac{\varepsilon}{8R_y\sqrt{\lambda_{\max}(A^\top A)}},$$

holds for all $p = 1, \ldots, p(l)$ simultaneously and, in particular, we get that with probability at least $1 - \beta$

$$\|\hat{x}^l - x^l\|_2 \le \frac{\varepsilon}{8R_y\sqrt{\lambda_{\max}(A^\top A)}}. \tag{4.192}$$

It implies that with probability at least $1 - \beta$

$$\begin{aligned}\|A\hat{x}^l - Ax^l\|_2 &\le \|A\|_2 \cdot \|\hat{x}^l - x^l\|_2 \\ &\overset{(4.192)}{\le} \sqrt{\lambda_{\max}(A^\top A)}\frac{\varepsilon}{8R_y\sqrt{\lambda_{\max}(A^\top A)}} = \frac{\varepsilon}{8R_y},\end{aligned} \tag{4.193}$$

and due to triangle inequality with probability $\ge 1 - \beta$

$$\|A\hat{x}^l\|_2 \ge \|Ax^l\|_2 - \|A\hat{x}^l - Ax^l\|_2 \overset{(4.193)}{\ge} \|Ax^l\|_2 - \frac{\varepsilon}{8R_y}. \tag{4.194}$$

Applying Demyanov-Danskin's theorem, L'-smoothness of φ with $L' = 1/\ $ and $\varepsilon \le \mu\ R_y^2$ we obtain that with probability at least $1 - \beta$

$$\begin{aligned} k\hat{x}^l k_2 &= k\, r\,\varphi(A^\top y^{l;p(l)})k_2 \le k\, r\,\varphi(A^\top y^{l;p(l)}) - r\,\varphi(A^\top y^*)k_2 + k\, r\,\varphi(A^\top y^*)k_2 \\ &\le L' k A^\top y^{l;p(l)} - A^\top y^* k_2 + k x(A^\top y^*)k_2 \le \frac{\sqrt{\lambda_{\max}(A^\top A)}}{\mu} k y^{l;p(l)} - y^* k_2 + R_x \\ &\overset{(4.56)}{\le} \frac{\sqrt{\lambda_{\max}(A^\top A)}\varepsilon}{\mu\mu\ R_y} + R_x \le \left(\frac{\sqrt{\lambda_{\max}(A^\top A)}}{\mu} + \frac{R_x}{R_y}\right) R_y \end{aligned} \tag{4.195}$$

and also

$$\begin{aligned} k x^l k_2 &\le k x^l - \hat{x}^l k_2 + k\hat{x}^l k_2 \\ &\overset{(4.192)+(4.195)}{\le} \left(\frac{\mu}{8\sqrt{\lambda_{\max}(A^\top A)}} + \frac{\sqrt{\lambda_{\max}(A^\top A)}}{\mu} + \frac{R_x}{R_y}\right) R_y. \end{aligned} \tag{4.196}$$

That is, we proved that with probability at least $1 - \beta$ points $\hat{x}^l$ and x^l lie in the ball $B_{R_f}(0)$. In this ball function f is L_f-Lipschitz continuous, therefore, with probability at least $1 - \beta$

$$\begin{aligned} f(\hat{x}^l) &= f(x^l) + f(\hat{x}^l) - f(x^l) \ge f(x^l) - j f(\hat{x}^l) - f(x^l) j \\ &\ge f(x^l) - L_f k\hat{x}^l - x^l k_2 \overset{(4.192)}{\ge} f(x^l) - \frac{\varepsilon L_f}{8R_y\sqrt{\lambda_{\max}(A^\top A)}}. \end{aligned} \tag{4.197}$$

Combining inequalities (4.191), (4.194) and (4.197) and using union bound we get that with probability at least $1 - 4\beta$

$$f(x^l) - f(x^*) \le \left(2 + \frac{L_f}{8R_y\sqrt{\lambda_{\max}(A^\top A)}}\right)\varepsilon, \quad k A x^l k \le \frac{9\varepsilon}{8R_y}.$$

Finally, in order to get the bound for the total number of oracle calls from (4.58) we use (4.55) together with $\sigma^2 = \sigma_x^2 \lambda_{\max}(A^\top A)$ and (4.125).

4.9.8. Missing Proofs from Section 4.5.3

Proof of Lemma 4.5.1

We prove (4.62) by induction. For $k = 0$ this inequality is trivial since $A_k = \frac{1}{L}$, $\mathcal{y}^1 = y^0$ and $z^0 = \mathcal{y}^0$. Next, assume that (4.62) holds for some $k \ge 0$ and prove it for $k + 1$. By definition of $g_{k+1}(z)$ we have

$$\begin{aligned} \mathcal{g}_{k+1}(z^{k+1}) &= \mathcal{g}_k(z^{k+1}) \\ &\quad + \alpha_{k+1}\left(\psi(\mathcal{y}^{k+1}) + h\mathcal{r}\ \ (\mathcal{y}^{k+1},\ \ ^{k+1}), z^{k+1} - \mathcal{y}^{k+1} i + \frac{\mu}{2} k z^{k+1} - \mathcal{y}^{k+1} k_2^2\right). \end{aligned} \tag{4.198}$$

Since $g_k(z)$ is $(1 + A_k\mu)$-strongly convex we can estimate the first term in the r.h.s. of the previous inequality as follows:

$$\begin{aligned} g_k(z^{k+1}) &\geq g_k(z) + \frac{1 + A_k\mu}{2}\|z^{k+1} - z^k\|_2^2 \\ &\overset{(4.62)}{\geq} A_k\psi(y^k) + \frac{1 + A_k\mu}{2}\|z^{k+1} - z^k\|_2^2 \\ &\quad + \sum_{l=0}^{k-1}\frac{A_l\mu}{2}\|y^l - x^{l+1}\|_2^2 - \sum_{l=0}^{k}\frac{\alpha_l}{2\mu}\left\|\nabla g_l(x^l, u^l) - \nabla\psi(x^l)\right\|_2^2 \end{aligned}$$

Applying μ-strong convexity of ψ and the relation

$$\begin{aligned} y^{k+1} &= \frac{A_k y^k + \alpha_{k+1}z^{k+1}}{A_{k+1}} = \frac{A_k y^k + \alpha_{k+1}z^k}{A_{k+1}} + \frac{\alpha_{k+1}}{A_{k+1}}\left(z^{k+1} - z^k\right) \\ &= x^{k+1} + \frac{\alpha_{k+1}}{A_{k+1}}\left(z^{k+1} - z^k\right) \end{aligned}$$

to the previous inequality we get

$$\begin{aligned} g_k(z^{k+1}) &\geq A_k\psi(x^{k+1}) + \langle\nabla\psi(x^{k+1}), A_k(y^k - x^{k+1})\rangle + \frac{A_k\mu}{2}\|y^k - x^{k+1}\|_2^2 \\ &\quad + \frac{A_{k+1}^2(1 + A_k\mu)}{2\alpha_{k+1}^2}\|y^{k+1} - x^{k+1}\|_2^2 + \sum_{l=0}^{k-1}\frac{A_l\mu}{2}\|y^l - x^{l+1}\|_2^2 \\ &\quad - \sum_{l=0}^{k}\frac{\alpha_l}{2\mu}\left\|\nabla g_l(x^l, u^l) - \nabla\psi(x^l)\right\|_2^2. \end{aligned} \tag{4.199}$$

Next, we use (4.199) in (4.198) together with relations $A_{k+1} = A_k + \alpha_{k+1}$, $A_{k+1}(1 + A_k\mu) = \alpha_{k+1}^2 L$ and $A_k(y^k - x^{k+1}) + \alpha_{k+1}(z^{k+1} - x^{k+1}) = A_{k+1}(y^{k+1} - x^{k+1})$:

$$\begin{aligned} g_{k+1}(z^{k+1}) &\geq A_{k+1}\psi(x^{k+1}) + \langle\nabla\psi(x^{k+1}), A_k(y^k - x^{k+1}) + \alpha_{k+1}(z^{k+1} - x^{k+1})\rangle \\ &\quad + \frac{A_{k+1}^2(1 + A_k\mu)}{2\alpha_{k+1}^2}\|y^{k+1} - x^{k+1}\|_2^2 + \sum_{l=0}^{k}\frac{A_l\mu}{2}\|y^l - x^{l+1}\|_2^2 \\ &\quad - \sum_{l=0}^{k}\frac{\alpha_l}{2\mu}\left\|\nabla g_l(x^l, u^l) - \nabla\psi(x^l)\right\|_2^2 \\ &\quad + \alpha_{k+1}\left\langle\nabla g_{k+1}(x^{l+1}, u^{l+1}) - \nabla\psi(x^{l+1}), z^{k+1} - x^{k+1}\right\rangle \\ &\quad + \frac{\alpha_{k+1}\mu}{2}\|z^{k+1} - x^{k+1}\|_2^2 \\ &= A_{k+1}\left(\psi(x^{k+1}) + \langle\nabla\psi(x^{k+1}), y^{k+1} - x^{k+1}\rangle + \frac{L}{2}\|y^{k+1} - x^{k+1}\|_2^2\right) \\ &\quad + \sum_{l=0}^{k}\frac{A_l\mu}{2}\|y^l - x^{l+1}\|_2^2 - \sum_{l=0}^{k}\frac{\alpha_l}{2\mu}\left\|\nabla g_l(x^l, u^l) - \nabla\psi(x^l)\right\|_2^2 \\ &\quad + \alpha_{k+1}\left\langle\nabla g_{k+1}(x^{l+1}, u^{l+1}) - \nabla\psi(x^{l+1}), z^{k+1} - x^{k+1}\right\rangle \\ &\quad + \frac{\alpha_{k+1}\mu}{2}\|z^{k+1} - x^{k+1}\|_2^2. \end{aligned}$$

From L -smoothness of ψ we have

$$\psi(\cdot^{k+1}) + h\, r\, \psi(\cdot^{k+1}), y^{k+1} - \cdot^{k+1}i + \frac{L}{2}k y^{k+1} - \cdot^{k+1}k_2^2 \geq \psi(y^{k+1}).$$

Next, Fenchel-Young inequality (see inequality (4.123)) implies that

$$\left\langle \tau \quad (\cdot^{l+1}, \quad ^{l+1}) - \quad r\, \psi(\cdot^{l+1}), z^{k+1} - \cdot^{k+1} \right\rangle$$
$$\geq -\tfrac{1}{2\ \psi} \left\| \tau \quad (\cdot^{l+1}, \quad ^{l+1}) - \quad r\, \psi(\cdot^{l+1}) \right\|_2^2 - \tfrac{\psi}{2} k z^{k+1} - \cdot^{k+1} k_2^2.$$

Putting all together and rearranging the terms we get

$$\cdot_{k+1}(z^{k+1}) \geq A_{k+1}\psi(y^{k+1}) + \sum_{l=0}^{k} \frac{A_l \mu}{2} k y^l - \cdot^{l+1} k_2^2 - \sum_{l=0}^{k+1} \frac{\alpha_l}{2\mu} \left\| \tau \quad (\cdot^l, \quad ^l) - \quad r\, \psi(\cdot^l) \right\|_2^2.$$

Proof of Lemma 4.5.2

The idea behind the proof of this lemma is exactly the same as for Lemma 4.9.8. We start with applying Cauchy-Schwarz inequality to the second and the third terms, i.e.

$$\begin{aligned} h\delta(R_k + \widetilde{R}_k) &\leq Dh^2\delta^2 + \frac{R_k^2}{4D} + Dh^2\delta^2 + \frac{\widetilde{R}_k^2}{4D} = 2Dh^2\delta^2 + \frac{R_k^2 + \widetilde{R}_k^2}{4D}, \\ u h\eta^k, a^k + \alpha^k i &\leq u k\eta^k k_2 \cdot k a^k k_2 + u k\eta^k k_2 \cdot k\alpha^k k_2 \leq u k\eta^k k_2 R_k + u k\eta^k k_2 \widetilde{R}_k \\ &\leq u^2 D k\eta^k k_2^2 + \frac{R_k^2}{4D} + u^2 D k\eta^k k_2^2 + \frac{\widetilde{R}_k^2}{4D} \leq 2u^2 D k\eta^k k_2^2 + \frac{R_k^2 + \widetilde{R}_k^2}{4D}, \end{aligned}$$

in the right-hand side of (4.63):

$$\begin{aligned} A_l R_l^2 + \sum_{k=0}^{l-1} A_k \widetilde{R}_k^2 \quad \leq \quad & A + 2Dh^2\delta^2 \underbrace{\sum_{k=0}^{l-1} \alpha_{k+1}}_{A_l} + \frac{1}{2D} \sum_{k=0}^{l-1} \alpha_{k+1}(R_k^2 + \widetilde{R}_k^2) \\ & + \left(c + 2Du^2\right) \sum_{k=0}^{l-1} \alpha_{k+1} k\eta^k k_2^2. \end{aligned} \tag{4.200}$$

Using Lemma 4.9.5 we get that with probability at least $1 - \ _{\overline{N}}$

$$\begin{aligned} k\eta^k k_2 \quad &\leq \quad {}^{p}\overline{2}\left(1 + \sqrt{3\ln\frac{N}{\beta}}\right)\sigma_k \leq {}^{p}\overline{2}\left(1 + \sqrt{3\ln\frac{N}{\beta}}\right) \frac{{}^{p}\overline{C\varepsilon}}{N\left(1 + \sqrt{3\ln\frac{N}{}}\right)} \\ &= \quad {}^{p}\overline{2C\varepsilon}. \end{aligned} \tag{4.201}$$

Using union bound and $\alpha_{k+1} \leq DA_k$ we get that with probability $\geq 1 - \beta$ inequalities

$$\begin{aligned} A_l R_l^2 + \sum_{k=0}^{l-1} A_k \widetilde{R}_k^2 \quad &\leq \quad A + 2Dh^2\delta^2 A_l + \frac{1}{2}\sum_{k=0}^{l-1} A_k(R_k^2 + \widetilde{R}_k^2) + 2C\left(c + 2Du^2\right) A_l\varepsilon, \\ A_l R_l^2 + \frac{1}{2}\sum_{k=0}^{l-1} A_k \widetilde{R}_k^2 \quad &\leq \quad A + 2Dh^2\delta^2 A_l + \frac{1}{2}\sum_{k=0}^{l-1} A_k R_k^2 + 2C\left(c + 2Du^2\right) A_l\varepsilon \end{aligned} \tag{4.202}$$

hold for all $l = 1, \ldots, N$ simultaneously. Therefore, with probability $\geq 1 - \beta$ the inequality

$$\begin{aligned} A_l R_l^2 &\leq A + 2Dh^2\delta^2 A_l + 2C\left(c + 2Du^2\right) A_l\varepsilon + \frac{1}{2}\sum_{k=0}^{l-1} A_k R_k^2 \\ &\leq \frac{3}{2}A + 2Dh^2\delta^2 \underbrace{\left(A_l + \frac{1}{2}A_{l-1}\right)}_{\leq \frac{3}{2}A_l} + 2C\left(c + 2Du^2\right)\varepsilon \underbrace{\left(A_l + \frac{1}{2}A_{l-1}\right)}_{\leq \frac{3}{2}A_l} + \frac{3}{2}\cdot\frac{1}{2}\sum_{k=0}^{l-2} A_k R_k^2 \\ &\leq \frac{3}{2}\left(A + 2Dh^2\delta^2 A_l + 2C\left(c + 2Du^2\right) A_l\varepsilon + \frac{1}{2}\sum_{k=0}^{l-2} A_k R_k^2\right), \end{aligned}$$

holds for all $l = 1, \ldots, N$ simultaneously. Unrolling the recurrence we get that with probability $\geq 1 - \beta$

$$A_l R_l^2 \leq \left(\frac{3}{2}\right)^l \left(A + 2Dh^2\delta^2 A_l + 2C\left(c + 2Du^2\right) A_l\varepsilon\right),$$

for all $l = 1, \ldots, N$. We emphasize that it is very rough estimate, but as for the convex case we show next that such a bound does not spoil the final result too much. It implies that with probability $\geq 1 - \beta$

$$\sum_{k=0}^{l-1} A_k R_k^2 \leq l\left(\frac{3}{2}\right)^l \left(A + 2Dh^2\delta^2 A_l + 2C\left(c + 2Du^2\right) A_l\varepsilon\right), \tag{4.203}$$

for all $l = 1, \ldots, N$ simultaneously. Moreover, since (4.202) holds we have in the same probability event that inequalities

$$\sum_{k=0}^{l-1} A_k \widetilde{R}_k^2 \leq \left(l\left(\frac{3}{2}\right)^l + 2\right)\left(A + 2Dh^2\delta^2 A_l + 2C\left(c + 2Du^2\right) A_l\varepsilon\right) \tag{4.204}$$

hold with probability $\geq 1 - \beta$ for all $l = 1, \ldots, N$ simultaneously with (4.203). Next we apply delicate result from [92] which is presented in Section 4.9.3 as Lemma 4.9.4. We consider random variables $\xi^k = \alpha_{k+1}\hbar\eta^k, a^k + \varpi^k\imath$. Note that $\mathbf{E}\left[\xi^k \jmath\, \xi^0, \ldots, \xi^{k-1}\right] = \alpha_{k+1}\left\langle \mathbf{E}\left[\eta^k \jmath\, \eta^0, \ldots, \eta^{k-1}\right], a^k\right\rangle = 0$ and

$$\begin{aligned} \mathbf{E}\left[\exp\left(\frac{(\xi^k)^2}{2\sigma_k^2\alpha_{k+1}^2(R_k^2 + \widetilde{R}_k^2)}\right) \jmath\, \xi^0, \ldots, \xi^{k-1}\right] &\leq \mathbf{E}\left[\exp\left(\frac{\alpha_{k+1}^2 \hbar\eta^k \hbar_2^2 \hbar a^k + \varpi^k \hbar_2^2}{2\sigma_k^2\alpha_{k+1}^2(R_k^2 + \widetilde{R}_k^2)}\right) \jmath\, \eta^0, \ldots, \eta^{k-1}\right] \\ &= \mathbf{E}\left[\exp\left(\frac{\hbar\eta^k \hbar_2^2}{\sigma_k^2}\right) \jmath\, \eta^0, \ldots, \eta^{k-1}\right] \leq \exp(1) \end{aligned}$$

due to Cauchy-Schwarz inequality and assumptions of the lemma. If we denote $\eth_k^2 = 2\sigma_k^2\alpha_{k+1}^2(R_k^2 + \widetilde{R}_k^2)$ and apply Lemma 4.9.4 with

$$B = 8HCDR_0^2\left(N\left(\frac{3}{2}\right)^N + 1\right)\left(A + 2Dh^2G^2R_0^2 + 2C\left(c + 2Du^2\right)HR_0^2\right)$$

and $b = \theta_0^2$, we get that for all $l = 1, \dots, N$ with probability $\geq 1 - \frac{}{N}$

$$\text{either} \sum_{k=0}^{l-1} \theta_k^2 \geq B \text{ or } \left| \sum_{k=0}^{l-1} \xi^k \right| \leq C_1 \sqrt{\sum_{k=0}^{l-1} \theta_k^2 \left(\ln\left(\frac{N}{\beta}\right) + \ln\ln\left(\frac{B}{b}\right)\right)}$$

with some constant $C_1 > 0$ which does not depend on B or b. Using union bound we obtain that with probability $\geq 1 - \beta$

$$\text{either} \sum_{k=0}^{l-1} \theta_k^2 \geq B \text{ or } \left| \sum_{k=0}^{l-1} \xi^k \right| \leq C_1 \sqrt{\sum_{k=0}^{l-1} \theta_k^2 \left(\ln\left(\frac{N}{\beta}\right) + \ln\ln\left(\frac{B}{b}\right)\right)}$$

and it holds for all $l = 1, \dots, N$ simultaneously. Note that $\alpha_{k+1} \leq A_{k+1}$, $\varepsilon \leq \frac{HR_0^2}{A_N}$, $\delta \leq \frac{GR_0}{N\sqrt{A_N}}$ and with probability at least $1 - \beta$

$$\begin{aligned}
\sum_{k=0}^{l-1} \theta_k^2 &= 2\sum_{k=0}^{l-1} \sigma_k^2 \alpha_{k+1}^2 (R_k^2 + \widetilde{R}_k^2) \leq \frac{2C\varepsilon}{N^2 \left(1 + \sqrt{3\ln\frac{N}{}}\right)^2} \sum_{k=0}^{l-1} A_{k+1} \cdot DA_k (R_k^2 + \widetilde{R}_k^2) \\
&\leq 2\varepsilon CDA_N \sum_{k=0}^{l-1} A_k (R_k^2 + \widetilde{R}_k^2) \\
&\overset{(4.203)+(4.204)}{\leq} 4\varepsilon CDA_N \left(l \left(\frac{3}{2}\right)^l + 1 \right) \left(A + 2Dh^2\delta^2 A_l + 2C\left(c + 2Du^2\right) A_l \varepsilon \right) \\
&\leq 4HCDR_0^2 \left(N \left(\frac{3}{2}\right)^N + 1 \right) \left(A + 2Dh^2 G^2 R_0^2 + 2C\left(c + 2Du^2\right) HR_0^2 \right) \\
&= \frac{B}{2}
\end{aligned}$$

for all $l = 1, \dots, N$ simultaneously. Using union bound again we get that with probability $\geq 1 - 2\beta$ the inequality

$$\left| \sum_{k=0}^{l-1} \xi^k \right| \leq C_1 \sqrt{\sum_{k=0}^{l-1} \theta_k^2 \left(\ln\left(\frac{N}{\beta}\right) + \ln\ln\left(\frac{B}{b}\right)\right)} \tag{4.205}$$

holds for all $l = 1, \dots, N$ simultaneously.

Note that we also proved that (4.201) is in the same event together with (4.205) and holds with probability $\geq 1 - 2\beta$. Putting all together in (4.63), we get that with probability

at least $1 - 2\beta$ the inequality

$$A_l R_l^2 + \sum_{k=0}^{l-1} A_k \widetilde{R}_k^2 \overset{(4.63)}{\leq} A + h\delta \sum_{k=0}^{l-1} \alpha_{k+1}(R_k + \widetilde{R}_k) + u \sum_{k=0}^{l-1} \alpha_{k+1} \langle h\eta^k, a^k + \widetilde{a}^k \rangle + c \sum_{k=0}^{l-1} \alpha_{k+1} \|h\eta^k\|_2^2$$

$$\overset{(4.201)+(4.205)}{\leq} A + h\delta \sum_{k=0}^{l-1} \alpha_{k+1}(R_k + \widetilde{R}_k) + uC_1 \sqrt{\sum_{k=0}^{l-1} \theta_k^2 \left(\ln\left(\frac{N}{\beta}\right) + \ln\ln\left(\frac{B}{b}\right) \right)} + 2cC\varepsilon A_l$$

holds for all $l = 1, \ldots, N$ simultaneously. For brevity, we introduce new notation: $g(N) = \frac{\ln\left(\frac{N}{\beta}\right) + \ln\ln\left(\frac{B}{b}\right)}{\left(1 + \sqrt{3\ln\left(\frac{N}{\beta}\right)}\right)^2} \approx 1$ (neglecting constant factor). Using our assumptions $\sigma_k^2 \leq \frac{C''}{N^2\left(1+\sqrt{3\ln\left(\frac{N}{\beta}\right)}\right)^2}$, $\varepsilon \leq \frac{HR_0^2}{A_N}$, $\delta \leq \frac{GR_0}{N\sqrt{A_N}}$ and definition $\theta_k^2 = 2\sigma_k^2 \alpha_{k+1}^2 (R_k^2 + \widetilde{R}_k^2)$ we obtain that with probability at least $1 - 2\beta$ the inequality

$$\begin{aligned} A_l R_l^2 + \sum_{k=0}^{l-1} A_k \widetilde{R}_k^2 &\leq A + h\delta \sum_{k=0}^{l-1} \alpha_{k+1}(R_k + \widetilde{R}_k) + u \sum_{k=0}^{l-1} \alpha_{k+1} \langle h\eta^k, a^k + \widetilde{a}^k \rangle \\ &\quad + c \sum_{k=0}^{l-1} \alpha_{k+1} \|h\eta^k\|_2^2 \\ &\leq A + \frac{hGR_0}{N\sqrt{A_N}} \sum_{k=0}^{l-1} \alpha_{k+1}(R_k + \widetilde{R}_k) \\ &\quad + \frac{uC_1 R_0 \sqrt{2HCg(N)}}{N\sqrt{A_N}} \sqrt{\sum_{k=0}^{l-1} \alpha_{k+1}^2 (R_k^2 + \widetilde{R}_k^2)} + 2cHCR_0^2 \\ &\leq \left(\frac{A}{R_0^2} + 2cHC \right) R_0^2 \\ &\quad + \frac{\left(hG + uC_1\sqrt{2HCg(N)} \right) R_0}{N\sqrt{A_N}} \sum_{k=0}^{l-1} \alpha_{k+1}(R_k + \widetilde{R}_k) \end{aligned} \tag{4.206}$$

holds for all $l = 1, \ldots, N$ simultaneously, where in the last row we applied well-known inequality: $\sqrt{\sum_{i=1}^m a_i^2} \leq \sum_{i=1}^m a_i$ for $a_i \geq 0$, $i = 1, \ldots, m$. Next we use Lemma 4.9.13 with $A = \frac{A}{R_0^2} + 2cHC$, $B = hG + uC_1\sqrt{2HCg(N)}$, $r_k = R_k$, $\widetilde{r}_k = \widetilde{R}_k$ and get that with probability at least $1 - 2\beta$ inequalities

$$R_l \leq \frac{JR_0}{\sqrt{A_l}}, \quad \widetilde{R}_{l-1} \leq \frac{JR_0}{\sqrt{A_{l-1}}}$$

hold for all $l = 1, \dots, N$ simultaneously with

$$J = \max\left\{\sqrt{A_0}, \frac{3B_1D + \sqrt{9B_1^2D^2 + \frac{4A}{R_0^2} + 8cHC}}{2}\right\}, \; B_1 = hG + uC_1\sqrt{2HCg(N)}.$$

It implies that with probability at least $1 - 2\beta$ the inequality

$$\begin{aligned}
A + h\delta\sum_{k=0}^{l-1}\alpha_{k+1}(R_k + \widetilde{R}_k) + u\sum_{k=0}^{l-1}\alpha_{k+1}\langle\eta^k, a^k + a\rangle + c\sum_{k=0}^{l-1}\alpha_{k+1}\|\eta^k\|_2^2 \\
&\le \left(\frac{A}{R_0^2} + 2cHC\right)R_0^2 + \frac{2J\left(hG+uC_1\sqrt{2HCg(N)}\right)R_0^2}{N\sqrt{A_N}}\sum_{k=0}^{l-1}\sqrt{A_{k+1}} \\
&\le A + \left(2cHC + \frac{2JD\left(hG+uC_1\sqrt{2HCg(N)}\right)}{N\sqrt{A_N}}\sum_{k=0}^{l-1}\sqrt{A_k}\right)R_0^2 \\
&\le A + \left(2cHC + \frac{2JD\left(hG+uC_1\sqrt{2HCg(N)}\right)}{N\sqrt{A_N}}l\sqrt{A_{l-1}}\right)R_0^2 \\
&\le A + \left(2cHC + 2JD\left(hG + uC_1\sqrt{2HCg(N)}\right)\right)R_0^2
\end{aligned}$$

holds for all $l = 1, \dots, N$ simultaneously.

Proof of Theorem 4.5.6

From Lemma 4.5.1 we have

$$A_k\psi(y^k) \le g_k(z^k) - \sum_{l=0}^{k-1}\frac{A_l\mu}{2}\|y^l - y^{l+1}\|_2^2 + \sum_{l=0}^{k}\frac{\alpha_l}{2\mu}\left\|\nabla f(y^l, \xi^l) - \nabla\psi(y^l)\right\|_2^2 \quad (4.207)$$

for all $k \ge 0$. By definition of z^k we get that

$$\begin{aligned}
g_k(z^k) &= \min_{z\in\mathbb{R}^n}\left\{\frac{1}{2}\|z - z^0\|_2^2 + \sum_{l=0}^{k}\alpha_l\left(\psi(y^l) + \langle\nabla f(y^l, \xi^l), z - y^l\rangle + \frac{\mu}{2}\|z - y^l\|_2^2\right)\right\} \\
&\le \frac{1}{2}\|y^* - z^0\|_2^2 + \sum_{l=0}^{k}\alpha_l\left(\psi(y^l) + \langle\nabla f(y^l, \xi^l), y^* - y^l\rangle + \frac{\mu}{2}\|y^* - y^l\|_2^2\right) \\
&= \frac{1}{2}\|y^* - z^0\|_2^2 + \sum_{l=0}^{k}\alpha_l\left(\psi(y^l) + \langle\nabla\psi(y^l), y^* - y^l\rangle + \frac{\mu}{2}\|y^* - y^l\|_2^2\right) \\
&\quad + \sum_{l=0}^{k}\alpha_l\langle\nabla f(y^l, \xi^l) - \nabla\psi(y^l), y^* - y^l\rangle \\
&\le \frac{1}{2}\|y^* - y^0\|_2^2 + A_k\psi(y^*) + \sum_{l=0}^{k}\alpha_l\langle\nabla f(y^l, \xi^l) - \nabla\psi(y^l), y^* - y^l\rangle, \qquad (4.208)
\end{aligned}$$

where the last inequality follows from μ -strong convexity of ψ and $A_k = \sum_{l=0}^{k}\alpha_l$. For brevity, we introduce new notation: $R_k \overset{\text{def}}{=} \|y^k - y^*\|_2$ and $\widetilde{R}_k \overset{\text{def}}{=} \|y^k - y^{k+1}\|_2$ for all $k \ge 0$.

Using this and another consequence of strong convexity, i.e. $\psi(y) - \psi(y^*) \geq \frac{\mu}{2}\|y - y^*\|_2^2$, we obtain

$$\begin{aligned}\frac{A_k\mu}{2}R_k^2 + \sum_{l=0}^{k-1}\frac{A_l\mu}{2}\widetilde{R}_l^2 &\leq A_k\left(\psi(y^k) - \psi(y^*)\right) + \sum_{l=0}^{k-1}\frac{A_l\mu}{2}\widetilde{R}_l^2 \\ &\overset{(4.207)+(4.208)}{\leq} \frac{1}{2}R_0^2 + \sum_{l=0}^{k}\alpha_l\left\langle \tau\left(g^l, \xi^l\right) - \nabla\psi(g^l), y^* - g^l\right\rangle \\ &\quad + \sum_{l=0}^{k}\frac{\alpha_l}{2\mu}\left\|\tau\left(g^l, \xi^l\right) - \nabla\psi(g^l)\right\|_2^2. \quad (4.209)\end{aligned}$$

From Cauchy-Schwarz inequality and the well-known fact that $\|a + b\|_2^2 \leq 2a^2 + 2b^2$ for all $a, b \in \mathbb{R}^n$ we have

$$\begin{aligned}\left\langle \tau\left(g^l, \xi^l\right) - \nabla\psi(g^l), y^* - g^l\right\rangle &= \left\langle \mathbf{E}\left[\tau\left(g^l, \xi^l\right)\right] - \nabla\psi(g^l), y^* - g^l\right\rangle \\ &\quad + \left\langle \tau\left(g^l, \xi^l\right) - \mathbf{E}\left[\tau\left(g^l, \xi^l\right)\right], y^* - g^l\right\rangle \\ &\overset{(4.42)}{\leq} \delta\|y^* - g^l\|_2 + \left\langle \tau\left(g^l, \xi^l\right) - \mathbf{E}\left[\tau\left(g^l, \xi^l\right)\right], y^* - g^l\right\rangle, \\ \left\|\tau\left(g^l, \xi^l\right) - \nabla\psi(g^l)\right\|_2^2 &\leq 2\left\|\mathbf{E}\left[\tau\left(g^l, \xi^l\right)\right] - \nabla\psi(g^l)\right\|_2^2 \\ &\quad + 2\left\|\tau\left(g^l, \xi^l\right) - \mathbf{E}\left[\tau\left(g^l, \xi^l\right)\right]\right\|_2^2 \\ &\overset{(4.42)}{\leq} 2\delta^2 + 2\left\|\tau\left(g^l, \xi^l\right) - \mathbf{E}\left[\tau\left(g^l, \xi^l\right)\right]\right\|_2^2\end{aligned}$$

for all $l \geq 0$. Next, we introduce new notation

$$\begin{aligned}\mathcal{A} &\overset{\text{def}}{=} \frac{1}{2}R_0^2 + \delta\alpha_0 R_0 + \frac{A_N\delta^2}{\mu} + \alpha_0\left\langle \tau\left(g^0, \xi^0\right) - \mathbf{E}\left[\tau\left(g^0, \xi^0\right)\right], y^* - g^0\right\rangle \\ &\quad + \frac{\alpha_0}{\mu}\left\|\tau\left(g^0, \xi^0\right) - \mathbf{E}\left[\tau\left(g^0, \xi^0\right)\right]\right\|_2^2. \quad (4.210)\end{aligned}$$

Putting all together in (4.209) we get

$$\begin{aligned}\frac{A_k\mu}{2}R_k^2+\sum_{l=0}^{k-1}\frac{A_l\mu}{2}\widetilde{R}_l^2 &\le \frac{1}{2}R_0^2+\delta\sum_{l=0}^{k}\alpha_l\|y^*-y^l\|_2\\ &\quad+\sum_{l=0}^{k}\alpha_l\left\langle \nabla f(y^l,\xi^l)-\mathbf{E}\left[\nabla f(y^l,\xi^l)\right],y^*-y^l\right\rangle\\ &\quad+\frac{\delta^2}{\mu}\sum_{l=0}^{k}\alpha_l+\frac{1}{\mu}\sum_{l=0}^{k}\alpha_l\left\|\nabla f(y^l,\xi^l)-\mathbf{E}\left[\nabla f(y^l,\xi^l)\right]\right\|_2^2\\ &\le A+\delta\sum_{l=0}^{k-1}\alpha_{l+1}\|y^*-y^{l+1}\|_2\\ &\quad+\sum_{l=0}^{k-1}\alpha_{l+1}\left\langle \nabla f(y^{l+1},\xi^{l+1})-\mathbf{E}\left[\nabla f(y^{l+1},\xi^{l+1})\right],y^*-y^{l+1}\right\rangle\\ &\quad+\frac{1}{\mu}\sum_{l=0}^{k-1}\alpha_{l+1}\left\|\nabla f(y^{l+1},\xi^{l+1})-\mathbf{E}\left[\nabla f(y^{l+1},\xi^{l+1})\right]\right\|_2^2 \quad (4.211)\end{aligned}$$

To simplify previous inequality we define new vectors $a^l \overset{\text{def}}{=} y^*-y^l$, $\alpha^l \overset{\text{def}}{=} y^l-y^{l+1}$, $\eta^l \overset{\text{def}}{=} \nabla f(y^{l+1},\xi^{l+1})-\mathbf{E}\left[\nabla f(y^{l+1},\xi^{l+1})\right]$ for all $l\ge 0$. Note that $\|a^l\|_2=R_l$, $\|\alpha^l\|_2=\widetilde{R}_l$ and $\alpha^0=y^0-y^1=0$. Using this we can rewrite (4.211) in the following form:

$$\begin{aligned}A_kR_k^2+\sum_{l=0}^{k-1}A_l\widetilde{R}_l^2 &\le A+\frac{2\delta}{\mu}\sum_{l=0}^{k-1}\alpha_{l+1}(R_l+\widetilde{R}_l)+\frac{2}{\mu}\sum_{l=0}^{k-1}\alpha_{l+1}\langle\eta^l,a^l+\alpha^l\rangle\\ &\quad+\frac{2}{\mu^2}\sum_{l=0}^{k-1}\alpha_{l+1}\|\eta^l\|_2^2, \quad (4.212)\end{aligned}$$

where we used $A \overset{\text{def}}{=} \frac{2A}{\psi}$ and triangle inequality, i.e. $\|y^*-y^{l+1}\|_2\le\|y^*-y^l\|_2+\|y^l-y^{l+1}\|_2=R_l+\widetilde{R}_l$. Next, we apply Lemma 4.5.2 with $h=u=\frac{2}{\psi}$, $c=\frac{2}{\frac{2}{\psi}}$ and get that with probability at least $1-2\beta$

$$R_N^2\le\frac{J^2R_0^2}{A_N} \quad (4.213)$$

where

$$g(N)=\frac{\ln\left(\frac{N}{\beta}\right)+\ln\ln\left(\frac{B}{b}\right)}{\left(1+\sqrt{3\ln\left(\frac{N}{\beta}\right)}\right)^2},\quad b=\frac{2\sigma_1^2\alpha_1^2R_0^2}{r_1},\quad D\overset{(4.229)}{=}1+\frac{\mu}{L}+\sqrt{1+\frac{\mu}{L}},$$

$$\begin{aligned}B &= 8HC\left(\frac{L}{\mu}\right)^{3/2}DR_0^2\left(N\left(\frac{3}{2}\right)^N+1\right)\left(A+2Dh^2G^2R_0^2\right.\\ &\quad\left.+2C\left(\frac{L}{\mu}\right)^{3/2}\left(c+2Du^2\right)HR_0^2\right),\end{aligned}$$

$$J = \max\left\{\sqrt{A_0}, \frac{3B_1D + \sqrt{9B_1^2D^2 + \frac{4A}{R_0^2} + 8cHC\left(\frac{L_\psi}{\psi}\right)^{3=2}}}{2}\right\},$$

$$B_1 = hG + uC_1\sqrt{2HC\left(\frac{L}{\mu}\right)^{3=2} g(N)}$$

and C_1 is some positive constant. However, J depends on A which is stochastic. That is, to finish the proof we need first to get an upper bound for A. Recall that $A = \frac{2A}{\psi}$ and

$$A \overset{(4.210)}{=} \frac{R_0^2}{\mu} + \frac{2\delta\alpha_0 R_0}{\mu} + \frac{2A_N\delta^2}{\mu^2} + \frac{2\alpha_0}{\mu}\left\langle \nabla \quad (y^0, \ {}^0) - \mathbf{E}\left[\nabla \quad (y^0, \ {}^0)\right], y^* - y^0\right\rangle + \frac{2\alpha_0}{\mu^2}\left\| \nabla \quad (y^0, \ {}^0) - \mathbf{E}\left[\nabla \quad (y^0, \ {}^0)\right]\right\|_2^2. \tag{4.214}$$

Lemma 4.9.5 implies that

$$\mathbf{P}\left\{\left\| \nabla \quad (y^0, \ {}^0) - \mathbf{E}\left[\nabla \quad (y^0, \ {}^0)\right]\right\|_2 \geq \sqrt{2}(1 + \sqrt{\gamma})\sqrt{\frac{\sigma^2}{r_0}}\right\} \leq \exp\left(-\frac{\gamma^2}{3}\right).$$

Taking $\gamma = \sqrt{3\ln\frac{1}{}}$ and using $r_0 \geq \left(\frac{\psi}{L_\psi}\right)^{3=2} \frac{N^2 \ \frac{2}{\psi}\left(1+\sqrt{3\ln\frac{N}{\beta}}\right)^2}{C''}$, $\varepsilon \leq \frac{HR_0^2}{A_N}$ we get that with probability at least $1 - \beta$

$$\begin{aligned}
\left\langle \nabla \quad (y^0, \ {}^0) - \mathbf{E}\left[\nabla \quad (y^0, \ {}^0)\right], y^* - y^0\right\rangle &\leq \left\| \nabla \quad (y^0, \ {}^0) - \mathbf{E}\left[\nabla \quad (y^0, \ {}^0)\right]\right\|_2 \cdot \|y^* - y^0\|_2 \\
&\leq \left(\frac{L}{\mu}\right)^{3=4} \frac{\sqrt{2C\varepsilon}R_0}{N} \\
&\leq \left(\frac{L}{\mu}\right)^{3=4} \frac{\sqrt{2CH}R_0^2}{N\sqrt{A_N}}, \qquad (4.215)\\
\left\| \nabla \quad (y^0, \ {}^0) - \mathbf{E}\left[\nabla \quad (y^0, \ {}^0)\right]\right\|_2^2 &\leq \left(\frac{L}{\mu}\right)^{3=2} 2C\varepsilon \leq \left(\frac{L}{\mu}\right)^{3=2} \frac{2CHR_0^2}{N^2A_N}. \qquad (4.216)
\end{aligned}$$

From this and $\delta \leq \frac{GR_0}{N\sqrt{A_N}}$ we obtain that with probability $\geq 1 - \beta$

$$\begin{aligned}
A &\overset{(4.214)+(4.215)+(4.216)}{\leq} \hat{A}R_0^2, \\
\hat{A} &\overset{\text{def}}{=} \frac{1}{\mu} + \frac{2G}{L\ \mu\ N\sqrt{A_N}} + \frac{2G^2}{\mu^2N^2} + \left(\frac{L}{\mu}\right)^{3=4} \frac{2^{\sqrt{2CH}}}{L\ \mu\ N\sqrt{A_N}} \\
&\quad + \left(\frac{L}{\mu}\right)^{3=2} \frac{4CH}{L\ \mu^2N^2A_N}.
\end{aligned}$$

Using union bound we get that with probability at least $1 - 3\beta$

$$R_N^2 \leq \frac{\hat{J}^2R_0^2}{A_N},$$

where

$$\hat{g}(N) = \frac{\ln\left(\frac{N}{}\right) + \ln\ln\left(\frac{\hat{B}}{b}\right)}{\left(1 + \sqrt{3\ln\left(\frac{N}{}\right)}\right)^2},$$

$$\hat{B} = 8HC\left(\frac{L}{\mu}\right)^{3=2} DR_0^4\left(N\left(\frac{3}{2}\right)^N + 1\right)\left(\hat{A} + 2Dh^2G^2 + 2C\left(\frac{L}{\mu}\right)^{3=2}\left(c + 2Du^2\right)H\right),$$

$$\hat{J} = \max\left\{\sqrt{A_0}, \frac{3\hat{B}_1 D + \sqrt{9\hat{B}_1^2 D^2 + 4\hat{A} + 8cHC\left(\frac{L_\psi}{\psi}\right)^{3=2}}}{2}\right\},$$

$$\hat{B}_1 = hG + uC_1\sqrt{2HC\left(\frac{L}{\mu}\right)^{3=2}\hat{g}(N)}.$$

Note that

$$A_k \overset{(4.228)}{\geq} \frac{1}{L}\left(1 + \frac{1}{2}\sqrt{\frac{\mu}{L}}\right)^{2k}.$$

It means that in order to achieve $R_N^2 \leq \varepsilon$ with probability at least $1 - 3\beta$ the method requires $N = \hat{O}\left(\sqrt{\frac{L_\psi}{\psi}}\ln\frac{1}{"}\right)$ iterations and

$$\sum_{k=0}^{N} r_k = \widetilde{O}\left(\max\left\{\sqrt{\frac{L}{\mu}}, \frac{\sigma^2}{\varepsilon}\ln\frac{1}{\beta}\right\}\ln\frac{1}{\varepsilon}\right)$$

oracle calls where $\widetilde{O}(\cdot)$ hides polylogarithmic factors depending on L , μ , R_0, ε and β.

Proof of Corollary 4.5.5

Corollary 4.5.4 implies that with probability at least $1 - 3\beta$

$$ky^N k_2 \leq 2R_y, \quad k\,r\,\psi(y^N)k_2 \leq \frac{\varepsilon}{R_y}$$

and the total number of oracle calls to get this is of order (4.76). Together with Theorem 4.5.2 it gives us that with probability at least $1 - 3\beta$

$$f(\hat{x}^N) - f(x^*) \leq 2\hat{\varepsilon}, \quad kA\hat{x}^N k_2 \leq \frac{\hat{\varepsilon}}{R_y}, \tag{4.217}$$

where $\hat{x}^N \overset{\text{def}}{=} x(A^> y^N)$. It remains to show that x^N and $\hat{x}^N$ are *close* to each other with high probability. Lemma 4.9.5 states that

$$\mathbf{P}\left\{\left\|x^N - \mathbf{E}\left[x^N \,j\, y^N\right]\right\|_2 \geq (\sqrt{2} + \sqrt{2\gamma})\sqrt{\frac{\sigma_x^2}{r_N}}\, j\, y^N\right\} \leq \exp\left(-\frac{\gamma^2}{3}\right).$$

Taking $\gamma = \sqrt{3 \ln \frac{1}{\beta}}$ and using $r_N \geq \frac{1}{C} \frac{\frac{2}{\psi} R_y^2 \left(1 + \sqrt{3 \ln \frac{1}{\beta}}\right)^2}{n^2}$ we get that with probability at least $1 - \beta$

$$\begin{aligned}
\left\| x^N - \mathbf{E}\left[x^N \mid y^N\right] \right\|_2 &\leq \sqrt{2C \frac{\sigma_x^2 \varepsilon^2}{\sigma^2 R_y^2}} = \frac{\sqrt{2C}\varepsilon}{R_y \sqrt{\lambda_{\max}(A^\top A)}}, \\
\left\| x^N - \hat{x}^N \right\|_2 &\leq \left\| x^N - \mathbf{E}\left[x^N \mid y^N\right] \right\|_2 + \left\| \mathbf{E}\left[x^N \mid y^N\right] - \hat{x}^N \right\|_2 \\
&\overset{(4.37)}{\leq} \frac{\sqrt{2C}\varepsilon}{R_y \sqrt{\lambda_{\max}(A^\top A)}} + \frac{G_1 \varepsilon}{N R_y} \\
&\leq \left(\sqrt{\frac{2C}{\lambda_{\max}(A^\top A)}} + G_1 \right) \frac{\varepsilon}{R_y}. \qquad (4.218)
\end{aligned}$$

It implies that with probability at least $1 - \beta$

$$\begin{aligned}
\|Ax^N - A\hat{x}^N\|_2 &\leq \|A\|_2 \cdot \|x^N - \hat{x}^N\|_2 \\
&\overset{(4.218)}{\leq} \sqrt{\lambda_{\max}(A^\top A)} \left(\sqrt{\frac{2C}{\lambda_{\max}(A^\top A)}} + G_1 \right) \frac{\varepsilon}{R_y} \\
&= \left(\sqrt{2C} + G_1 \sqrt{\lambda_{\max}(A^\top A)} \right) \frac{\varepsilon}{R_y}, \qquad (4.219)
\end{aligned}$$

and due to triangle inequality with probability $\geq 1 - \beta$

$$\begin{aligned}
\|A\hat{x}^N\|_2 &\geq \|Ax^N\|_2 - \|A\hat{x}^N - Ax^N\|_2 \\
&\overset{(4.219)}{\geq} \|Ax^N\|_2 - \left(\sqrt{2C} + G_1 \sqrt{\lambda_{\max}(A^\top A)} \right) \frac{\varepsilon}{R_y}. \qquad (4.220)
\end{aligned}$$

Applying Demyanov-Danskin theorem and L'-smoothness of φ with $L' = 1/\mu$ we obtain that with probability at least $1 - \beta$

$$\begin{aligned}
\|\hat{x}^N\|_2 &= \|r\,\varphi(A^\top y^N)\|_2 \leq \|r\,\varphi(A^\top y^N) - r\,\varphi(A^\top y^*)\|_2 + \|r\,\varphi(A^\top y^*)\|_2 \\
&\leq L' \|Ax^N - Ax^*\|_2 + \|x(A^\top y^*)\|_2 \leq \frac{\sqrt{\lambda_{\max}(A^\top A)}}{\mu} \|y^N - y^*\|_2 + R_x \\
&\overset{(4.71)}{\leq} \frac{\sqrt{\lambda_{\max}(A^\top A)}\varepsilon}{\mu R_y} + R_x \qquad (4.221)
\end{aligned}$$

and also

$$\begin{aligned}
\|x^N\|_2 &\leq \|x^N - \hat{x}^N\|_2 + \|\hat{x}^N\|_2 \\
&\overset{(4.218)+(4.221)}{\leq} \left(\sqrt{\frac{2C}{\lambda_{\max}(A^\top A)}} + G_1 + \frac{\sqrt{\lambda_{\max}(A^\top A)}}{\mu} \right) \frac{\varepsilon}{R_y} + R_x. \qquad (4.222)
\end{aligned}$$

That is, we proved that with probability at least $1 - \beta$ points $\hat{x}^l$ and x^l lie in the ball $B_{R_f}(0)$. In this ball function f is L_f-Lipschitz continuous, therefore, with probability at

least $1 - \beta$

$$\begin{aligned} f(\hat{x}^N) &= f(x^N) + f(\hat{x}^N) - f(x^N) \geq f(x^N) - |f(\hat{x}^N) - f(x^N)| \\ &\geq f(x^N) - L_f \|\hat{x}^N - x^N\|_2 \\ &\overset{(4.218)}{\geq} f(x^N) - \left(\sqrt{\frac{2C}{\lambda_{\max}(A^\top A)}} + G_1\right) \frac{L_f \varepsilon}{R_y}. \end{aligned} \tag{4.223}$$

Combining inequalities (4.217), (4.220) and (4.223) and using union bound we get that with probability at least $1 - 4\beta$

$$\begin{aligned} f(x^N) - f(x^*) &\leq \left(2 + \left(\sqrt{\frac{2C}{\lambda_{\max}(A^\top A)}} + G_1\right) \frac{L_f}{R_y}\right) \varepsilon, \\ \|Ax^N\|_2 &\leq \left(1 + \sqrt{2C} + G_1 \sqrt{\lambda_{\max}(A^\top A)}\right) \frac{\varepsilon}{R_y}. \end{aligned}$$

Finally, in order to get the bound for the total number of oracle calls from (4.76) we use (4.70) together with $\sigma^2 = \sigma_x^2 \lambda_{\max}(A^\top A)$ and (4.125).

4.9.9. Technical Results

Lemma 4.9.9. *For the sequence $\alpha_{k+1} \geq 0$ such that*

$$A_{k+1} = A_k + \alpha_{k+1}, \quad A_{k+1} = 2L\alpha_{k+1}^2 \tag{4.224}$$

we have for all $k \geq 0$

$$\alpha_{k+1} \leq \widetilde{\alpha}_{k+1} \overset{def}{=} \frac{k+2}{2L}. \tag{4.225}$$

Moreover, $A_k = \left(\frac{N^2}{L}\right)$.

Proof. We prove (4.225) by induction. For $k = 0$ equation (4.224) gives us $\alpha_1 = 2L\alpha_1^2 \ ( )$ $\alpha_1 = \frac{1}{2L}$. Next we assume that (4.225) holds for all $k \leq l - 1$ and prove it for $k = l$:

$$2L\alpha_{l+1}^2 \overset{(4.224)}{=} \sum_{i=1}^{l+1} \alpha_i \overset{(4.225)}{\leq} \alpha_{l+1} + \frac{1}{2L} \sum_{i=1}^{l} (i+1) = \alpha_{l+1} + \frac{l(l+3)}{4L}.$$

This quadratic inequality implies that $\alpha_{k+1} \leq \frac{1+\sqrt{4k^2+12k+1}}{4L} \leq \frac{1+\sqrt{(2k+3)^2}}{4L} \leq \frac{2k+4}{4L} = \frac{k+2}{2L}$.

Finally, the relation $A_k = \left(\frac{N^2}{L}\right)$ is proved in Lemma 1 from [16] (see also [1]). □

Lemma 4.9.10 (See Lemma 3 from [95] and Lemma 4 from [96]). *For the sequence $\alpha_{k+1} \geq 0$ such that*

$$A_{k+1} = A_k + \alpha_{k+1}, \quad A_{k+1}(1 + A_k\mu) = L\alpha_{k+1}^2, \quad \alpha_0 = A_0 = \frac{1}{L} \tag{4.226}$$

we have for all $k \geq 0$

$$\alpha_{k+1} = \frac{1 + A_k\mu}{2L} + \sqrt{\frac{(1 + A_k\mu)^2}{4L^2} + \frac{A_k(1 + A_k\mu)}{L}}, \tag{4.227}$$

$$A_k \geq \frac{1}{L}\left(1 + \frac{1}{2}\sqrt{\frac{\mu}{L}}\right)^{2k}, \tag{4.228}$$

$$\alpha_{k+1} \leq \left(1 + \frac{\mu}{L} + \sqrt{1 + \frac{\mu}{L}}\right) A_k. \tag{4.229}$$

Proof. If we solve quadratic equation $A_{k+1}(1 + A_k\mu) = L\alpha_{k+1}^2$, $A_{k+1} = A_k + \alpha_{k+1}$ with respect to α_{k+1}, we will get (4.227). Inequality (4.228) was established in Lemma 3 from [95] and Lemma 4 from [96]. It remains to prove (4.229). Since $\sqrt{a^2 + b^2} \leq a + b$ for all $a, b \geq 0$ and $A_k \geq A_0 = \frac{1}{L}$ we have

$$\begin{aligned}
\alpha_{k+1} &\overset{(4.227)}{=} \frac{1 + A_k\mu}{2L} + \sqrt{\frac{(1 + A_k\mu)^2}{4L^2} + \frac{A_k(1 + A_k\mu)}{L}} \\
&\leq \frac{1}{2L} + \frac{\mu}{2L}A_k + \frac{1 + A_k\mu}{2L} + \sqrt{\frac{A_k}{L} + \frac{\mu}{L}A_k^2} \\
&\leq \frac{1}{L} + \frac{\mu}{L}A_k + A_k\sqrt{1 + \frac{\mu}{L}} = \left(1 + \frac{\mu}{L} + \sqrt{1 + \frac{\mu}{L}}\right) A_k.
\end{aligned}$$

□

Lemma 4.9.11. *Let* $A, B, D, r_0, r_1, \ldots, r_N$*, where* $N \geq 1$*, be non-negative numbers such that*

$$\frac{1}{2}r_l^2 \leq Ar_0^2 + \frac{Dr_0}{(N+1)^2}\sum_{k=0}^{l-1}(k+2)r_k + B\frac{r_0}{N}\sqrt{\sum_{k=0}^{l-1}(k+2)r_k^2}, \quad \forall l = 1, \ldots, N. \tag{4.230}$$

Then for all $l = 0, \ldots, N$ *we have*

$$r_l \leq Cr_0, \tag{4.231}$$

where C *is such positive number that* $C^2 \geq \max\{2A + 2(B + D)C, 1\}$*, i.e. one can choose* $C = \max\{B + D + \sqrt{(B + D)^2 + 2A}, 1\}$*.*

Proof. We prove (4.231) by induction. For $l = 0$ the inequality $r_l \leq Cr_0$ trivially follows

since $C \geq 1$. Next we assume that (4.231) holds for some $l < N$ and prove it for $l+1$:

$$\begin{aligned} r_{l+1} &\overset{(4.230)}{\leq} \sqrt{2}\sqrt{Ar_0^2 + \frac{Dr_0}{(N+1)^2}\sum_{k=0}^{l}(k+2)r_k + B\frac{r_0}{N}\sqrt{\sum_{k=0}^{l}(k+2)r_k^2}} \\ &\overset{(4.231)}{\leq} r_0\sqrt{2}\sqrt{A + \frac{DC}{(N+1)^2}\sum_{k=0}^{l}(k+2) + \frac{BC}{N}\sqrt{\sum_{k=0}^{l}(k+2)}} \\ &\leq r_0\sqrt{2}\sqrt{A + \frac{DC}{(N+1)^2}\frac{(l+1)(l+2)}{2} + \frac{BC}{N}\sqrt{\frac{(l+1)(l+2)}{2}}} \\ &\leq r_0\sqrt{2}\sqrt{A + DC + \frac{BC}{N}\sqrt{\frac{N(N+1)}{2}}} \leq r_0\underbrace{\sqrt{2A + 2(B+D)C}}_{\leq C} \leq Cr_0. \end{aligned}$$

□

Lemma 4.9.12. *Let $C, r_0, r_1, \ldots, r_N$, where $N \geq 1$, be non-negative numbers such that*

$$r_l^2 \leq r_0^2 + \frac{2C}{(N+1)^{3/2}}\sum_{k=0}^{l-1}(k+2)^{1/2}r_{k+1}^2, \quad \forall l = 1, \ldots, N, \tag{4.232}$$

and C 2 $(0, 1/4)$. Then for all $l = 0, \ldots, N$ we have

$$r_l \leq 2r_0, \tag{4.233}$$

Proof. We prove (4.233) by induction. For $l = 0$ the inequality $r_l \leq 2r_0$ trivially follows. Next we assume that (4.233) holds for some $l \leq N-1$ and prove it for $l+1$. From (4.232), $C < 1/4$, $N \geq 1$ and $l \leq N-1$ we have

$$\begin{aligned} \frac{3}{4}r_{l+1}^2 &\leq \left(1 - \frac{2C(l+2)^{1/2}}{(N+1)^{3/2}}\right)r_{l+1}^2 \\ &\overset{(4.232)}{\leq} r_0^2 + \frac{2C}{(N+1)^{3/2}}\sum_{k=0}^{l-1}(k+2)^{1/2}r_{k+1}^2 \\ &\overset{(4.233)}{\leq} r_0^2 + \frac{1}{2(N+1)^{3/2}}l\cdot(l+1)^{1/2}\cdot 4r_0^2 \leq 3r_0^2, \end{aligned}$$

which implies $r_{l+1} \leq 2r_0$. □

Lemma 4.9.13. *Let $A, B, D, r_0, r_1, \ldots, r_N, \tilde{r}_0, \tilde{r}_1, \ldots, \tilde{r}_N, \alpha_0, \alpha_1, \ldots, \alpha_N$, where $N \geq 1$, be non-negative numbers such that*

$$A_l r_l^2 + \sum_{k=0}^{l-1} A_k \tilde{r}_k^2 \leq Ar_0^2 + \frac{Br_0}{N\sqrt{A_N}}\sum_{k=0}^{l-1}\alpha_{k+1}(r_k + \tilde{r}_k), \quad \forall l = 1, \ldots, N, \tag{4.234}$$

where $\tilde{r}_0 = 0$, $A_0 = \alpha_0 > 0$, $A_l = A_{l-1} + \alpha_l$ *and* $\alpha_l \le DA_{l-1}$ *for* $l = 1, \dots, N$ *and* $D \ge 1$. *Then for all* $l = 1, \dots, N$ *we have*

$$r_l \le \frac{Cr_0}{\sqrt{A_l}}, \quad \tilde{r}_{l-1} \le \frac{Cr_0}{\sqrt{A_{l-1}}} \tag{4.235}$$

and $r_0 \le \frac{Cr_0}{\sqrt{A_0}}$ *where* C *is such positive number that*

$$C \ge \max\left\{\sqrt{A_0}, \frac{BD}{2} + \sqrt{\frac{B^2D^2}{4} + A + 2BCD}\right\},$$

i.e. one can choose $C = \max\left\{\sqrt{A_0}, \frac{3BD + \sqrt{9B^2D^2 + 4A}}{2}\right\}$.

Proof. We prove (4.235) by induction. For $l = 1$ the inequality $\tilde{r}_0 \le \frac{Cr_0}{\sqrt{A_0}}$ trivially follows since $\tilde{r}_0 = 0$. What is more, (4.234) implies that

$$A_1 r_1^2 \le Ar_0^2 + \frac{B\alpha_1 r_0^2}{N\sqrt{A_N}} \quad \Rightarrow \quad r_1 \le r_0\sqrt{\frac{A}{A_1} + \frac{BDA_0}{A_1 N\sqrt{A_N}}} \le r_0\sqrt{\frac{A + BD\sqrt{A_0}}{A_1}} \le \frac{Cr_0}{\sqrt{A_1}},$$

since $C \ge \sqrt{A_0}$ and $C \ge \sqrt{A + BCD} \ge \sqrt{A + BD\sqrt{A_0}}$. Note that we also have $r_0 \le \frac{Cr_0}{\sqrt{A_0}}$. Next we assume that (4.235) holds for some $l \le N - 1$ and prove it for $l + 1$:

$$\begin{aligned}
A_l\tilde{r}_l^2 &\overset{(4.234)}{\le} Ar_0^2 + \frac{Br_0}{N\sqrt{A_N}}\sum_{k=0}^{l}\alpha_{k+1}(r_k + \tilde{r}_k) \\
&\overset{(4.235)}{\le} Ar_0^2 + \frac{BCr_0^2}{N\sqrt{A_N}}\sum_{k=0}^{l}\frac{\alpha_{k+1}}{\sqrt{A_k}} + \frac{BCr_0^2}{N\sqrt{A_N}}\sum_{k=0}^{l-1}\frac{\alpha_{k+1}}{\sqrt{A_k}} + \frac{Br_0\alpha_{l+1}\tilde{r}_l}{N\sqrt{A_N}} \\
&\le Ar_0^2 + \frac{BCDr_0^2}{N\sqrt{A_N}}\sum_{k=0}^{l}\sqrt{A_k} + \frac{BCDr_0^2}{N\sqrt{A_N}}\sum_{k=0}^{l-1}\sqrt{A_k} + \frac{BDr_0A_l\tilde{r}_l}{\sqrt{A_N}} \\
&\le Ar_0^2 + \frac{BCDr_0^2}{N\sqrt{A_N}}(l+1)\sqrt{A_l} + \frac{BCDr_0^2}{N\sqrt{A_N}}l\sqrt{A_{l-1}} + \frac{BDr_0A_l\tilde{r}_l}{\sqrt{A_N}} \\
&\le (A + 2BCD)r_0^2 + \frac{BDr_0A_l\tilde{r}_l}{\sqrt{A_N}} \\
0 &\ge \tilde{r}_l^2 - \frac{BDr_0\tilde{r}_l}{\sqrt{A_N}} - \frac{(A + 2BCD)r_0^2}{A_l}.
\end{aligned}$$

From this we have that $\tilde{r}_l$ is not greater than the biggest root of the quadratic equation corresponding to the last inequality, i.e.

$$\begin{aligned}
\tilde{r}_l &\le \frac{BDr_0}{2\sqrt{A_N}} + \sqrt{\frac{B^2D^2r_0^2}{4A_N} + \frac{(A + 2BCD)r_0^2}{A_l}} \\
&\le \underbrace{\left(\frac{BD}{2} + \sqrt{\frac{B^2D^2}{4} + A + 2BCD}\right)}_{\le C}\frac{r_0}{\sqrt{A_l}} \le \frac{Cr_0}{\sqrt{A_l}}.
\end{aligned}$$

It implies that

$$\begin{aligned} A_{l+1}r_{l+1}^2 &\overset{(4.234)}{\leq} Ar_0^2 + \frac{Br_0}{N\sqrt{A_N}}\sum_{k=0}^{l}\alpha_{k+1}(r_k + r_k) \\ &\overset{(4.235)}{\leq} Ar_0^2 + \frac{2BCr_0^2}{N\sqrt{A_N}}\sum_{k=0}^{l}\frac{\alpha_{k+1}}{\sqrt{A_k}} \\ &\leq Ar_0^2 + \frac{2BCDr_0^2}{N\sqrt{A_N}}(l+1)\sqrt{A_l} \leq Ar_0^2 + 2BCDr_0^2, \\ r_{l+1} &\leq r_0\sqrt{\frac{A+2BCD}{A_{l+1}}} \leq \frac{Cr_0}{\sqrt{A_{l+1}}}. \end{aligned}$$

That is, we proved the statement of the lemma for

$$C \geq \max\left\{\sqrt{A_0}, \frac{BD}{2} + \sqrt{\frac{B^2D^2}{4} + A + 2BCD}\right\}.$$

In particular, via solving the equation

$$C = \frac{BD}{2} + \sqrt{\frac{B^2D^2}{4} + A + 2BCD}$$

w.r.t. C one can show that the choice $C = \max\left\{\sqrt{A_0}, \frac{3BD+\sqrt{9B^2D^2+4A}}{2}\right\}$ satisfies the assumption of the lemma on C.

□

Chapter 5

Stochastic Derivative Free Optimization Methods with Momentum

The theoretical results proposed in this chapter were obtained by the author of this thesis in [97].

5.1. Introduction

In this paper, we consider the following minimization problem

$$\min_{x \in \mathbb{R}^n} f(x), \tag{5.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is "smooth"but not necessarily a convex function in a Derivative-Free Optimization (DFO) setting where only function evaluations are possible. The function f is bounded from below by $f(x^*)$ where x^* is a minimizer. Lastly and throughout the paper, we assume that f is L-smooth.

DFO. In DFO setting [98, 99], the derivatives of the objective function f are not accessible. That is they are either impractical to evaluate, noisy (function f is noisy) [100] or they are simply not available at all. In standard applications of DFO, evaluations of f are only accessible through simulations of black-box engine or software as in reinforcement learning and continuous control environments [101]. This setting of optimization problems appears also in applications from computational medicine [102] and fluid dynamics [103–105] to localization [106, 107] and continuous control [108, 109] to name a few.

The literature on DFO for solving (5.1) is long and rich. The first approaches were based on deterministic direct search (DDS) and they span half a century of work [110–112]. However, for DDS methods complexity bounds have only been established recently by the work of Vicente and coauthors [113, 114]. In particular, the work of Vicente [113] showed the first complexity results on non-convex f and the results were extended to better complexities when f is convex [114]. However, there have been several variants of DDS, including randomized approaches [115–120]. Only very recently, complexity bounds have also been derived for randomized methods [27, 121–124]. For instance, the work of [121, 124] imposes a decrease condition on whether to accept or reject a step of a set

of random directions. Moreover, [125] derived new complexity bounds when the random directions are normally distributed vectors for both smooth and non-smooth f. They proposed both accelerated and non-accelerated zero-order (ZO) methods. Accelerated derivative-free methods in the case of inexact oracle information was proposed in [15]. An extension of [125] for non-Euclidean proximal setup was proposed by [126] for the smooth stochastic convex optimization with inexact oracle. In [127, 128] authors also consider acceleration of ZO methods and, in particular, develop the method called SARP, proved that its convergence rate is not worse than for non-accelerated ZO methods and showed that in some cases it works even better.

More recently and closely related to our work, [129] proposed a new randomized direct search method called *Stochastic Three Points* (STP). At each iteration k STP generates a random search direction s_k according to a certain probability law and compares the objective function at three points: current iterate x_k, a point in the direction of s_k and a point in the direction of $-s_k$ with a certain step size α_k. The method then chooses the best of these three points as the new iterate:

$$x_{k+1} = \arg\min\{f(x_k), f(x_k + \alpha_k s_k), f(x_k - \alpha_k s_k)\}.$$

The key properties of STP are its simplicity, generality and practicality. Indeed, the update rule for STP makes it extremely simple to implement, the proofs of convergence results for STP are short and clear and assumptions on random search directions cover a lot of strategies of choosing decent direction and even some of first-order methods fit the STP scheme which makes it a very flexible in comparison with other zeroth-order methods (e.g. two-point evaluations methods like in [125], [27], [123], [126] that try to approximate directional derivatives along random direction at each iteration). Motivated by these properties of STP we focus on further developing of this method.

Momentum. Heavy ball momentum[1] is a special technique introduced by Polyak in 1964 [130] to get faster convergence to the optimum for the first-order methods. In the original paper, Polyak proved that his method converges *locally* with $O\left(\sqrt{L/\mu}\log 1/\varepsilon\right)$ rate for twice continuously differentiable μ-strongly convex and L-smooth functions. Despite the long history of this approach, there is still an open question whether heavy ball method converges to the optimum *globally* with accelerated rate when the objective function is twice continuous differentiable, L-smooth and μ-strongly convex. For this class of functions,

[1] We will refer to this as momentum.

only non-accelerated global convergence was proved [131] and for the special case of quadratic strongly convex and L-smooth functions Lessard et. al. [132] recently proved asymptotic accelerated global convergence. However, heavy ball method performs well in practice and, therefore, is widely used. One can find more detailed survey of the literature about heavy ball momentum in [133].

Importance Sampling. Importance sampling has been celebrated and extensively studied in stochastic gradient based methods [134] or in coordinate based methods [135]. Only very recently, [136] proposed, `STP_IS`, the first DFO algorithm with importance sampling. In particular, under coordinate-wise smooth function, they show that sampling coordinate directions, can be generalized to arbitrary directions, with probabilities proportional to the function coordinate smoothness constants, improves the leading constant by the same factor typically gained in gradient based methods.

Contributions. Our contributions can be summarized into three folds.

- **First ZO method with heavy ball momentum.** Motivated by practical effectiveness of first-order momentum heavy ball method, we introduce momentum into `STP` method and propose new DFO algorithm with heavy ball momentum (`SMTP`). We summarized the method in Algorithm 9, with theoretical guarantees for non-convex, convex and strongly convex functions under generic sampling directions D. We emphasize that the `SMTP` with momentum is not a straightforward generalization of `STP` and Polyak's method and requires insights from virtual iterates analysis from [137].

  To the best of our knowledge it is the first analysis of derivative-free method with heavy ball momentum, i.e. we show that the same momentum trick that works for the first order method could be applied for zeroth-order methods as well.

- **First ZO method with both heavy ball momentum and importance sampling.** In order to get more gain from momentum in the case when the sampling directions are coordinate directions and the objective function is coordinate-wise L-smooth (see Assumption 5.3.1), we consider importance sampling to the above method. In fact, we propose the first zeroth-order momentum method with importance sampling (`SMTP_IS`) summarized in Algorithm 10 with theoretical guarantees for non-convex, convex and strongly convex functions. The details and proofs are left for Section 5.3 and Appendix 5.7.3.

Algorithm 9 SMTP: Stochastic Momentum Three Points

Require: learning rates $(\gamma^k)_{k\geq 0}$, starting point $x^0 \in \mathbb{R}^d$, D — distribution on $\mathbb{R}^d$, $0 \leq \beta < 1$ — momentum parameter

Set $v^{-1} = 0$ and $z^0 = x^0$

for $k = 0, 1, \ldots$ **do**

  Sample $s^k \sim D$

  Let $v_+^k = \beta v^{k-1} + s^k$ and $v_-^k = \beta v^{k-1} - s^k$

  Let $x_+^{k+1} = x^k - \gamma^k v_+^k$ and $x_-^{k+1} = x^k - \gamma^k v_-^k$

  Let $z_+^{k+1} = x_+^{k+1} - \frac{k}{1-\beta} v_+^k$ and $z_-^{k+1} = x_-^{k+1} - \frac{k}{1-\beta} v_-^k$

  Set $z^{k+1} = \arg\min \left\{ f(z^k), f(z_+^{k+1}), f(z_-^{k+1}) \right\}$

  Set $x^{k+1} = \begin{cases} x_+^{k+1}, & \text{if } z^{k+1} = z_+^{k+1} \\ x_-^{k+1}, & \text{if } z^{k+1} = z_-^{k+1} \\ x^k, & \text{if } z^{k+1} = z^k \end{cases}$ and $v^{k+1} = \begin{cases} v_+^{k+1}, & \text{if } z^{k+1} = z_+^{k+1} \\ v_-^{k+1}, & \text{if } z^{k+1} = z_-^{k+1} \\ v^k, & \text{if } z^{k+1} = z^k \end{cases}$

end for

- **Practicality.** We conduct extensive experiments on continuous control tasks from the MuJoCo suite [101] following recent success of DFO compared to model-free reinforcement learning [108, 109]. We achieve with SMTP_IS the state-of-the-art results on across all tested environments on the continuous control outperforming DFO [108] and policy gradient methods [138, 139].

We provide more detailed comparison of SMTP and SMTP_IS in Section 5.4 of the Appendix.

5.2. Stochastic Momentum Three Points (SMTP)

Our analysis of SMTP is based on the following key assumption.

Assumption 5.2.1. *The probability distribution D on $\mathbb{R}^n$ satisfies the following properties:*

1. *The quantity $\gamma_D \overset{def}{=} \mathbf{E}_{s \sim D} \|s\|_2^2$ is finite.*
2. *There is a constant $\mu_D > 0$ for a norm $\|\cdot\|_D$ in $\mathbb{R}^n$ such that for all $g \in \mathbb{R}^n$*

$$\mathbf{E}_{s \sim D} |\langle g, s \rangle| \geq \mu_D \|g\|_D. \tag{5.2}$$

Assumptions on f	SMTP Complexity	Theorem	Importance Sampling	SMTP_IS Complexity	Theorem
None	$\frac{2r_0 L_{\langle D \rangle}}{\varepsilon^2_{\langle D \rangle}}$	5.2.1	$p_i = \frac{L_i}{\sum_{i=1}^n L_i}$	$\frac{2r_0 n \sum_{i=1}^n L_i}{\varepsilon^2}$	5.3.1
Convex, $R_0 < \infty$	$\frac{1}{\varepsilon} \frac{L_{\langle D \rangle} R_0^2}{\varepsilon^2_{\langle D \rangle}} \ln\left(\frac{2r_0}{\varepsilon}\right)$	5.2.2	$p_i = \frac{L_i}{\sum_{i=1}^n L_i}$	$\frac{R_0^2 n \sum_{i=1}^n L_i}{\varepsilon} \ln\left(\frac{2r_0}{\varepsilon}\right)$	5.3.2
μ-strongly convex	$\frac{L_{\langle D \rangle}}{\mu^2_{\langle D \rangle}} \ln\left(\frac{2r_0}{\varepsilon}\right)$	5.2.5	$p_i = \frac{L_i}{\sum_{i=1}^n L_i}$	$\frac{\sum_{i=1}^n L_i}{\mu} \ln\left(\frac{2r_0}{\varepsilon}\right)$	5.3.5

Table 5.1: Summary of the new derived complexity results of SMTP and SMTP_IS. The complexities for SMTP are under a generic sampling distribution D satisfying Assumption 5.2.1 while for SMTP_IS are under an arbitrary discrete sampling from a set of coordinate directions following [136] where we propose an importance sampling that improves the leading constant marked in red. Note that $r_0 = f(x_0) - f(x_*)$ and that all assumptions listed are in addition to L-smoothness. Complexity means number of iterations in order to guarantee $\mathbf{E}\|\nabla f(\overline{z}^K)\|_D \leq \varepsilon$ for the non-convex case, $\mathbf{E}\left[f(z^K) - f(x^*)\right] \leq \varepsilon$ for convex and strongly convex cases. $R_0 < \infty$ is the radius in $\|\cdot\|_D^*$-norm of a bounded level set where the exact definition is given in Assumption 5.2.2. We notice that for SMTP_IS $\|\cdot\|_D = \|\cdot\|_1$ and $\|\cdot\|_D^* = \|\cdot\|_\infty$ in non-convex and convex cases and $\|\cdot\|_D = \|\cdot\|_2$ in the strongly convex case.

Some examples of distributions that meet above assumption are described in Lemma 3.4 from [129]. For convenience we provide the statement of the lemma in the Appendix (see Lemma 5.7.4).

Recall that one possible view on STP [129] is as following. If we substitute gradient $\nabla f(x^k)$ in the update rule for the gradient descent $x^{k+1} = x^k - \gamma^k \nabla f(x^k)$ by $\pm s^k$ where s^k is sampled from distribution D satisfied Assumption 5.2.1 and then select x^{k+1} as the best point in terms of functional value among $x^k, x^k - \gamma^k s^k, x^k + \gamma^k s^k$ we will get exactly STP method. However, gradient descent is not the best algorithm to solve unconstrained smooth minimization problems and the natural idea is to try to perform the same substitution-trick with more efficient first-order methods than gradient descent.

We put our attention on Polyak's heavy ball method where the update rule could be written in the following form:

$$v^k = \beta v^{k-1} + \nabla f(x^k), \quad x^{k+1} = x^k - \gamma^k v^k. \tag{5.3}$$

As in STP, we substitute $\nabla f(x^k)$ by $\pm s^k$ and consider new sequences $\{v_+^k\}_{k\geq 0}$ and

$\{v_-^k\}_{k\geq 0}$ defined in the Algorithm 9. However, it is not straightforward how to choose next x^{k+1} and v^k and the virtual iterates analysis [137] hints the update rule. We consider new iterates $z_+^{k+1} = x_+^{k+1} - \frac{k}{1-}v_+^k$ and $z_-^{k+1} = x_-^{k+1} - \frac{k}{1-}v_-^k$ and define z^{k+1} as $\arg\min\left\{f(z^k), f(z_+^{k+1}), f(z_-^{k+1})\right\}$. Next we update x^{k+1} and v^k in order to have the same relationship between z^{k+1}, x^{k+1} and v^k as between z_+^{k+1}, x_+^{k+1} and v_+^k and z_-^{k+1}, x_-^{k+1} and v_-^k. Such scheme allows easily apply virtual iterates analysis and and generalize Key Lemma from [129] which is the main tool in the analysis of STP.

By definition of z^{k+1}, we get that the sequence $\{f(z^k)\}_{k\geq 0}$ is monotone:

$$f(z^{k+1}) \leq f(z^k) \qquad \forall k \geq 0. \tag{5.4}$$

Now, we establish the key result which will be used to prove the main complexity results and remaining theorems in this section.

Lemma 5.2.1. *Assume that f is L-smooth and D satisfies Assumption 5.2.1. Then for the iterates of* SMTP *the following inequalities hold:*

$$f(z^{k+1}) \leq f(z^k) - \frac{\gamma^k}{1-\beta}\langle \nabla f(z^k), s^k\rangle + \frac{L(\gamma^k)^2}{2(1-\beta)^2}\|s^k\|_2^2 \tag{5.5}$$

and

$$\mathbf{E}_{S^k \sim D}\left[f(z^{k+1})\right] \leq f(z^k) - \frac{\gamma^k \mu_D}{1-\beta}\|\nabla f(z^k)\|_D + \frac{L(\gamma^k)^2\gamma_D}{2(1-\beta)^2}. \tag{5.6}$$

5.2.1. Non-Convex Case

In this section, we show our complexity results for Algorithm 9 in the case when f is allowed to be non-convex. In particular, we show that SMTP in Algorithm 9 guarantees complexity bounds with the same order as classical bounds, i.e. $1/\sqrt{K}$ where K is the number of iterations, in the literature. We notice that query complexity (i.e. number of oracle calls) of SMTP coincides with its iteration complexity up to numerical constant factor. For clarity and completeness, proofs are left for the appendix.

Theorem 5.2.1. *Let Assumption 5.2.1 be satisfied and function f be L-smooth. Let* SMTP *with $\gamma^k \equiv \gamma > 0$ produce points $\{z^0, z^1, \ldots, z^{K-1}\}$ and $\overline{z}^K$ is chosen uniformly at random among them. Then*

$$\mathbf{E}\left[\|\nabla f(\overline{z}^K)\|_D\right] \leq \frac{(1-\beta)(f(x^0) - f(x^*))}{K\gamma\mu_D} + \frac{L\gamma\gamma_D}{2\mu_D(1-\beta)}. \tag{5.7}$$

Moreover, if we choose $\gamma = \frac{p}{K}$ *the complexity* (5.7) *reduces to*

$$\mathbf{E}\left[k \nabla f(\overline{z}^K)k_D\right] \leq p\frac{1}{K}\left(\frac{(1-\beta)(f(z^0)-f(x^*))}{\gamma_0\mu_D} + \frac{L\gamma_0\gamma_D}{2\mu_D(1-\beta)}\right). \tag{5.8}$$

Then $\gamma_0 = \sqrt{\frac{2(1-\beta)^2(f(x^0)-f(x^*))}{L\gamma_D}}$ *minimizes the right-hand side of* (5.8) *and for this choice we have*

$$\mathbf{E}\left[k \nabla f(\overline{z}^K)k_D\right] \leq \frac{\sqrt{2(f(x^0)-f(x^*))L\gamma_D}}{\mu_D}p\frac{}{K}. \tag{5.9}$$

In other words, the above theorem states that SMTP converges no worse than STP for non-convex problems to the stationary point. In the next sections we also show that theoretical convergence guarantees for SMTP are not worse than for STP for convex and strongly convex problems. However, in practice SMTP significantly outperforms STP. So, the relationship between SMTP and STP correlates with the known in the literature relationship between Polyak's heavy ball method and gradient descent.

5.2.2. Convex Case

In this section, we present our complexity results for Algorithm 9 when f is convex. In particular, we show that this method guarantees complexity bounds with the same order as classical bounds, i.e. $1/K$, in the literature. We will need the following additional assumption in the sequel.

Assumption 5.2.2. *We assume that* f *is convex, has a minimizer* x^* *and has bounded level set at* x^0*:*

$$R_0 \overset{def}{=} \max\left\{kx - x^*k_D^* \mid f(x) \leq f(x^0)\right\} < +\infty, \tag{5.10}$$

where $k\xi k_D^* \overset{def}{=} \max\{\langle \xi, x\rangle \mid kxk_D \leq 1\}$ *defines the dual norm to* $k\cdot k_D$*.*

From the above assumption and Cauchy-Schwartz inequality we get the following implication:

$$f(x) \leq f(x_0) \Rightarrow f(x) - f(x_*) \leq \langle \nabla f(x), x - x^*\rangle \leq k \nabla f(x)k_D kx - x^*k_D^* \leq R_0 k \nabla f(x)k_D,$$

which implies

$$k \nabla f(x)k_D \geq \frac{f(x) - f(x^*)}{R_0} \qquad \forall x : f(x) \leq f(x_0). \tag{5.11}$$

Theorem 5.2.2 (Constant stepsize). *Let Assumptions 5.2.1 and 5.2.2 be satisfied and f be L-smooth. If we set $\gamma^k \equiv \gamma < \frac{(1-\beta)R_0}{D}$, then for the iterates of* SMTP *method the following inequality holds:*

$$\mathbf{E}\left[f(z^k) - f(x^*)\right] \leq \left(1 - \frac{\gamma\mu_D}{(1-\beta)R_0}\right)^k \left(f(x^0) - f(x^*)\right) + \frac{L\gamma\gamma_D R_0}{2(1-\beta)\mu_D}. \tag{5.12}$$

If we choose $\gamma = \frac{"(1-\beta)D}{L D R_0}$ for some $0 < \varepsilon \leq \frac{L D R_0^2}{2D}$ and run SMTP *for $k = K$ iterations where*

$$K = \frac{1}{\varepsilon}\frac{L\gamma_D R_0^2}{\mu_D^2} \ln\left(\frac{2(f(x^0) - f(x^*))}{\varepsilon}\right), \tag{5.13}$$

then we will get $\mathbf{E}\left[f(z^K)\right] - f(x^*) \leq \varepsilon$.

In order to get rid of factor $\ln\frac{2(f(x^0)-f(x^*))}{"}$ in the complexity we consider decreasing stepsizes.

Theorem 5.2.3 (Decreasing stepsizes). *Let Assumptions 5.2.1 and 5.2.2 be satisfied and function f be L-smooth. If we set $\gamma^k = \frac{2}{k+}$, where $\alpha = \frac{D}{(1-\beta)R_0}$ and $\theta \geq \frac{2}{}$, then for the iterates of* SMTP *method the following inequality holds:*

$$\mathbf{E}\left[f(z^k)\right] - f(x^*) \leq \frac{1}{\eta k + 1}\max\left\{f(x^0) - f(x^*), \frac{2L\gamma_D}{\alpha\theta(1-\beta)^2}\right\}, \tag{5.14}$$

where $\eta \overset{def}{=} -$. Then, if we choose $\gamma^k = \frac{2}{^2k+2}$ where $\alpha = \frac{D}{(1-\beta)R_0}$ and run SMTP *for $k = K$ iterations where*

$$K = \frac{1}{\varepsilon}\cdot\frac{2R_0^2}{\mu_D^2}\max\left\{(1-\beta)^2(f(x^0) - f(x^*)), L\gamma_D\right\} - \frac{2(1-\beta)^2R_0^2}{\mu_D^2}, \qquad \varepsilon > 0, \tag{5.15}$$

we get $\mathbf{E}\left[f(z^K)\right] - f(x^*) \leq \varepsilon$.

We notice that if we choose β sufficiently close to 1, we will obtain from the formula (5.15) that $K \approx \frac{2R_0^2 L D}{" 2D}$.

5.2.3. Strongly Convex Case

In this section we present our complexity results for Algorithm 9 when f is μ-strongly convex.

Assumption 5.2.3. *We assume that f is μ-strongly convex with respect to the norm $\|\cdot\|_D^*$:*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}(\|y - x\|_D^*)^2, \quad \forall x, y \in \mathbb{R}^n. \tag{5.16}$$

It is well known that strong convexity implies

$$\|\nabla f(x)\|_D^2 \geq 2\mu\left(f(x) - f(x^*)\right). \tag{5.17}$$

Theorem 5.2.4 (Solution-dependent stepsizes). *Let Assumptions 5.2.1 and 5.2.3 be satisfied and function f be L-smooth. If we set $\gamma^k = \frac{(1-\theta_k)}{L}\|\nabla f(z^k)\|_D\sqrt{2\mu(f(z^k) - f(x^*))}$ for some $\theta_k \in (0,2)$ such that $\theta = \inf_{k\geq 0}\{2\theta_k - \gamma_D\theta_k^2\} \in (0, L/(\mu_D^2))$, then for the iterates of* SMTP*, the following inequality holds:*

$$\mathbf{E}\left[f(z^k)\right] - f(x^*) \leq \left(1 - \frac{\theta\mu_D^2\mu}{L}\right)^k \left(f(x^0) - f(x^*)\right). \tag{5.18}$$

Then, If we run SMTP *for $k = K$ iterations where*

$$K = \frac{\kappa}{\theta\mu_D^2}\ln\left(\frac{f(x^0) - f(x^*)}{\varepsilon}\right), \qquad \varepsilon > 0, \tag{5.19}$$

where $\kappa \overset{def}{=} \frac{L}{\mu}$ is the condition number of the objective, we will get $\mathbf{E}\left[f(z^K)\right] - f(x^*) \leq \varepsilon$.

Note that the previous result uses stepsizes that depends on the optimal solution $f(x^*)$ which is often not known in practice. The next theorem removes this drawback without spoiling the convergence rate. However, we need an additional assumption on the distribution D and one extra function evaluation.

Assumption 5.2.4. *We assume that for all $s \in D$ we have $\|s\|_2 = 1$.*

Theorem 5.2.5 (Solution-free stepsizes). *Let Assumptions 5.2.1, 5.2.3 and 5.2.4 be satisfied and function f be L-smooth. If additionally we compute $f(z^k + ts^k)$, set $\gamma^k = (1-\theta_k)|f(z^k+ts^k) - f(z^k)|/(Lt)$ for $t > 0$ and assume that D is such that $\mu_D^2 \leq L/\mu$, then for the iterates of* SMTP *the following inequality holds:*

$$\mathbf{E}\left[f(z^k)\right] - f(x^*) \leq \left(1 - \frac{\mu_D^2\mu}{L}\right)^k \left(f(x^0) - f(x^*)\right) + \frac{L^2t^2}{8\mu_D^2\mu}. \tag{5.20}$$

Moreover, for any $\varepsilon > 0$ if we set t such that

$$0 < t \leq \sqrt{\frac{4\varepsilon\mu_D^2\mu}{L^2}}, \tag{5.21}$$

and run SMTP *for $k = K$ iterations where*

$$K = \frac{\kappa}{\mu_D^2}\ln\left(\frac{2(f(x^0) - f(x^*))}{\varepsilon}\right), \tag{5.22}$$

where $\kappa \overset{def}{=} \frac{L}{\mu}$ is the condition number of f, we will have $\mathbf{E}\left[f(z^K)\right] - f(x^*) \leq \varepsilon$.

5.3. Stochastic Momentum Three Points with Importance Sampling (SMTP_IS)

In this section we consider another assumption, in a similar spirit to [136], on the objective.

Assumption 5.3.1 (Coordinate-wise L-smoothness). *We assume that the objective f has coordinate-wise Lipschitz gradient, with Lipschitz constants $L_1, \dots, L_n > 0$, i.e.*

$$f(x + he_i) \leq f(x) + \nabla_i f(x)h + \frac{L_i}{2}h^2, \qquad \forall x \in \mathbb{R}^n, h \in \mathbb{R}, \tag{5.23}$$

where $\nabla_i f(x)$ is i-th partial derivative of f at the point x.

For this kind of problems we modify SMTP and present STMP_IS method in Algorithm 10. In general, the idea behind methods with importance sampling and, in particular, behind SMTP_IS is to adjust probabilities of sampling in such a way that gives better convergence guarantees. In the case when f satisfies coordinate-wise L-smoothness and Lipschitz constants L_i are known it is natural to sample direction $s^k = e_i$ with probability depending on L_i (e.g. proportional to L_i). One can find more detailed discussion of the importance sampling in [134] and [135].

Now, we establish the key result which will be used to prove the main complexity results of STMP_IS.

Lemma 5.3.1. *Assume that f satisfies Assumption 5.3.1. Then for the iterates of* STMP_IS *the following inequalities hold:*

$$f(z^{k+1}) \leq f(z^k) - \frac{\gamma_i^k}{1-\beta}|\nabla_{i_k} f(z^k)| + \frac{L_{i_k}(\gamma_i^k)^2}{2(1-\beta)^2} \tag{5.24}$$

and

$$\mathbf{E}_{s^k \sim D}\left[f(z^{k+1})\right] \leq f(z^k) - \frac{1}{1-\beta}\mathbf{E}\left[\gamma_i^k |\nabla_{i_k} f(z^k)| \mid z^k\right] + \frac{1}{2(1-\beta)^2}\mathbf{E}\left[L_{i_k}(\gamma_i^k)^2 \mid z^k\right]. \tag{5.25}$$

5.3.1. Non-convex Case

Theorem 5.3.1. *Assume that f satisfies Assumption 5.3.1. Let* STMP_IS *with $\gamma_i^k = \frac{\gamma}{w_{i_k}}$ for some $\gamma > 0$ produce points $\{z^0, z^1, \dots, z^{K-1}\}$ and $\overline{z}^K$ is chosen uniformly at random*

Algorithm 10 SMTP_IS: Stochastic Momentum Three Points with Importance Sampling

Require: stepsize parameters $w_1, \dots, w_n > 0$, probabilities $p_1, \dots, p_n > 0$ summing to 1, starting point $x^0 \in \mathbb{R}^n$, $0 \le \beta < 1$ — momentum parameter

Set $v^{-1} = 0$ and $z^0 = x^0$

for $k = 0, 1, \dots$ **do**

  Select $i_k = i$ with probability $p_i > 0$

  Choose stepsize γ_i^k proportional to $\frac{1}{w_{i_k}}$

  Let $v_+^k = \beta v^{k-1} + e_{i_k}$ and $v_-^k = \beta v^{k-1} - e_{i_k}$

  Let $x_+^{k+1} = x^k - \gamma_i^k v_+^k$ and $x_-^{k+1} = x^k - \gamma_i^k v_-^k$

  Let $z_+^{k+1} = x_+^{k+1} - \frac{\gamma_i^k}{1-\beta} v_+^k$ and $z_-^{k+1} = x_-^{k+1} - \frac{\gamma_i^k}{1-\beta} v_-^k$

  Set $z^{k+1} = \arg\min \left\{ f(z^k), f(z_+^{k+1}), f(z_-^{k+1}) \right\}$

  Set $x^{k+1} = \begin{cases} x_+^{k+1}, & \text{if } z^{k+1} = z_+^{k+1} \\ x_-^{k+1}, & \text{if } z^{k+1} = z_-^{k+1} \\ x^k, & \text{if } z^{k+1} = z^k \end{cases}$ and $v^{k+1} = \begin{cases} v_+^{k+1}, & \text{if } z^{k+1} = z_+^{k+1} \\ v_-^{k+1}, & \text{if } z^{k+1} = z_-^{k+1} \\ v^k, & \text{if } z^{k+1} = z^k \end{cases}$

end for

among them. Then

$$\mathbf{E}\left[\|\nabla f(\bar{z}^K)\|_1\right] \le \frac{(1-\beta)(f(x^0) - f(x^*))}{K\gamma \min\limits_{i=1,\dots,n} \frac{p_i}{w_i}} + \frac{\gamma}{2(1-\beta) \min\limits_{i=1,\dots,n} \frac{p_i}{w_i}} \sum_{i=1}^n \frac{L_i p_i}{w_i^2}. \tag{5.26}$$

Moreover, if we choose $\gamma = \frac{\gamma_0}{K}$, *then*

$$\mathbf{E}\left[\|\nabla f(\bar{z}^K)\|_1\right] \le \frac{1}{K \min\limits_{i=1,\dots,n} \frac{p_i}{w_i}} \left(\frac{(1-\beta)(f(x^0) - f(x^*))}{\gamma_0} + \frac{\gamma_0}{2(1-\beta)} \sum_{i=1}^n \frac{L_i p_i}{w_i^2} \right). \tag{5.27}$$

Note that if we choose $\gamma_0 = \sqrt{\frac{2(1-\beta)^2 (f(x^0) - f(x^*))}{\sum_{i=1}^n \frac{L_i p_i}{w_i^2}}}$ *in order to minimize right-hand side of* (5.27), *we will get*

$$\mathbf{E}\left[\|\nabla f(\bar{z}^K)\|_1\right] \le \frac{\sqrt{2 (f(x^0) - f(x^*)) \sum_{i=1}^n \frac{L_i p_i}{w_i^2}}}{K \min\limits_{i=1,\dots,n} \frac{p_i}{w_i}}. \tag{5.28}$$

Note that for $p_i = L_i / \sum_i L_i$ *with* $w_i = L_i$ *we have that the rates improves to*

$$\mathbf{E}\left[\|\nabla f(\bar{z}^K)\|_1\right] \le \frac{\sqrt{2(f(x^0) - f(x^*)) n \sum_{i=1}^n L_i}}{K}. \tag{5.29}$$

5.3.2. Convex Case

As for SMTP to tackle convex problems by SMTP_IS we use Assumption 5.2.2 with $\|\cdot\|_D = \|\cdot\|_1$. Note that in this case $R_0 = \max\{\|x - x^*\|_1 \mid f(x) \le f(x^0)\}$.

Theorem 5.3.2 (Constant stepsize). *Let Assumptions 5.2.2 and 5.3.1 be satisfied. If we set $\gamma_i^k = \frac{\gamma}{w_{i_k}}$ such that $0 < \gamma \le \frac{(1-\beta)R_0}{\min\limits_{i=1,\dots,n} \frac{p_i}{w_i}}$, then for the iterates of* SMTP_IS *method the following inequality holds:*

$$\mathbf{E}\left[f(z^k) - f(x^*)\right] \le \left(1 - \frac{\gamma \min\limits_{i=1,\dots,n} \frac{p_i}{w_i}}{(1-\beta)R_0}\right)^k \left(f(z^0) - f(x^*)\right) + \frac{\gamma R_0}{2(1-\beta)\min\limits_{i=1,\dots,n} \frac{p_i}{w_i}} \sum_{i=1}^n \frac{L_i p_i}{w_i^2}. \tag{5.30}$$

Moreover, if we choose $\gamma = \frac{(1-\beta)\min\limits_{i=1,\dots,n} \frac{p_i}{w_i}}{R_0 \sum\limits_{i=1}^n \frac{L_i p_i}{w_i^2}}$ for some $0 < \varepsilon \le \frac{R_0^2 \sum\limits_{i=1}^n \frac{L_i p_i}{w_i^2}}{\min\limits_{i=1,\dots,n} \frac{p_i^2}{w_i^2}}$ and run SMTP_IS *for $k = K$ iterations where*

$$K = \frac{1}{\varepsilon} \frac{R_0^2 \sum\limits_{i=1}^n \frac{L_i p_i}{w_i^2}}{\min\limits_{i=1,\dots,n} \frac{p_i^2}{w_i^2}} \ln\left(\frac{2(f(x^0) - f(x^*))}{\varepsilon}\right), \tag{5.31}$$

we will get $\mathbf{E}\left[f(z^K)\right] - f(x^*) \le \varepsilon$. Moreover, for $p_i = L_i / \sum_i L_i$ with $w_i = L_i$, the rate improves to

$$K = \frac{1}{\varepsilon} R_0^2 d \sum_{i=1}^n L_i \ln\left(\frac{2(f(x^0) - f(x^*))}{\varepsilon}\right). \tag{5.32}$$

Theorem 5.3.3 (Decreasing stepsizes). *Let Assumptions 5.2.2 and 5.3.1 be satisfied. If we set $\gamma_i^k = \frac{k}{w_{i_k}}$ and $\gamma^k = \frac{2}{k+}$, where $\alpha = \frac{\min\limits_{i=1,\dots,n} \frac{p_i}{w_i}}{(1-)R_0}$ and $\theta \ge \frac{2}{}$, then for the iterates of* SMTP_IS *method the following inequality holds:*

$$\mathbf{E}\left[f(z^k)\right] - f(x^*) \le \frac{1}{\eta k + 1} \max\left\{f(x^0) - f(x^*), \frac{2}{\alpha\theta(1-\beta)^2} \sum_{i=1}^n \frac{L_i p_i}{w_i^2}\right\}, \tag{5.33}$$

where $\eta \stackrel{def}{=} -$. Moreover, if we choose $\gamma^k = \frac{2}{^2 k + 2}$ where $\alpha = \frac{\min\limits_{i=1,\dots,n} \frac{p_i}{w_i}}{(1-)R_0}$ and run SMTP_IS *for $k = K$ iterations where*

$$K = \frac{1}{\varepsilon} \cdot \frac{2R_0^2}{\min\limits_{i=1,\dots,d} \frac{p_i^2}{w_i^2}} \max\left\{(1-\beta)^2(f(x^0) - f(x^*)), \sum_{i=1}^n \frac{L_i p_i}{w_i^2}\right\} - \frac{2(1-\beta)^2 R_0^2}{\min\limits_{i=1,\dots,n} \frac{p_i^2}{w_i^2}}, \qquad \varepsilon > 0, \tag{5.34}$$

we will get $\mathbf{E}\left[f(z^K)\right] - f(x^*) \le \varepsilon$.

5.3.3. Strongly Convex Case

Theorem 5.3.4 (Solution-dependent stepsizes). *Let Assumptions 5.2.3 (with $k \cdot k_D = k \cdot k_1$) and 5.3.1 be satisfied. If we set $\gamma_i^k = \frac{(1-\ )_k \min_{i=1,\dots,n} \frac{p_i}{w_i}}{w_{i_k} \sum_{i=1}^n \frac{L_i p_i}{w_i^2}} \sqrt{2\mu(f(z^k) - f(x^*))}$ for some $\theta_k \in (0,2)$ such that $\theta = \inf_{k \geq 0}\{2\theta_k - \theta_k^2\} \in \left(0, \frac{\sum_{i=1}^n \frac{L_i p_i}{w_i^2}}{\min_{i=1,\dots,n} \frac{p_i^2}{w_i^2}}\right)$, then for the iterates of* SMTP_IS *method the following inequality holds:*

$$\mathbf{E}\left[f(z^k)\right] - f(x^*) \leq \left(1 - \frac{\theta\mu \min_{i=1,\dots,n} \frac{p_i^2}{w_i^2}}{\sum_{i=1}^n \frac{L_i p_i}{w_i^2}}\right)^k \left(f(x^0) - f(x^*)\right). \tag{5.35}$$

If we run SMTP_IS *for $k = K$ iterations where*

$$K = \frac{\sum_{i=1}^n \frac{L_i p_i}{w_i^2}}{\theta\mu \min_{i=1,\dots,n} \frac{p_i^2}{w_i^2}} \ln\left(\frac{f(x^0) - f(x^*)}{\varepsilon}\right), \qquad \varepsilon > 0, \tag{5.36}$$

we will get $\mathbf{E}\left[f(z^K)\right] - f(x^*) \leq \varepsilon$.

The previous result based on the choice of γ^k which depends on the $f(z^k) - f(x^*)$ which is often unknown in practice. The next theorem does not have this drawback and makes it possible to obtain the same rate of convergence as in the previous theorem using one extra function evaluation.

Theorem 5.3.5 (Solution-free stepsizes). *Let Assumptions 5.2.3 (with $k \cdot k_D = k \cdot k_2$) and 5.3.1 be satisfied. If additionally we compute $f(z^k + te_{i_k})$, set $\gamma_i^k = \frac{(1-\ )|f(z^k + te_{i_k}) - f(z^k)|}{L_{i_k} t}$ for $t > 0$, then for the iterates of* SMTP_IS *method the following inequality holds:*

$$\mathbf{E}\left[f(z^k)\right] - f(x^*) \leq \left(1 - \mu \min_{i=1,\dots,n} \frac{p_i}{L_i}\right)^k \left(f(x^0) - f(x^*)\right) + \frac{t^2}{8\mu \min_{i=1,\dots,n} \frac{p_i}{L_i}} \sum_{i=1}^n p_i L_i. \tag{5.37}$$

Moreover, for any $\varepsilon > 0$ if we set t such that

$$0 < t \leq \sqrt{\frac{4\varepsilon\mu \min_{i=1,\dots,n} \frac{p_i}{L_i}}{\sum_{i=1}^n p_i L_i}}, \tag{5.38}$$

and run SMTP_IS *for $k = K$ iterations where*

$$K = \frac{1}{\mu \min_{i=1,\dots,n} \frac{p_i}{L_i}} \ln\left(\frac{2(f(x^0) - f(x^*))}{\varepsilon}\right), \tag{5.39}$$

Assumptions on f	SMTP Compleixty	Theorem	Importance Sampling	SMTP_IS Complexity	Theorem
None	$\frac{r_0 n L}{\varepsilon^2}$	5.2.1	$p_i = \frac{L_i}{\sum_{i=1}^n L_i}$	$\frac{2r_0 n \sum_{i=1}^n L_i}{\varepsilon^2}$	5.3.1
Convex, $R_0 < 1$	$\frac{R_{0,\ell_2}^2 n L}{2\varepsilon} \ln\left(\frac{2r_0}{\varepsilon}\right)$	5.2.2	$p_i = \frac{L_i}{\sum_{i=1}^n L_i}$	$\frac{R_{0,\ell_1}^2 n \sum_{i=1}^n L_i}{\varepsilon} \ln\left(\frac{2r_0}{\varepsilon}\right)$	5.3.2
μ-strongly convex	$\frac{nL}{2} \ln\left(\frac{2r_0}{\varepsilon}\right)$	5.2.5	$p_i = \frac{L_i}{\sum_{i=1}^n L_i}$	$\frac{\sum_{i=1}^n L_i}{\mu} \ln\left(\frac{2r_0}{\varepsilon}\right)$	5.3.5

Table 5.2: Comparison of SMTP with $D = N\left(0, \frac{I}{n}\right)$ and SMTP_IS with $p_i = L_i / \sum_{i=1}^n L_i$. Here $r_0 = f(x^0) - f(x^*)$, R_{0,ℓ_2} corresponds to the R_0 from Assumption 5.7.3 with $\|\cdot\|_D = \|\cdot\|_2$ and R_{0,ℓ_1} corresponds to the R_0 from Assumption 5.7.3 with $\|\cdot\|_D = \|\cdot\|_1$.

we will get $\mathbf{E}\left[f(z^K)\right] - f(x^*) \le \varepsilon$. *Moreover, note that for* $p_i = L_i / \sum_i^n L_i$ *with* $w_i = L_i$, *the rate improves to*

$$K = \frac{\sum_{i=1}^n L_i}{\mu} \ln\left(\frac{2(f(x^0) - f(x^*))}{\varepsilon}\right). \tag{5.40}$$

5.4. Comparison of SMTP and SMTP_IS

Here we compare SMTP when D is normal distribution with zero mean and $\frac{I}{n}$ covariance matrix with SMTP_IS with probabilities $p_i = L_i / \sum_{i=1}^n L_i$. We choose such a distribution for SMTP since it shows the best dimension dependence among other distributions considered in Lemma 5.7.4. Note that if f satisfies Assumption 5.3.1, it is L-smooth with $L = \max_{i=1,\ldots,n} L_i$. So, we always have that $\sum_{i=1}^n L_i \le nL$. Table 5.2 summarizes complexities in this case.

We notice that for SMTP we have $\|\cdot\|_D = \|\cdot\|_2$. That is why one needs to compare SMTP with SMTP_IS accurately. At the first glance, Table 5.2 says that for non-convex and convex cases we get an extra n factor in the complexity of SMTP_IS when $L_1 = \ldots = L_n = L$. However, it is natural since we use different norms for SMTP and SMTP_IS. In the non-convex case for SMTP we give number of iterations in order to guarantee $\mathbf{E}\left[\|\nabla f(\bar{z}^K)\|_2\right] \le \varepsilon$ while for SMTP_IS we provide number of iterations in order to guarantee $\mathbf{E}\left[\|\nabla f(\bar{z}^K)\|_1\right] \le \varepsilon$. From Holder's inequality $\|\cdot\|_1 \le \sqrt{n}\|\cdot\|_2$ and, therefore, in order to have $\mathbf{E}\left[\|\nabla f(\bar{z}^K)\|_1\right] \le \varepsilon$ for SMTP we need to ensure that $\mathbf{E}\left[\|\nabla f(\bar{z}^K)\|_2\right] \le \frac{\varepsilon}{\sqrt{n}}$. That is, to guarantee $\mathbf{E}\left[\|\nabla f(\bar{z}^K)\|_1\right] \le \varepsilon$ SMTP for aforementioned distribution needs to perform $\frac{r_0 n^2 L}{\varepsilon^2}$ iterations.

Analogously, in the convex case using Cauchy-Schwartz inequality $\|\cdot\|_2 \le \sqrt{n}\|\cdot\|_1$

we have that $R_{0,\ell_2} \leq \sqrt{n} R_{0,\ell_1}$. Typically this inequality is tight and if we assume that $R_{0,\ell_1} \geq C \frac{R_{0,\ell_2}}{\sqrt{n}}$, we will get that SMTP_IS complexity is $\frac{R_{0,\ell_2}^2 \sum_{i=1}^n L_i}{\varepsilon} \ln\left(\frac{2r_0}{\varepsilon}\right)$ up to constant factor.

That is, in all cases SMTP_IS shows better complexity than SMTP up to some constant factor.

5.5. Experiments

Experimental Setup. We conduct extensive experiments on challenging non-convex problems on the continuous control task from the MuJoCO suit [101]. In particular, we address the problem of model-free control of a dynamical system. Policy gradient methods for model-free reinforcement learning algorithms provide an off-the-shelf model-free approach to learn how to control a dynamical system and are often benchmarked in a simulator. We compare our proposed momentum stochastic three points method SMTP and the momentum with importance sampling version SMTP_IS against state-of-art DFO based methods as STP_IS [136] and ARS [108]. Moreover, we also compare against classical policy gradient methods as TRPO [138] and NG [139]. We conduct experiments on several environments with varying difficulty Swimmer-v1, Hopper-v1, HalfCheetah-v1, Ant-v1, and Humanoid-v1.

Note that due to the stochastic nature of problem where f is stochastic, we use the mean of the function values of $f(x^k), f(x_+^k)$ and $f(x_-^k)$, see Algorithm 9, over K observations. Similar to the work in [136], we use $K = 2$ for Swimmer-v1, $K = 4$ for both Hopper-v1 and HalfCheetah-v1, $K = 40$ for Ant-v1 and Humanoid-v1. Similar to [136], these values were chosen based on the validation performance over the grid that is $K \in \{1, 2, 4, 8, 16\}$ for the smaller dimensional problems Swimmer-v1, Hopper-v1, HalfCheetah-v1 and $K \in \{20, 40, 80, 120\}$ for larger dimensional problems Ant-v1, and Humanoid-v1. As for the momentum term, for SMTP we set $\beta = 0.5$. For SMTP_IS, as the smoothness constants are not available for continuous control, we use the coordinate smoothness constants of a θ parameterized smooth function $\hat{f}$ (multi-layer perceptron) that estimates f. In particular, consider running any DFO for n steps; with the queried sampled $\{x_i, f(x_i)\}_{i=1}^n$, we estimate f by solving $\theta_{n+1} = \operatorname{argmin} \sum_i (f(x_i) - \hat{f}(x_i; \theta))^2$. See [136] for further implementation details as we follow the same experimental procedure. In contrast to STP_IS, our method (SMTP) does not required sampling from directions in

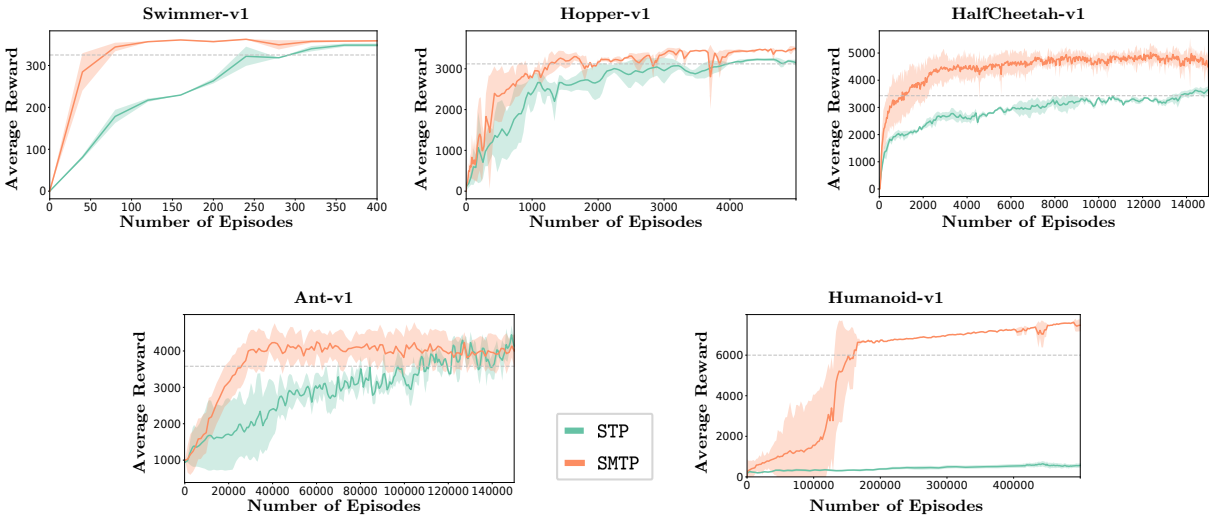

Figure 5.1: SMTP is far superior to STP on all 5 different MuJoCo tasks particularly on the high dimensional Humanoid-v1 problem. The horizontal dashed lines are the thresholds used in Table 5.3 to demonstrate complexity of each method.

the canonical basis; hence, we use directions from standard Normal distribution in each iteration. For SMTP_IS, we follow a similar procedure as [136] and sample from columns of a random matrix B.

Similar to the standard practice, we perform all experiments with 5 different initialization and measure the average reward, in continuous control we are maximizing the reward function f, and best and worst run per iteration. We compare algorithms in terms of reward vs. sample complexity.

Comparison Against STP. Our method improves sample complexity of STP and STP_IS significantly. Especially for high dimensional problems like Ant-v1 and Humanoid-v1, sample efficiency of SMTP is at least as twice as the STP. Moreover, SMTP_IS helps in some experiments by improving over SMTP. However, this is not consistent in all environments. We believe this is largely due to the fact that SMTP_IS can only handle sampling from canonical basis similar to STP_IS.

Comparison Against State-of-The-Art. We compare our method with state-of-the-art DFO and policy gradient algorithms. For the environments, Swimmer-v1, Hopper-v1, HalfCheetah-v1 and Ant-v1, our method outperforms the state-of-the-art results. Whereas for Humanoid-v1, our methods results in a comparable sample complexity.

Table 5.3: For each MuJoCo task, we report the average number of episodes required to achieve a predefined reward threshold. Results for our method is averaged over five random seeds, the rest is copied from [108] (N/A means the method failed to reach the threshold. UNK means the results is unknown since they are not reported in the literature.)

	Threshold	STP	STP_{IS}	SMTP	SMTP_{IS}	ARS(V1-t)	ARS(V2-t)	NG-lin	TRPO-nn
Swimmer-v1	325	320	110	80	100	100	427	1450	N/A
Hopper-v1	3120	3970	2400	1264	1408	51840	1973	13920	10000
HalfCheetah-v1	3430	13760	4420	1872	1624	8106	1707	11250	4250
Ant-v1	3580	107220	43860	19890	14420	58133	20800	39240	73500
Humanoid-v1	6000	N/A	530200	161230	207160	N/A	142600	130000	UNK

5.6. Conclusion

We have proposed, SMTP, the first heavy ball momentum DFO based algorithm with convergence rates for non-convex, convex and strongly convex functions under generic sampling direction. We specialize the sampling to the set of coordinate bases and further improve rates by proposing a momentum and importance sampling version SMPT_IS with new convergence rates for non-convex, convex and strongly convex functions too. We conduct large number of experiments on the task of controlling dynamical systems. We outperform two different policy gradient methods and achieve comparable or better performance to the best DFO algorithm (ARS) on the respective environments.

5.7. Missing Proofs, Technical Lemmas and Auxiliary Results

5.7.1. Preliminaries

We first list the main assumptions.

Assumption 5.7.1. *(L-smoothness) We say that f is L-smooth if:*

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \quad \forall x, y \in \mathbb{R}^n. \tag{5.41}$$

Assumption 5.7.2. *The probability distribution D on $\mathbb{R}^n$ satisfies the following properties:*

1. *The quantity $\gamma_D \overset{def}{=} \mathbf{E}_{s\ D}\|s\|_2^2$ is positive and finite.*
2. *There is a constant $\mu_D > 0$ and norm $\|\cdot\|_D$ on $\mathbb{R}^n$ such that for all $g \in \mathbb{R}^n$*

$$\mathbf{E}_{s\ D}|\langle g, s\rangle| \geq \mu_D\|g\|_D. \tag{5.42}$$

We establish the key lemma which will be used to prove the theorems stated in the paper.

Lemma 5.7.1. *Assume that f is L-smooth and D satisfies Assumption 5.7.2. Then for the iterates of* SMTP *the following inequalities hold:*

$$f(z^{k+1}) \leq f(z^k) - \frac{\gamma^k}{1-\beta}\langle \nabla f(z^k), s^k\rangle + \frac{L(\gamma^k)^2}{2(1-\beta)^2}\|s^k\|_2^2 \tag{5.43}$$

and

$$\mathbf{E}_{S^k\ D}\left[f(z^{k+1})\right] \leq f(z^k) - \frac{\gamma^k \mu_D}{1-\beta}\|\nabla f(z^k)\|_D + \frac{L(\gamma^k)^2\gamma_D}{2(1-\beta)^2}. \tag{5.44}$$

Proof. By induction one can show that

$$z^k = x^k - \frac{\gamma^k\beta}{1-\beta}v^{k-1}. \tag{5.45}$$

That is, for $k = 0$ this recurrence holds and update rules for z^k, x^k and v^{k-1} do not brake it. From this we get

$$\begin{aligned} z_+^{k+1} &= x_+^{k+1} - \frac{\gamma^k\beta}{1-\beta}v_+^k = x^k - \gamma^k v_+^k - \frac{\gamma^k\beta}{1-\beta}v_+^k \\ &= x^k - \frac{\gamma^k}{1-\beta}v_+^k = x^k - \frac{\gamma^k\beta}{1-\beta}v^{k-1} - \frac{\gamma^k}{1-\beta}s^k \\ &\overset{(5.45)}{=} z^k - \frac{\gamma^k}{1-\beta}s^k. \end{aligned}$$

Similarly,

$$\begin{aligned} z_-^{k+1} &= x_-^{k+1} - \frac{\gamma^k\beta}{1-\beta}v_-^k = x^k - \gamma^k v_-^k - \frac{\gamma^k\beta}{1-\beta}v_-^k \\ &= x^k - \frac{\gamma^k}{1-\beta}v_-^k = x^k - \frac{\gamma^k\beta}{1-\beta}v^{k-1} + \frac{\gamma^k}{1-\beta}s^k \\ &\overset{(5.45)}{=} z^k + \frac{\gamma^k}{1-\beta}s^k. \end{aligned}$$

It implies that

$$\begin{aligned} f(z_+^{k+1}) &\overset{(3.3)}{\leq} f(z^k) + \langle \nabla f(z^k), z_+^{k+1} - z_k\rangle + \frac{L}{2}\|z_+^{k+1} - z^k\|_2^2 \\ &= f(z^k) - \frac{\gamma^k}{1-\beta}\langle \nabla f(z^k), s^k\rangle + \frac{L(\gamma^k)^2}{2(1-\beta)^2}\|s^k\|_2^2 \end{aligned}$$

and

$$f(z_-^{k+1}) \leq f(z^k) + \frac{\gamma^k}{1-\beta}\langle \nabla f(z^k), s^k\rangle + \frac{L(\gamma^k)^2}{2(1-\beta)^2}\|s^k\|_2^2.$$

Unifying these two inequalities we get

$$f(z^{k+1}) \leq \min\{f(z_+^{k+1}), f(z_-^{k+1})\} = f(z^k) - \frac{\gamma^k}{1-\beta}\langle \nabla f(z^k), s^k\rangle_D + \frac{L(\gamma^k)^2}{2(1-\beta)^2}\|s^k\|_D^2,$$

which proves (5.43). Finally, taking the expectation $\mathbf{E}_{s^k\ D}$ of both sides of the previous inequality and invoking Assumption 5.7.2, we obtain

$$\mathbf{E}_{s^k\ D}\left[f(z^{k+1})\right] \leq f(z^k) - \frac{\gamma^k\mu_D}{1-\beta}\|\nabla f(z^k)\|_D + \frac{L(\gamma^k)^2\gamma_D}{2(1-\beta)^2}.$$

□

5.7.2. Missing Proofs from Section 5.2

Non-Convex Case

Theorem 5.7.1. *Let Assumptions 5.7.1 and 5.7.2 be satisfied. Let* SMTP *with* $\gamma^k \equiv \gamma > 0$ *produce points* $\{z^0, z^1, \ldots, z^{K-1}\}$ *and* $\overline{z}^K$ *is chosen uniformly at random among them. Then*

$$\mathbf{E}\left[\|\nabla f(\overline{z}^K)\|_D\right] \leq \frac{(1-\beta)(f(x^0)-f(x^*))}{K\gamma\mu_D} + \frac{L\gamma\gamma_D}{2\mu_D(1-\beta)}. \tag{5.46}$$

Moreover, if we choose $\gamma = \frac{p^0}{\overline{K}}$ *the complexity* (5.46) *reduces to*

$$\mathbf{E}\left[\|\nabla f(\overline{z}^K)\|_D\right] \leq \frac{1}{\overline{K}}\left(\frac{(1-\beta)(f(z^0)-f(x^*))}{\gamma_0\mu_D} + \frac{L\gamma_0\gamma_D}{2\mu_D(1-\beta)}\right). \tag{5.47}$$

Then $\gamma_0 = \sqrt{\frac{2(1-\beta)^2(f(x^0)-f(x^*))}{L\gamma_D}}$ *minimizes the right-hand side of* (5.47) *and for this choice we have*

$$\mathbf{E}\left[\|\nabla f(\overline{z}^K)\|_D\right] \leq \frac{\sqrt{2(f(x^0)-f(x^*))L\gamma_D}}{\mu_D}\frac{p^0}{\overline{K}}. \tag{5.48}$$

Proof. Taking full expectation from both sides of inequality (5.44) we get

$$\mathbf{E}\left[\|\nabla f(z^k)\|_D\right] \leq \frac{(1-\beta)\mathbf{E}\left[f(z^k)-f(z^{k+1})\right]}{\gamma\mu_D} + \frac{L\gamma\gamma_D}{2\mu_D(1-\beta)}.$$

Further, summing up the results for $k = 0, 1, \ldots, K-1$, dividing both sides of the obtained inequality by K and using tower property of the mathematical expectation we get

$$\mathbf{E}\left[\|\nabla f(\overline{z}^K)\|_D\right] = \frac{1}{K}\sum_{k=0}^{K-1}\mathbf{E}\left[\|\nabla f(z^k)\|_D\right] \leq \frac{(1-\beta)(f(z^0)-f(x^*))}{K\gamma\mu_D} + \frac{L\gamma\gamma_D}{2\mu_D(1-\beta)}.$$

The last part where $\gamma = \frac{p^0}{\overline{K}}$ is straightforward. □

Convex Case

Assumption 5.7.3. *We assume that f is convex, has a minimizer x^* and has bounded level set at x^0:*

$$R_0 \overset{def}{=} \max\left\{\|x - x^*\|_D^* \mid f(x) \le f(x^0)\right\} < +\infty, \tag{5.49}$$

where $\|\xi\|_D^* \overset{def}{=} \max\{\langle \xi, x\rangle \mid \|x\|_D \le 1\}$ defines the dual norm to $\|\cdot\|_D$.

Theorem 5.7.2 (Constant stepsize). *Let Assumptions 5.7.1, 5.7.2 and 5.7.3 be satisfied. If we set $\gamma^k \equiv \gamma < \frac{(1-\beta)R_0}{L_D}$, then for the iterates of* SMTP *method the following inequality holds:*

$$\mathbf{E}\left[f(z^k) - f(x^*)\right] \le \left(1 - \frac{\gamma\mu_D}{(1-\beta)R_0}\right)^k \left(f(x^0) - f(x^*)\right) + \frac{L\gamma\gamma_D R_0}{2(1-\beta)\mu_D}. \tag{5.50}$$

If we choose $\gamma = \frac{\varepsilon(1-\beta)}{L_D R_0}$ for some $0 < \varepsilon \le \frac{L_D R_0^2}{2}$ and run SMTP *for $k = K$ iterations where*

$$K = \frac{1}{\varepsilon}\frac{L\gamma_D R_0^2}{\mu_D^2} \ln\left(\frac{2(f(x^0) - f(x^*))}{\varepsilon}\right), \tag{5.51}$$

then we will get $\mathbf{E}\left[f(z^K)\right] - f(x^*) \le \varepsilon$.

Proof. From the (5.44) and monotonicity of $\{f(z^k)\}_{k\ge 0}$ we have

$$\begin{aligned}
\mathbf{E}_{s\sim D}\left[f(z^{k+1})\right] &\le f(z^k) - \frac{\gamma\mu_D}{1-\beta}\|\nabla f(z^k)\|_D^* + \frac{L\gamma^2\gamma_D}{2(1-\beta)^2} \\
&\overset{(5.11)}{\le} f(z^k) - \frac{\gamma\mu_D}{(1-\beta)R_0}(f(z^k) - f(x^*)) + \frac{L\gamma^2\gamma_D}{2(1-\beta)^2}.
\end{aligned}$$

Taking full expectation, subtracting $f(x^*)$ from the both sides of the previous inequality and using the tower property of mathematical expectation we get

$$\mathbf{E}\left[f(z^{k+1}) - f(x^*)\right] \le \left(1 - \frac{\gamma\mu_D}{(1-\beta)R_0}\right)\mathbf{E}\left[f(z^k) - f(x^*)\right] + \frac{L\gamma^2\gamma_D}{2(1-\beta)^2}. \tag{5.52}$$

Since $\gamma < \frac{(1-\beta)R_0}{L_D}$ the term $1 - \frac{\gamma\mu_D}{(1-\beta)R_0}$ is positive and we can unroll the recurrence (5.52):

$$\begin{aligned}
\mathbf{E}\left[f(z^k) - f(x^*)\right] &\le \left(1 - \frac{\gamma\mu_D}{(1-\beta)R_0}\right)^k \left(f(z^0) - f(x^*)\right) + \frac{L\gamma^2\gamma_D}{2(1-\beta)^2}\sum_{l=0}^{k-1}\left(1 - \frac{\gamma\mu_D}{(1-\beta)R_0}\right)^l \\
&\le \left(1 - \frac{\gamma\mu_D}{(1-\beta)R_0}\right)^k \left(f(x^0) - f(x^*)\right) + \frac{L\gamma^2\gamma_D}{2(1-\beta)^2}\sum_{l=0}^{1}\left(1 - \frac{\gamma\mu_D}{(1-\beta)R_0}\right)^l \\
&\le \left(1 - \frac{\gamma\mu_D}{(1-\beta)R_0}\right)^k \left(f(x^0) - f(x^*)\right) + \frac{L\gamma^2\gamma_D}{2(1-\beta)^2}\cdot\frac{(1-\beta)R_0}{\gamma\mu_D} \\
&= \left(1 - \frac{\gamma\mu_D}{(1-\beta)R_0}\right)^k \left(f(x^0) - f(x^*)\right) + \frac{L\gamma\gamma_D R_0}{2(1-\beta)\mu_D}.
\end{aligned}$$

Lastly, putting $\gamma = \frac{\varepsilon(1-\beta)\mu_D}{L\gamma_D R_0}$ and $k = K$ from (5.51) in (5.50) we have

$$\begin{aligned}\mathbf{E}[f(z^K)] - f(x^*) &= \left(1 - \frac{\varepsilon\mu_D^2}{L\gamma_D R_0^2}\right)^K \left(f(x^0) - f(x^*)\right) + \frac{\varepsilon}{2} \\ &\leq \exp\left\{-K \cdot \frac{\varepsilon\mu_D^2}{L\gamma_D R_0^2}\right\}\left(f(x^0) - f(x^*)\right) + \frac{\varepsilon}{2} \\ &\overset{(5.51)}{=} \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.\end{aligned}$$

□

Next we use technical lemma from [39]. We provide the original proof for completeness.

Lemma 5.7.2 (Lemma 6 from [39]). *Let a sequence $\{a^k\}_{k\geq 0}$ satisfy inequality $a^{k+1} \leq (1-\gamma^k\alpha)a^k + (\gamma^k)^2 N$ for any positive $\gamma^k \leq \gamma_0$ with some constants $\alpha > 0, N > 0, \gamma_0 > 0$. Further, let $\theta \geq \frac{2}{\gamma_0}$ and take C such that $N \leq \frac{\alpha}{4}C$ and $a_0 \leq C$. Then, it holds*

$$a^k \leq \frac{C}{\frac{\alpha}{2}k + 1}$$

if we set $\gamma^k = \frac{2}{\alpha k + \theta}$.

Proof. We will show the inequality for a^k by induction. Since inequality $a_0 \leq C$ is one of our assumptions, we have the initial step of the induction. To prove the inductive step, consider

$$a^{k+1} \leq (1-\gamma^k\alpha)a^k + (\gamma^k)^2 N \leq \left(1 - \frac{2\alpha}{\alpha k + \theta}\right)\frac{\theta C}{\alpha k + \theta} + \theta\alpha\frac{C}{(\alpha k + \theta)^2}.$$

To show that the right-hand side is upper bounded by $\frac{C}{\frac{\alpha}{2}(k+1)+\theta}$, one needs to have, after multiplying both sides by $(\alpha k + \theta)(\alpha k + \alpha + \theta)(\theta C)^{-1}$,

$$\left(1 - \frac{2\alpha}{\alpha k + \theta}\right)(\alpha k + \alpha + \theta) + \alpha\frac{\alpha k + \alpha + \theta}{\alpha k + \theta} \leq \alpha k + \theta,$$

which is equivalent to

$$\alpha - \alpha\frac{\alpha k + \alpha + \theta}{\alpha k + \theta} \leq 0.$$

The last inequality is trivially satisfied for all $k \geq 0$. □

Theorem 5.7.3 (Decreasing stepsizes). *Let Assumptions 5.7.1, 5.7.2 and 5.7.3 be satisfied. If we set $\gamma^k = \frac{2}{\alpha k + \theta}$, where $\alpha = \frac{\mu_D}{(1-\beta)R_0}$ and $\theta \geq \frac{2}{\gamma_0}$, then for the iterates of* SMTP *method the following inequality holds:*

$$\mathbf{E}\left[f(z^k)\right] - f(x^*) \leq \frac{1}{\eta k + 1}\max\left\{f(x^0) - f(x^*), \frac{2L\gamma_D}{\alpha\theta(1-\beta)^2}\right\}, \tag{5.53}$$

where $\eta \overset{def}{=} -$. *Then, if we choose* $\gamma^k = \frac{2}{2k+2}$ *where* $\alpha = \frac{D}{(1-)R_0}$ *and run* SMTP *for* $k = K$ *iterations where*

$$K = \frac{1}{\varepsilon} \cdot \frac{2R_0^2}{\mu_D^2} \max\left\{(1-\beta)^2(f(x^0) - f(x^*)), L\gamma_D\right\} - \frac{2(1-\beta)^2 R_0^2}{\mu_D^2}, \qquad \varepsilon > 0, \tag{5.54}$$

we get $\mathbf{E}\left[f(z^K)\right] - f(x^*) \leq \varepsilon$.

Proof. In (5.52) we proved that

$$\mathbf{E}\left[f(z^{k+1}) - f(x^*)\right] \leq \left(1 - \frac{\gamma\mu_D}{(1-\beta)R_0}\right)\mathbf{E}\left[f(z^k) - f(x^*)\right] + \frac{L\gamma^2\gamma_D}{2(1-\beta)^2}.$$

Having that, we can apply Lemma 5.7.2 to the sequence $\mathbf{E}\left[f(z^k) - f(x^*)\right]$. The constants for the lemma are: $N = \frac{L D}{2(1-)^2}$, $\alpha = \frac{D}{(1-)R_0}$ and $C = \max\left\{f(x^0) - f(x^*), \frac{2L D}{(1-)^2}\right\}$. Lastly, choosing $\gamma^k = \frac{2}{2k+2}$ is equivalent to the choice $\theta = \frac{2}{}$. In this case, we have $\alpha\theta = 2$, $C = \max\left\{f(x^0) - f(x^*), \frac{L D}{(1-)^2}\right\}$ and $\eta = - = \frac{2}{2} = \frac{2}{2(1-)^2 R_0^2}$. Putting these parameters and K from (5.54) in the (5.53) we get the result. □

Strongly Convex Case

Assumption 5.7.4. *We assume that* f *is* μ*-strongly convex with respect to the norm* $\|\cdot\|_D^*$:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}(\|y - x\|_D^*)^2, \quad \forall x, y \in \mathbb{R}^n. \tag{5.55}$$

It is well known that strong convexity implies

$$\|\nabla f(x)\|_D^2 \geq 2\mu\left(f(x) - f(x^*)\right). \tag{5.56}$$

Theorem 5.7.4 (Solution-dependent stepsizes). *Let Assumptions 5.7.1, 5.7.2 and 5.7.4 be satisfied. If we set* $\gamma^k = \frac{(1-)_k D}{L}\sqrt{2\mu(f(z^k) - f(x^*))}$ *for some* $\theta_k \in (0, 2)$ *such that* $\theta = \inf_{k\geq 0}\{2\theta_k - \gamma_D\theta_k^2\} \in \left(0, \frac{L}{2D}\right)$, *then for the iterates of* SMTP *the following inequality holds:*

$$\mathbf{E}\left[f(z^k)\right] - f(x^*) \leq \left(1 - \frac{\theta\mu_D^2\mu}{L}\right)^k \left(f(x^0) - f(x^*)\right). \tag{5.57}$$

If we run SMTP *for* $k = K$ *iterations where*

$$K = \frac{\kappa}{\theta\mu_D^2}\ln\left(\frac{f(x^0) - f(x^*)}{\varepsilon}\right), \qquad \varepsilon > 0, \tag{5.58}$$

where $\kappa \overset{def}{=} \frac{L}{}$ *is the condition number of the objective, we will get* $\mathbf{E}\left[f(z^K)\right] - f(x^*) \leq \varepsilon$.

Proof. From (5.44) and $\gamma^k = \frac{k\ D}{L}\sqrt{2\mu(f(x^k) - f(x^*))}$ we have

$$\begin{aligned}
\mathbf{E}_{S^k\ D}\left[f(z^{k+1})\right] - f(x^*) &\leq f(z^k) - f(x^*) - \frac{\gamma^k \mu_D}{1-\beta} k\, r\, f(z^k) k_D + \frac{L(\gamma^k)^2\gamma_D}{2(1-\beta)^2} \\
&\overset{(5.56)}{\leq} f(z^k) - f(x^*) - \frac{\gamma^k \mu_D}{1-\beta}\sqrt{2\mu(f(z^k) - f(x^*))} \\
&\quad + \frac{\gamma_D \theta_k^2 \mu_D^2 \mu}{L}(f(z^k) - f(x^*)) \\
&\leq f(z^k) - f(x^*) - \frac{2\theta^k \mu_D^2 \mu}{L}(f(z^k) - f(x^*)) \\
&\quad + \frac{\gamma_D \theta_k^2 \mu_D^2 \mu}{L}(f(z^k) - f(x^*)) \\
&\leq \left(1 - (2\theta_k - \gamma_D\theta_k^2)\frac{\mu_D^2\mu}{L}\right)(f(z^k) - f(x^*)).
\end{aligned}$$

Using $\theta = \inf_{k\geq 0}\{2\theta_k - \gamma_D\theta_k^2 g\ 2\ \left(0, \frac{L}{\frac{2}{D}}\right)$ and taking the full expectation from the previous inequality we get

$$\begin{aligned}
\mathbf{E}\left[f(z^{k+1}) - f(x^*)\right] &\leq \left(1 - \frac{\theta\mu_D^2\mu}{L}\right)\mathbf{E}\left[f(z^k) - f(x^*)\right] \\
&\leq \left(1 - \frac{\theta\mu_D^2\mu}{L}\right)^{k+1}\left(f(x^0) - f(x^*)\right).
\end{aligned}$$

Lastly, from (5.57) we have

$$\begin{aligned}
\mathbf{E}\left[f(z^K)\right] - f(x^*) &\leq \left(1 - \frac{\theta\mu_D^2\mu}{L}\right)^K\left(f(x^0) - f(x^*)\right) \\
&\leq \exp\left\{-K\frac{\theta\mu_D^2\mu}{L}\right\}\left(f(x^0) - f(x^*)\right) \\
&\overset{(5.58)}{\leq} \varepsilon.
\end{aligned}$$

□

Assumption 5.7.5. *We assume that for all s D we have $ksk_2 = 1$.*

Theorem 5.7.5 (Solution-free stepsizes). *Let Assumptions 5.7.1, 5.7.2, 5.7.4 and 5.7.5 be satisfied. If additionally we compute $f(z^k + ts^k)$, set $\gamma^k = \frac{(1-\ )jf(z^k+ts^k)-f(z^k)j}{Lt}$ for $t > 0$ and assume that D is such that $\mu_D^2 \leq \frac{L}{}$, then for the iterates of* SMTP *the following inequality holds:*

$$\mathbf{E}\left[f(z^k)\right] - f(x^*) \leq \left(1 - \frac{\mu_D^2\mu}{L}\right)^k\left(f(x^0) - f(x^*)\right) + \frac{L^2t^2}{8\mu_D^2\mu}. \tag{5.59}$$

Moreover, for any $\varepsilon > 0$ if we set t such that

$$0 < t \leq \sqrt{\frac{4\varepsilon\mu_D^2\mu}{L^2}}, \tag{5.60}$$

and run SMTP *for* $k = K$ *iterations where*

$$K = \frac{\kappa}{\mu_D^2} \ln\left(\frac{2(f(x^0) - f(x^*))}{\varepsilon}\right), \tag{5.61}$$

where $\kappa \overset{def}{=} \frac{L}{}$ *is the condition number of* f*, we will have* $\mathbf{E}\left[f(z^K)\right] - f(x^*) \leq \varepsilon$*.*

Proof. Recall that from (5.43) we have

$$f(z^{k+1}) \leq f(z^k) - \frac{\gamma^k}{1-\beta}\langle \nabla f(z^k), s^k\rangle + \frac{L(\gamma^k)^2}{2(1-\beta)^2}.$$

If we minimize the right hand side of the previous inequality as a function of γ^k, we will get that the optimal choice in this sense is $\gamma^k_{\text{opt}} = \frac{(1-\beta)\langle \nabla f(z^k); s^k\rangle}{L}$. However, this stepsize is impractical for derivative-free optimization, since it requires to know $\nabla f(z^k)$. The natural way to handle this is to approximate directional derivative $\langle \nabla f(z^k), s^k\rangle$ by finite difference $\frac{f(z^k+ts^k)-f(z^k)}{t}$ and that is what we do. We choose $\gamma^k = \frac{(1-\beta)[f(z^k+ts^k)-f(z^k)]}{Lt} = \frac{(1-\beta)\langle \nabla f(z^k); s^k\rangle}{L} + \frac{(1-\beta)[f(z^k+ts^k)-f(z^k)]}{Lt} - \frac{(1-\beta)\langle \nabla f(z^k); s^k\rangle}{L} \overset{\text{def}}{=} \gamma^k_{\text{opt}} + \delta^k$. From this we get

$$f(z^{k+1}) \leq f(z^k) - \frac{\langle \nabla f(z^k), s^k\rangle^2}{2L} + \frac{L}{2(1-\beta)^2}(\delta^k)^2.$$

Next we estimate $|\delta^k|$:

$$\begin{aligned}
|\delta^k| &= \frac{(1-\beta)}{Lt}\left|[f(z^k + ts^k) - f(z^k)] - \langle \nabla f(z^k), ts^k\rangle\right| \\
&\leq \frac{(1-\beta)}{Lt}\left|f(z^k + ts^k) - f(z^k) - \langle \nabla f(z^k), ts^k\rangle\right| \\
&\overset{(3.3)}{\leq} \frac{(1-\beta)}{Lt}\cdot\frac{L}{2}\|ts^k\|_2^2 = \frac{(1-\beta)t}{2}.
\end{aligned}$$

It implies that

$$\begin{aligned}
f(z^{k+1}) &\leq f(z^k) - \frac{\langle \nabla f(z^k), s^k\rangle^2}{2L} + \frac{L}{2(1-\beta)^2}\cdot\frac{(1-\beta)^2t^2}{4} \\
&= f(z^k) - \frac{\langle \nabla f(z^k), s^k\rangle^2}{2L} + \frac{Lt^2}{8}
\end{aligned}$$

and after taking full expectation from the both sides of the obtained inequality we get

$$\mathbf{E}\left[f(z^{k+1}) - f(x^*)\right] \leq \mathbf{E}\left[f(z^k) - f(x^*)\right] - \frac{1}{2L}\mathbf{E}\left[\langle \nabla f(z^k), s^k\rangle^2\right] + \frac{Lt^2}{8}.$$

Note that from the tower property of mathematical expectation and Jensen's inequality we have

$$\begin{aligned}
\mathbf{E}\left[\langle \nabla f(z^k), s^k\rangle^2\right] &= \mathbf{E}\left[\mathbf{E}_{s^k\sim D}\left[\langle \nabla f(z^k), s^k\rangle^2 \mid z^k\right]\right] \\
&\geq \mathbf{E}\left[\left(\mathbf{E}_{s^k\sim D}\left[\langle \nabla f(z^k), s^k\rangle \mid z^k\right]\right)^2\right] \\
&\overset{(5.42)}{\geq} \mathbf{E}\left[\mu_D^2\|\nabla f(z^k)\|_2^2\right] \overset{(5.56)}{\geq} 2\mu_D^2\mu\mathbf{E}\left[f(z^k) - f(x^*)\right].
\end{aligned}$$

Putting all together we get

$$\mathbf{E}\left[f(z^{k+1}) - f(x^*)\right] \leq \left(1 - \frac{\mu_D^2 \mu}{L}\right) \mathbf{E}\left[f(z^k) - f(x^*)\right] + \frac{Lt^2}{8}.$$

Due to $\mu_D^2 \leq \frac{L}{}$ we have

$$\begin{aligned}
\mathbf{E}\left[f(z^k) - f(x^*)\right] &\leq \left(1 - \frac{\mu_D^2 \mu}{L}\right)^k \left(f(x^0) - f(x^*)\right) + \frac{Lt^2}{8} \sum_{l=0}^{k-1} \left(1 - \frac{\mu_D^2 \mu}{L}\right)^l \\
&\leq \left(1 - \frac{\mu_D^2 \mu}{L}\right)^k \left(f(x^0) - f(x^*)\right) + \frac{Lt^2}{8} \sum_{l=0}^{1} \left(1 - \frac{\mu_D^2 \mu}{L}\right)^l \\
&= \left(1 - \frac{\mu_D^2 \mu}{L}\right)^k \left(f(x^0) - f(x^*)\right) + \frac{L^2 t^2}{8\mu_D^2 \mu}.
\end{aligned}$$

Lastly, from (5.59) we have

$$\begin{aligned}
\mathbf{E}\left[f(z^K)\right] - f(x^*) &\leq \left(1 - \frac{\mu_D^2 \mu}{L}\right)^K \left(f(x^0) - f(x^*)\right) + \frac{L^2 t^2}{8\mu_D^2 \mu} \\
&\overset{(5.60)}{\leq} \exp\left\{-K \frac{\mu_D^2 \mu}{L}\right\} \left(f(x^0) - f(x^*)\right) + \frac{\varepsilon}{2} \\
&\overset{(5.61)}{\leq} \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.
\end{aligned}$$

□

5.7.3. Missing Proofs from Section 5.3

Again by definition of z^{k+1} we get that the sequence $\{f(z^k)\}_{k \geq 0}$ is monotone:

$$f(z^{k+1}) \leq f(z^k) \qquad \forall k \geq 0. \tag{5.62}$$

Lemma 5.7.3. *Assume that f satisfies Assumption 5.3.1. Then for the iterates of* SMTP_IS *the following inequalities hold:*

$$f(z^{k+1}) \leq f(z^k) - \frac{\gamma_i^k}{1-\beta} \| \nabla_{i_k} f(z^k) \| + \frac{L_{i_k} (\gamma_i^k)^2}{2(1-\beta)^2} \tag{5.63}$$

and

$$\mathbf{E}_{S^k \ D}\left[f(z^{k+1})\right] \leq f(z^k) - \frac{1}{1-\beta} \mathbf{E}\left[\gamma_i^k \| \nabla_{i_k} f(z^k) \| \, | \, z^k\right] + \frac{1}{2(1-\beta)^2} \mathbf{E}\left[L_{i_k} (\gamma_i^k)^2 \, | \, z^k\right]. \tag{5.64}$$

Proof. In the similar way as in Lemma 5.7.1 one can show that

$$z^k = x^k - \frac{\gamma_i^k \beta}{1-\beta} v^{k-1} \tag{5.65}$$

and

$$z_+^{k+1} = z^k - \frac{\gamma_i^k}{1-\beta} e_{i_k},$$

$$z_-^{k+1} = z^k + \frac{\gamma_i^k}{1-\beta} e_{i_k}.$$

It implies that

$$f(z_+^{k+1}) \overset{(5.23)}{\leq} f(z^k) - \frac{\gamma_i^k}{1-\beta} \nabla_{i_k} f(z^k) + \frac{L_{i_k}(\gamma_i^k)^2}{2(1-\beta)^2}$$

and

$$f(z_-^{k+1}) \leq f(z^k) + \frac{\gamma_i^k}{1-\beta} \nabla_{i_k} f(z^k) + \frac{L_{i_k}(\gamma_i^k)^2}{2(1-\beta)^2}.$$

Unifying these two inequalities we get

$$f(z^{k+1}) \leq \min\{f(z_+^{k+1}), f(z_-^{k+1})\} = f(z^k) - \frac{\gamma_i^k}{1-\beta} |\nabla_{i_k} f(z^k)| + \frac{L_{i_k}(\gamma_i^k)^2}{2(1-\beta)^2},$$

which proves (5.63). Finally, taking the expectation $\mathbf{E}[\cdot \mid z^k]$ conditioned on z^k from the both sides of the previous inequality we obtain

$$\mathbf{E}\left[f(z^{k+1}) \mid z^k\right] \leq f(z^k) - \frac{1}{1-\beta}\mathbf{E}\left[\gamma_i^k |\nabla_{i_k} f(z^k)| \mid z^k\right] + \frac{1}{2(1-\beta)^2}\mathbf{E}\left[L_{i_k}(\gamma_i^k)^2 \mid z^k\right].$$

□

Non-convex Case

Theorem 5.7.6. *Assume that f satisfies Assumption 5.3.1. Let* SMTP_IS *with $\gamma_i^k = \frac{\gamma}{w_{i_k}}$ for some $\gamma > 0$ produce points $\{z^0, z^1, \ldots, z^{K-1}\}$ and $\overline{z}^K$ is chosen uniformly at random among them. Then*

$$\mathbf{E}\left[\|\nabla f(\overline{z}^K)\|_1\right] \leq \frac{(1-\beta)(f(x^0) - f(x^*))}{K\gamma \min\limits_{i=1,\ldots,n} \frac{p_i}{w_i}} + \frac{\gamma}{2(1-\beta)\min\limits_{i=1,\ldots,n} \frac{p_i}{w_i}} \sum_{i=1}^n \frac{L_i p_i}{w_i^2}. \tag{5.66}$$

Moreover, if we choose $\gamma = \frac{\gamma_0}{K}$, then

$$\mathbf{E}\left[\|\nabla f(\overline{z}^K)\|_1\right] \leq \frac{1}{K \min\limits_{i=1,\ldots,n} \frac{p_i}{w_i}} \left(\frac{(1-\beta)(f(x^0) - f(x^*))}{\gamma_0} + \frac{\gamma_0}{2(1-\beta)} \sum_{i=1}^n \frac{L_i p_i}{w_i^2}\right). \tag{5.67}$$

Note that if we choose $\gamma_0 = \sqrt{\frac{2(1-\ )^2(f(x^0)-f(x^*))}{\sum\limits_{i=1}^{n}\frac{L_i p_i}{w_i^2}}}$ *in order to minimize right-hand side of* (5.67)*, we will get*

$$\mathbf{E}\left[k\,r\,f(\overline{z}^K)k_1\right] \leq \frac{\sqrt{2\left(f(x^0)-f(x^*)\right)\sum\limits_{i=1}^{n}\frac{L_i p_i}{w_i^2}}}{p\overline{K}\min\limits_{i=1;:::;n}\frac{p_i}{w_i}}. \tag{5.68}$$

Note that for $p_i = {}^{L_i}/_{\sum_i L_i}$ *with* $w_i = L_i$ *we have that the rates improves to*

$$\mathbf{E}\left[k\,r\,f(\overline{z}^K)k_1\right] \leq \frac{\sqrt{2(f(x^0)-f(x^*))d\sum_{i=1}^{d}L_i}}{p\overline{K}}. \tag{5.69}$$

Proof. Recall that from (5.64) we have

$$\mathbf{E}\left[f(z^{k+1})\,j\,z^k\right] \leq f(z^k) - \frac{1}{1-\beta}\mathbf{E}\left[\gamma_i^k j\,r_{\,i_k}f(z^k)j\,j\,z^k\right] + \frac{1}{2(1-\beta)^2}\mathbf{E}\left[L_{i_k}(\gamma_i^k)^2\,j\,z^k\right]. \tag{5.70}$$

Using our choice $\gamma_i^k = \frac{}{w_{i_k}}$ we derive

$$\mathbf{E}\left[\gamma_i^k j\,r_{\,i_k}f(z^k)j\,j\,z^k\right] = \gamma\sum_{i=1}^{n}\frac{p_i}{w_i}j\,r_{\,i}f(z^k)j \geq \gamma k\,r\,f(z^k)k_1\min_{i=1;:::;n}\frac{p_i}{w_i}$$

and

$$\mathbf{E}\left[L_{i_k}(\gamma_i^k)^2\,j\,z^k\right] = \gamma^2\sum_{i=1}^{n}\frac{L_i p_i}{w_i^2}.$$

Putting it in (5.70) and taking full expectation from the both sides of obtained inequality we get

$$\mathbf{E}\left[f(z^{k+1})\right] \leq \mathbf{E}\left[f(z^k)\right] - \frac{\gamma\min\limits_{i=1;:::;n}\frac{p_i}{w_i}}{1-\beta}\mathbf{E}k\,r\,f(z^k)k_1 + \frac{\gamma^2}{2(1-\beta)^2}\sum_{i=1}^{n}\frac{L_i p_i}{w_i^2},$$

whence

$$k\,r\,f(z^k)k_1 \leq \frac{(1-\beta)\left(\mathbf{E}\left[f(z^k)\right]-\mathbf{E}\left[f(z^{k+1})\right]\right)}{\gamma\min\limits_{i=1;:::;n}\frac{p_i}{w_i}} + \frac{\gamma}{2(1-\beta)\min\limits_{i=1;:::;n}\frac{p_i}{w_i}}\sum_{i=1}^{n}\frac{L_i p_i}{w_i^2}.$$

Summing up previous inequality for $k = 0, 1, \ldots, K-1$ and dividing both sides of the result by K, we get

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbf{E}\left[k\,r\,f(z^k)k_1\right] \leq \frac{(1-\beta)(f(z^0)-f(x^*))}{K\gamma\min\limits_{i=1;:::;n}\frac{p_i}{w_i}} + \frac{\gamma}{2(1-\beta)\min\limits_{i=1;:::;n}\frac{p_i}{w_i}}\sum_{i=1}^{n}\frac{L_i p_i}{w_i^2}.$$

It remains to notice that $\frac{1}{K}\sum\limits_{k=0}^{K-1}\mathbf{E}\left[k\,r\,f(z^k)k_1\right] = \mathbf{E}\left[k\,r\,f(\overline{z}^K)k_1\right]$. The last part where $\gamma = \frac{p\,0}{K}$ is straightforward. □

Convex Case

Theorem 5.7.7 (Constant stepsize). *Let Assumptions 5.2.2 and 5.3.1 be satisfied. If we set* $\gamma_i^k = \frac{\gamma}{w_{i_k}}$ *such that* $0 < \gamma \leq \frac{(1-\beta)R_0}{\min\limits_{i=1,\dots,n} \frac{p_i}{w_i}}$, *then for the iterates of* SMTP_IS *method the following inequality holds:*

$$\mathbf{E}\left[f(z^k) - f(x^*)\right] \leq \left(1 - \frac{\gamma \min\limits_{i=1,\dots,n} \frac{p_i}{w_i}}{(1-\beta)R_0}\right)^k \left(f(z^0) - f(x^*)\right) + \frac{\gamma R_0}{2(1-\beta)\min\limits_{i=1,\dots,n} \frac{p_i}{w_i}} \sum_{i=1}^n \frac{L_i p_i}{w_i^2}. \tag{5.71}$$

Moreover, if we choose $\gamma = \frac{\varepsilon(1-\beta)\min\limits_{i=1,\dots,n} \frac{p_i}{w_i}}{R_0 \sum\limits_{i=1}^n \frac{L_i p_i}{w_i^2}}$ *for some* $0 < \varepsilon \leq \frac{R_0^2 \sum\limits_{i=1}^n \frac{L_i p_i}{w_i^2}}{\min\limits_{i=1,\dots,n} \frac{p_i^2}{w_i^2}}$ *and run* SMTP_IS *for* $k = K$ *iterations where*

$$K = \frac{1}{\varepsilon} \frac{R_0^2 \sum\limits_{i=1}^n \frac{L_i p_i}{w_i^2}}{\min\limits_{i=1,\dots,n} \frac{p_i^2}{w_i^2}} \ln\left(\frac{2(f(x^0) - f(x^*))}{\varepsilon}\right), \tag{5.72}$$

we will get $\mathbf{E}\left[f(z^K)\right] - f(x^*) \leq \varepsilon$. *Moreover, for* $p_i = L_i / \sum_i L_i$ *with* $w_i = L_i$, *the rate improves to*

$$K = \frac{1}{\varepsilon} R_0^2 n \sum_{i=1}^n L_i \ln\left(\frac{2(f(x^0) - f(x^*))}{\varepsilon}\right). \tag{5.73}$$

Proof. Recall that from (5.64) we have

$$\mathbf{E}\left[f(z^{k+1}) \,|\, z^k\right] \leq f(z^k) - \frac{1}{1-\beta}\mathbf{E}\left[\gamma_i^k \,|\, r_{i_k} f(z^k)| \,|\, z^k\right] + \frac{1}{2(1-\beta)^2}\mathbf{E}\left[L_{i_k}(\gamma_i^k)^2 \,|\, z^k\right]. \tag{5.74}$$

Using our choice $\gamma_i^k = \frac{\gamma}{w_{i_k}}$ we derive

$$\begin{aligned}\mathbf{E}\left[\gamma_i^k \, r_{i_k} f(z^k) \,|\, z^k\right] &= \gamma \sum_{i=1}^n \frac{p_i}{w_i} |r_i f(z^k)| \geq \gamma \|r f(z^k)\|_1 \min_{i=1,\dots,n} \frac{p_i}{w_i} \\ &\overset{(5.11)}{\geq} \frac{\gamma}{R_0} \min_{i=1,\dots,n} \frac{p_i}{w_i} \left(f(z^k) - f(x^*)\right)\end{aligned}$$

and

$$\mathbf{E}\left[L_{i_k}(\gamma_i^k)^2 \,|\, z^k\right] = \gamma^2 \sum_{i=1}^n \frac{L_i p_i}{w_i^2}.$$

Putting it in (5.74) and taking full expectation from the both sides of obtained inequality we get

$$\mathbf{E}\left[f(z^{k+1}) - f(x^*)\right] \leq \left(1 - \frac{\gamma \min\limits_{i=1,\dots,n} \frac{p_i}{w_i}}{(1-\beta)R_0}\right) \mathbf{E}\left[f(z^k) - f(x^*)\right] + \frac{\gamma^2}{2(1-\beta)^2} \sum_{i=1}^n \frac{L_i p_i}{w_i^2}. \tag{5.75}$$

Due to our choice of $\gamma \leq \frac{(1-\beta)R_0}{\min_{i=1,\dots,n} \frac{p_i}{w_i}}$ we have that the factor $\left(1 - \frac{\gamma}{(1-\beta)R_0} \min_{i=1,\dots,n} \frac{p_i}{w_i}\right)$ is non-negative and, therefore,

$$\begin{aligned}
\mathbf{E}\left[f(z^k) - f(x^*)\right] &\leq \left(1 - \frac{\gamma}{(1-\beta)R_0} \min_{i=1,\dots,n} \frac{p_i}{w_i}\right)^k \left(f(z^0) - f(x^*)\right) \\
&\quad + \left(\frac{\gamma^2}{2(1-\beta)^2} \sum_{i=1}^n \frac{L_i p_i}{w_i^2}\right) \sum_{l=0}^{k-1} \left(1 - \frac{\gamma}{(1-\beta)R_0} \min_{i=1,\dots,n} \frac{p_i}{w_i}\right)^l \\
&\leq \left(1 - \frac{\gamma}{(1-\beta)R_0} \min_{i=1,\dots,n} \frac{p_i}{w_i}\right)^k \left(f(z^0) - f(x^*)\right) \\
&\quad + \left(\frac{\gamma^2}{2(1-\beta)^2} \sum_{i=1}^n \frac{L_i p_i}{w_i^2}\right) \sum_{l=0}^{1} \left(1 - \frac{\gamma}{(1-\beta)R_0} \min_{i=1,\dots,n} \frac{p_i}{w_i}\right)^l \\
&\leq \left(1 - \frac{\gamma \min_{i=1,\dots,n} \frac{p_i}{w_i}}{(1-\beta)R_0}\right)^k \left(f(z^0) - f(x^*)\right) + \frac{\gamma R_0}{2(1-\beta) \min_{i=1,\dots,n} \frac{p_i}{w_i}} \sum_{i=1}^n \frac{L_i p_i}{w_i^2}.
\end{aligned}$$

Then, putting $\gamma = \frac{"(1-\beta) \min_{i=1,\dots,n} \frac{p_i}{w_i}}{R_0 \sum_{i=1}^d \frac{L_i p_i}{w_i^2}}$ and $k = K$ from (5.31) in (5.30) we have

$$\begin{aligned}
\mathbf{E}[f(z^K)] - f(x^*) &= \left(1 - \frac{\varepsilon \min_{i=1,\dots,n} \frac{p_i^2}{w_i^2}}{R_0^2 \sum_{i=1}^n \frac{L_i p_i}{w_i^2}}\right)^K \left(f(z^0) - f(x^*)\right) + \frac{\varepsilon}{2} \\
&\leq \exp\left\{-K \cdot \frac{\varepsilon \min_{i=1,\dots,n} \frac{p_i^2}{w_i^2}}{R_0^2 \sum_{i=1}^n \frac{L_i p_i}{w_i^2}}\right\} \left(f(z^0) - f(x^*)\right) + \frac{\varepsilon}{2} \\
&\overset{(5.31)}{=} \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.
\end{aligned}$$

□

Theorem 5.7.8 (Decreasing stepsizes). *Let Assumptions 5.2.2 and 5.3.1 be satisfied. If we set* $\gamma_i^k = \frac{k}{w_{i_k}}$ *and* $\gamma^k = \frac{2}{k+}$, *where* $\alpha = \frac{\min_{i=1,\dots,n} \frac{p_i}{w_i}}{(1-\beta)R_0}$ *and* $\theta \geq \frac{2}{}$, *then for the iterates of* SMTP_IS *method the following inequality holds:*

$$\mathbf{E}\left[f(z^k)\right] - f(x^*) \leq \frac{1}{\eta k + 1} \max\left\{f(x^0) - f(x^*), \frac{2}{\alpha\theta(1-\beta)^2} \sum_{i=1}^n \frac{L_i p_i}{w_i^2}\right\}, \tag{5.76}$$

where $\eta \overset{def}{=} -$. *Moreover, if we choose* $\gamma^k = \frac{2}{^2k+2}$ *where* $\alpha = \frac{\min_{i=1,\dots,n} \frac{p_i}{w_i}}{(1-\beta)R_0}$ *and run* SMTP_IS *for* $k = K$ *iterations where*

$$K = \frac{1}{\varepsilon} \cdot \frac{2R_0^2}{\min_{i=1,\dots,n} \frac{p_i^2}{w_i^2}} \max\left\{(1-\beta)^2(f(x^0) - f(x^*)), \sum_{i=1}^n \frac{L_i p_i}{w_i^2}\right\} - \frac{2(1-\beta)^2 R_0^2}{\min_{i=1,\dots,n} \frac{p_i^2}{w_i^2}}, \qquad \varepsilon > 0, \tag{5.77}$$

we will get $\mathbf{E}\left[f(z^K)\right] - f(x^*) \leq \varepsilon$.

Proof. In (5.75) we proved that

$$\mathbf{E}\left[f(z^{k+1}) - f(x^*)\right] \leq \left(1 - \frac{\gamma \min_{i=1,\dots,n} \frac{p_i}{w_i}}{(1-\beta)R_0}\right) \mathbf{E}\left[f(z^k) - f(x^*)\right] + \frac{\gamma^2}{2(1-\beta)^2} \sum_{i=1}^{n} \frac{L_i p_i}{w_i^2}.$$

Having that, we can apply Lemma 5.7.2 to the sequence $\mathbf{E}\left[f(z^k) - f(x^*)\right]$. The constants for the lemma are: $N = \frac{1}{2(1-\beta)^2} \sum_{i=1}^{n} \frac{L_i p_i}{w_i^2}$, $\alpha = \frac{\min_{i=1,\dots,n} \frac{p_i}{w_i}}{(1-\beta)R_0}$ and $C = \max\left\{f(x^0) - f(x^*), \frac{2}{(1-\beta)^2} \sum_{i=1}^{n} \frac{L_i p_i}{w_i^2}\right\}$. Lastly, note that choosing $\gamma^k = \frac{2}{\beta k+2}$ is equivalent to choice $\theta = \frac{2}{\beta}$. In this case we have $\alpha\theta = 2$ and $C = \max\left\{f(x^0) - f(x^*), \frac{1}{(1-\beta)^2} \sum_{i=1}^{n} \frac{L_i p_i}{w_i^2}\right\}$ and $\eta = \beta - \beta = \frac{\beta}{2} = \frac{\min_{i=1,\dots,n} \frac{p_i^2}{w_i^2}}{2(1-\beta)^2 R_0^2}$. Putting these parameters and K from (5.34) in the (5.33) we get the result. □

Strongly Convex Case

Theorem 5.7.9 (Solution-dependent stepsizes). *Let Assumptions 5.2.3 (with $\|\cdot\| = \|\cdot\|_D = \|\cdot\|_1$) and 5.3.1 be satisfied. If we set* $\gamma_i^k = \frac{(1-\beta)\beta^k \min_{i=1,\dots,n} \frac{p_i}{w_i}}{w_{i_k} \sum_{i=1}^{n} \frac{L_i p_i}{w_i^2}} \sqrt{2\mu(f(z^k) - f(x^*))}$ *for some* $\theta_k \in (0, 2)$ *such that* $\theta = \inf_{k \geq 0}\{2\theta_k - \theta_k^2\} \in \left(0, \frac{\sum_{i=1}^{n} \frac{L_i p_i}{w_i^2}}{\min_{i=1,\dots,n} \frac{p_i^2}{w_i^2}}\right]$, *then for the iterates of* SMTP_IS *method the following inequality holds:*

$$\mathbf{E}\left[f(z^k)\right] - f(x^*) \leq \left(1 - \frac{\theta\mu \min_{i=1,\dots,n} \frac{p_i^2}{w_i^2}}{\sum_{i=1}^{n} \frac{L_i p_i}{w_i^2}}\right)^k \left(f(x^0) - f(x^*)\right). \tag{5.78}$$

If we run SMTP_IS *for $k = K$ iterations where*

$$K = \frac{\sum_{i=1}^{n} \frac{L_i p_i}{w_i^2}}{\theta\mu \min_{i=1,\dots,n} \frac{p_i^2}{w_i^2}} \ln\left(\frac{f(x^0) - f(x^*)}{\varepsilon}\right), \qquad \varepsilon > 0, \tag{5.79}$$

we will get $\mathbf{E}\left[f(z^K)\right] - f(x^*) \leq \varepsilon$.

Proof. Recall that from (5.64) we have

$$\mathbf{E}\left[f(z^{k+1}) \mid z^k\right] \leq f(z^k) - \frac{1}{1-\beta}\mathbf{E}\left[\gamma_i^k \langle \nabla_{i_k} f(z^k), z^k \rangle \mid z^k\right] + \frac{1}{2(1-\beta)^2}\mathbf{E}\left[L_{i_k}(\gamma_i^k)^2 \mid z^k\right]. \tag{5.80}$$

Using our choice $\gamma_i^k = \frac{(1-\ )_k \min_{i=1,\dots,n} \frac{p_i}{w_i}}{w_{i_k} \sum_{i=1}^n \frac{L_i p_i}{w_i^2}} \sqrt{2\mu(f(z^k) - f(x^*))}$ we derive

$$\begin{aligned}
\mathbf{E}\left[\gamma_i^k \, r_{i_k} f(z^k) \, j \, z^k\right] &= \frac{(1-\beta)\theta_k \min_{i=1,\dots,n} \frac{p_i}{w_i}}{\sum_{i=1}^n \frac{L_i p_i}{w_i^2}} \sqrt{2\mu(f(z^k) - f(x^*))} \sum_{i=1}^n \frac{p_i}{w_i} j \, r_i f(z^k) j \\
&\geq \frac{(1-\beta)\theta_k \left(\min_{i=1,\dots,n} \frac{p_i}{w_i}\right)^2}{\sum_{i=1}^n \frac{L_i p_i}{w_i^2}} \sqrt{2\mu(f(z^k) - f(x^*))} k \, r \, f(z^k) k_1 \\
&\overset{(5.17)}{\geq} \frac{2(1-\beta)\theta_k \min_{i=1,\dots,n} \frac{p_i^2}{w_i^2}}{\sum_{i=1}^n \frac{L_i p_i}{w_i^2}} \mu(f(z^k) - f(x^*))
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{E}\left[L_{i_k} (\gamma_i^k)^2 \, j \, z^k\right] &= \frac{2(1-\beta)^2 \theta_k^2 \min_{i=1,\dots,n} \frac{p_i^2}{w_i^2}}{\left(\sum_{i=1}^n \frac{L_i p_i}{w_i^2}\right)^2} \mu(f(z^k) - f(x^*)) \sum_{i=1}^n \frac{L_i p_i}{w_i^2} \\
&= \frac{2(1-\beta)^2 \theta_k^2 \min_{i=1,\dots,n} \frac{p_i^2}{w_i^2}}{\sum_{i=1}^n \frac{L_i p_i}{w_i^2}} \mu(f(z^k) - f(x^*)).
\end{aligned}$$

Putting it in (5.80) and taking full expectation from the both sides of obtained inequality we get

$$\mathbf{E}\left[f(z^{k+1}) - f(x^*)\right] \leq \left(1 - (2\theta - \theta^2) \frac{\mu \min_{i=1,\dots,n} \frac{p_i^2}{w_i^2}}{\sum_{i=1}^n \frac{L_i p_i}{w_i^2}}\right) \mathbf{E}\left[f(z^k) - f(x^*)\right].$$

Using $\theta = \inf_{k \geq 0} f 2\theta_k - \theta_k^2 g \; 2 \left(0, \frac{\sum_{i=1}^n \frac{L_i p_i}{w_i^2}}{\min_{i=1,\dots,n} \frac{p_i^2}{w_i^2}}\right)$ we obtain

$$\begin{aligned}
\mathbf{E}\left[f(z^{k+1}) - f(x^*)\right] &\leq \left(1 - \frac{\theta\mu \min_{i=1,\dots,n} \frac{p_i^2}{w_i^2}}{\sum_{i=1}^n \frac{L_i p_i}{w_i^2}}\right) \mathbf{E}\left[f(z^k) - f(x^*)\right] \\
&\leq \left(1 - \frac{\theta\mu \min_{i=1,\dots,n} \frac{p_i^2}{w_i^2}}{\sum_{i=1}^n \frac{L_i p_i}{w_i^2}}\right)^{k+1} \left(f(x^0) - f(x^*)\right).
\end{aligned}$$

Lastly, from (5.35) we have

$$\begin{aligned}\mathbf{E}\left[f(z^K)\right] - f(x^*) &\leq \left(1 - \frac{\theta\mu \min\limits_{i=1,\dots,n} \frac{p_i^2}{w_i^2}}{\sum\limits_{i=1}^n \frac{L_i p_i}{w_i^2}}\right)^K \left(f(x^0) - f(x^*)\right) \\ &\leq \exp\left\{-K\frac{\theta\mu \min\limits_{i=1,\dots,n} \frac{p_i^2}{w_i^2}}{\sum\limits_{i=1}^n \frac{L_i p_i}{w_i^2}}\right\}\left(f(x^0) - f(x^*)\right) \\ &\overset{(5.36)}{\leq} \varepsilon.\end{aligned}$$

□

Theorem 5.7.10 (Solution-free stepsizes). *Let Assumptions 5.2.3 (with $k \cdot k_D = k \cdot k_2$) and 5.3.1 be satisfied. If additionally we compute $f(z^k + te_{i_k})$, set $\gamma_i^k = \frac{(1-\beta)|f(z^k + te_{i_k}) - f(z^k)|}{L_{i_k} t}$ for $t > 0$, then for the iterates of* SMTP_IS *method the following inequality holds:*

$$\mathbf{E}\left[f(z^k)\right] - f(x^*) \leq \left(1 - \mu \min_{i=1,\dots,n} \frac{p_i}{L_i}\right)^k \left(f(x^0) - f(x^*)\right) + \frac{t^2}{8\mu \min\limits_{i=1,\dots,n} \frac{p_i}{L_i}} \sum_{i=1}^n p_i L_i. \quad (5.81)$$

Moreover, for any $\varepsilon > 0$ if we set t such that

$$0 < t \leq \sqrt{\frac{4\varepsilon\mu \min\limits_{l=1,\dots,n} \frac{p_i}{L_i}}{\sum\limits_{i=1}^n p_i L_i}}, \quad (5.82)$$

and run SMTP_IS *for $k = K$ iterations where*

$$K = \frac{1}{\mu \min\limits_{i=1,\dots,n} \frac{p_i}{L_i}} \ln\left(\frac{2(f(x^0) - f(x^*))}{\varepsilon}\right), \quad (5.83)$$

we will get $\mathbf{E}\left[f(z^K)\right] - f(x^*) \leq \varepsilon$. Moreover, note that for $p_i = L_i / \sum_i L_i$ with $w_i = L_i$, the rate improves to

$$K = \frac{\sum_{i=1}^n L_i}{\mu} \ln\left(\frac{2(f(x^0) - f(x^*))}{\varepsilon}\right). \quad (5.84)$$

Proof. Recall that from (5.63) we have

$$f(z^{k+1}) \leq f(z^k) - \frac{\gamma_i^k}{1-\beta}|\nabla_{i_k} f(z^k)| + \frac{L_{i_k}(\gamma_i^k)^2}{2(1-\beta)^2}.$$

If we minimize the right hand side of the previous inequality as a function of γ_i^k, we will get that the optimal choice in this sense is $\gamma_{\text{opt}}^k = \frac{(1-\beta)|\nabla_{i_k} f(z^k)|}{L_{i_k}}$. However, this stepsize is impractical for derivative-free optimization, since it requires to know $\nabla_{i_k} f(z^k)$. The

natural way to handle this is to approximate directional derivative $\nabla_{i_k} f(z^k)$ by finite difference $\frac{f(z^k+te_{i_k})-f(z^k)}{t}$ and that is what we do. We choose $\gamma_i^k = \frac{(1-\beta)|f(z^k+te_{i_k})-f(z^k)|}{L_{i_k}t} = \frac{(1-\beta)|\nabla_{i_k} f(z^k)|}{L_{i_k}} + \frac{(1-\beta)|f(z^k+te_{i_k})-f(z^k)|}{L_{i_k}t} - \frac{(1-\beta)|\nabla_{i_k} f(z^k)|}{L_{i_k}} \stackrel{\text{def}}{=} \gamma_{\text{opt}}^k + \delta_i^k$. From this we get

$$f(z^{k+1}) \leq f(z^k) - \frac{|\nabla_{i_k} f(z^k)|^2}{2L_{i_k}} + \frac{L_{i_k}}{2(1-\beta)^2}(\delta_i^k)^2.$$

Next we estimate $|\delta_i^k|$:

$$\begin{aligned} |\delta_i^k| &= \frac{(1-\beta)}{L_{i_k}t}\left|f(z^k+te_{i_k}) - f(z^k) - |\nabla_{i_k} f(z^k)|t\right| \\ &\leq \frac{(1-\beta)}{L_{i_k}t}\left|f(z^k+te_{i_k}) - f(z^k) - \nabla_{i_k} f(z^k)t\right| \\ &\overset{(5.23)}{\leq} \frac{(1-\beta)}{L_{i_k}t}\cdot\frac{L_{i_k}t^2}{2} = \frac{(1-\beta)t}{2}. \end{aligned}$$

It implies that

$$\begin{aligned} f(z^{k+1}) &\leq f(z^k) - \frac{|\nabla_{i_k} f(z^k)|^2}{2L_{i_k}} + \frac{L_{i_k}}{2(1-\beta)^2}\cdot\frac{(1-\beta)^2t^2}{4} \\ &= f(z^k) - \frac{|\nabla_{i_k} f(z^k)|^2}{2L_{i_k}} + \frac{L_{i_k}t^2}{8} \end{aligned}$$

and after taking expectation $\mathbf{E}\left[\cdot \mid z^k\right]$ conditioned on z^k from the both sides of the obtained inequality we get

$$\mathbf{E}\left[f(z^{k+1}) \mid z^k\right] \leq f(z^k) - \frac{1}{2}\mathbf{E}\left[\frac{|\nabla_{i_k} f(z^k)|^2}{L_{i_k}} \mid z^k\right] + \frac{t^2}{8}\mathbf{E}\left[L_{i_k} \mid z^k\right].$$

Note that

$$\begin{aligned} \mathbf{E}\left[\frac{|\nabla_{i_k} f(z^k)|^2}{L_{i_k}} \mid z^k\right] &= \sum_{i=1}^n \frac{p_i}{L_i}|\nabla_i f(z^k)|^2 \\ &\geq \|\nabla f(z^k)\|_2^2 \min_{i=1,\ldots,n}\frac{p_i}{L_i} \\ &\overset{(5.56)}{\geq} 2\mu\left(f(z^k) - f(x^*)\right)\min_{i=1,\ldots,n}\frac{p_i}{L_i}, \end{aligned}$$

since $\|\cdot\|_D = \|\cdot\|_2$, and

$$\mathbf{E}\left[L_{i_k} \mid z^k\right] = \sum_{i=1}^n p_i L_i.$$

Putting all together we get

$$\mathbf{E}\left[f(z^{k+1}) \mid z^k\right] \leq f(z^k) - \mu\min_{i=1,\ldots,n}\frac{p_i}{L_i}\left(f(z^k) - f(x^*)\right) + \frac{t^2}{8}\sum_{i=1}^n p_i L_i.$$

Taking full expectation from the previous inequality we get

$$\mathbf{E}\left[f(z^{k+1}) - f(x^*)\right] \leq \left(1 - \mu \min_{i=1,\dots,n} \frac{p_i}{L_i}\right) \mathbf{E}\left[f(z^k) - f(x^*)\right] + \frac{t^2}{8} \sum_{i=1}^{n} p_i L_i.$$

Since $\mu \leq L_i$ for all $i = 1, \dots, n$ we have

$$\begin{aligned}
\mathbf{E}\left[f(z^k) - f(x^*)\right] &\leq \left(1 - \mu \min_{i=1,\dots,n} \frac{p_i}{L_i}\right)^k \left(f(x^0) - f(x^*)\right) \\
&\quad + \left(\frac{t^2}{8} \sum_{i=1}^{n} p_i L_i\right) \sum_{l=0}^{k-1} \left(1 - \mu \min_{i=1,\dots,n} \frac{p_i}{L_i}\right)^l \\
&\leq \left(1 - \mu \min_{i=1,\dots,n} \frac{p_i}{L_i}\right)^k \left(f(x^0) - f(x^*)\right) \\
&\quad + \left(\frac{t^2}{8} \sum_{i=1}^{n} p_i L_i\right) \sum_{l=0}^{1} \left(1 - \mu \min_{i=1,\dots,n} \frac{p_i}{L_i}\right)^l \\
&= \left(1 - \mu \min_{i=1,\dots,n} \frac{p_i}{L_i}\right)^k \left(f(x^0) - f(x^*)\right) + \frac{t^2}{8\mu \min_{i=1,\dots,n} \frac{p_i}{L_i}} \sum_{i=1}^{n} p_i L_i.
\end{aligned}$$

Lastly, from (5.37) we have

$$\begin{aligned}
\mathbf{E}\left[f(z^K)\right] - f(x^*) &\leq \left(1 - \mu \min_{i=1,\dots,n} \frac{p_i}{L_i}\right)^K \left(f(x^0) - f(x^*)\right) + \frac{t^2}{8\mu \min_{i=1,\dots,n} \frac{p_i}{L_i}} \sum_{i=1}^{n} p_i L_i \\
&\overset{(5.38)}{\leq} \exp\left\{-K\mu \min_{i=1,\dots,n} \frac{p_i}{L_i}\right\} \left(f(x^0) - f(x^*)\right) + \frac{\varepsilon}{2} \\
&\overset{(5.39)}{\leq} \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.
\end{aligned}$$

□

5.7.4. Auxiliary results

Lemma 5.7.4 (Lemma 3.4 from [129]). *Let $g \in \mathbb{R}^n$.*

1. *If D is the uniform distribution on the unit sphere in $\mathbb{R}^n$, then*

$$\gamma_D = 1 \quad \text{and} \quad \mathbf{E}_{s \sim D}\left[|\langle g, s\rangle|\right] \geq \frac{1}{\sqrt{2\pi n}} \|g\|_2. \tag{5.85}$$

   Hence, D satisfies Assumption 5.2.1 with $\gamma_D = 1$, $\|\cdot\|_D = \|\cdot\|_2$ and $\mu_D \geq \frac{1}{\sqrt{2n}}$.

2. *If D is the normal distribution with zero mean and identity over n as covariance matrix (i.e. $s \sim N(0, \frac{I}{n})$) then*

$$\gamma_D = 1 \quad \text{and} \quad \mathbf{E}_{s \sim D}\left[|\langle g, s\rangle|\right] = \frac{\sqrt{2}}{\sqrt{n\pi}} \|g\|_2. \tag{5.86}$$

   Hence, D satisfies Assumption 5.2.1 with $\gamma_D = 1$, $\|\cdot\|_D = \|\cdot\|_2$ and $\mu_D = \frac{\sqrt{2}}{\sqrt{n}}$.

3. *If D is the uniform distribution on $\{e_1, \ldots, e_n\}$, then*

$$\gamma_D = 1 \quad \text{and} \quad \mathbf{E}_{s \sim D} |\langle g, s \rangle| = \frac{1}{n} \|g\|_1. \tag{5.87}$$

Hence, D satisfies Assumption 5.2.1 with $\gamma_D = 1$, $k \cdot k_D = k \cdot k_1$ and $\mu_D = \frac{1}{n}$.

4. *If D is an arbitrary distribution on $\{e_1, \ldots, e_n\}$ given by $\mathbf{P}\{s = e_i\} = p_i > 0$, then*

$$\gamma_D = 1 \quad \text{and} \quad \mathbf{E}_{s \sim D} |\langle g, s \rangle| = \|g\|_D \stackrel{\text{def}}{=} \sum_{i=1}^{n} p_i |g_i|. \tag{5.88}$$

Hence, D satisfies Assumption 5.2.1 with $\gamma_D = 1$ and $\mu_D = 1$.

5. *If D is a distribution on $D = \{u_1, \ldots, u_n\}$ where $u_1, \ldots, u_n$ form an orthonormal basis of $\mathbb{R}^n$ and $\mathbf{P}\{s = u_i\} = p_i$, then*

$$\gamma_D = 1 \quad \text{and} \quad \mathbf{E}_{s \sim D} |\langle g, s \rangle| = \|g\|_D \stackrel{\text{def}}{=} \sum_{i=1}^{n} p_i |g_i|. \tag{5.89}$$

Hence, D satisfies Assumption 5.2.1 with $\gamma_D = 1$ and $\mu_D = 1$.

References

1. Nesterov Y. Introductory Lectures on Convex Optimization: a basic course. Kluwer Academic Publishers, Massachusetts, 2004.
2. Gorbunov E., Dvinskikh D., Gasnikov A. Optimal decentralized distributed algorithms for stochastic convex optimization // arXiv preprint arXiv:1911.07363. 2019.
3. Shalev-Shwartz S., Ben-David S. Understanding machine learning: From theory to algorithms. Cambridge university press, 2014.
4. Shapiro A., Dentcheva D., Ruszczyński A. Lectures on Stochastic Programming. Society for Industrial and Applied Mathematics, 2009. http://epubs.siam.org/doi/pdf/10.1137/1.9780898718751. URL: http://epubs.siam.org/doi/abs/10.1137/1.9780898718751.
5. Spokoiny V. et al. Parametric estimation. Finite sample theory // The Annals of Statistics. 2012. Vol. 40, no. 6. P. 2877–2909.
6. Cesa-Bianchi N., Conconi A., Gentile C. On the generalization ability of on-line learning algorithms. IEEE, 2004. Vol. 50. P. 2050–2057.
7. Shalev-Shwartz S., Shamir O., Srebro N., Sridharan K. Stochastic Convex Optimization. // COLT. 2009.
8. Feldman V., Vondrak J. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate // arXiv preprint arXiv:1902.10710. 2019.
9. Gower R. M., Loizou N., Qian X. et al. SGD: General Analysis and Improved Rates // arXiv preprint arXiv:1901.09401. 2019.
10. Nemirovski A., Juditsky A., Lan G., Shapiro A. Robust Stochastic Approximation Approach to Stochastic Programming // SIAM Journal on Optimization. 2009. Vol. 19, no. 4. P. 1574–1609. URL: https://doi.org/10.1137/070704277.
11. Nguyen L. M., Nguyen P. H., van Dijk M. et al. SGD and Hogwild! convergence without the bounded gradients assumption // arXiv preprint arXiv:1802.03801. 2018.
12. Robbins H., Monro S. A stochastic approximation method // Annals of Mathematical Statistics. 1951. Vol. 22. P. 400–407.
13. Vaswani S., Bach F., Schmidt M. Fast and Faster Convergence of SGD for Over-Parameterized Models and an Accelerated Perceptron // The 22nd International Conference on Artificial Intelligence and Statistics. 2019. P. 1195–1204.

14. Lan G. An optimal method for stochastic composite optimization // Mathematical Programming. 2012. — Jun. Vol. 133, no. 1. P. 365–397. Firs appeared in June 2008. URL: https://doi.org/10.1007/s10107-010-0434-y.
15. Dvurechensky P., Gasnikov A., Tiurin A. Randomized similar triangles method: A unifying framework for accelerated randomized optimization methods (coordinate descent, directional search, derivative-free method) // arXiv preprint arXiv:1707.08486. 2017.
16. Gasnikov A. V., Nesterov Y. E. Universal method for stochastic composite optimization problems // Computational Mathematics and Mathematical Physics. 2018. Vol. 58, no. 1. P. 48–64.
17. Nesterov Y. Lectures on convex optimization. Springer, 2018. Vol. 137.
18. Defazio A., Bach F., Lacoste-Julien S. SAGA: A Fast Incremental Gradient Method with Support for Non-strongly Convex Composite Objectives // Proceedings of the 27th International Conference on Neural Information Processing Systems. NIPS'14. Cambridge, MA, USA: MIT Press, 2014. P. 1646–1654. URL: http://dl.acm.org/citation.cfm?id=2968826.2969010.
19. Gorbunov E., Hanzely F., Richtárik P. A Unified Theory of SGD: Variance Reduction, Sampling, Quantization and Coordinate Descent // arXiv preprint arXiv:1905.11261. 2019.
20. Johnson R., Zhang T. Accelerating stochastic gradient descent using predictive variance reduction // Advances in neural information processing systems. 2013. P. 315–323.
21. Schmidt M., Le Roux N., Bach F. Minimizing finite sums with the stochastic average gradient // Mathematical Programming. 2017. Vol. 162, no. 1-2. P. 83–112.
22. Allen-Zhu Z. Katyusha: The First Direct Acceleration of Stochastic Gradient Methods // Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing. STOC 2017. New York, NY, USA: ACM, 2017. P. 1200–1205. arXiv:1603.05953. URL: http://doi.acm.org/10.1145/3055399.3055448.
23. Zhou K. Direct acceleration of SAGA using sampled negative momentum // arXiv preprint arXiv:1806.11048. 2018.
24. Zhou K., Shang F., Cheng J. A simple stochastic variance reduced algorithm with fast convergence rates // arXiv preprint arXiv:1806.11027. 2018.

25. Devolder O. Exactness, inexactness and stochasticity in first-order methods for large-scale convex optimization: Ph. D. thesis / ICTEAM and CORE, Université Catholique de Louvain. 2013.
26. Dvurechensky P., Gasnikov A. Stochastic Intermediate Gradient Method for Convex Problems with Stochastic Inexact Oracle // Journal of Optimization Theory and Applications. 2016. Vol. 171, no. 1. P. 121–145. URL: http://dx.doi.org/10.1007/s10957-016-0999-6.
27. Ghadimi S., Lan G. Stochastic first-and zeroth-order methods for nonconvex stochastic programming // SIAM Journal on Optimization. 2013. Vol. 23, no. 4. P. 2341–2368.
28. Bertsekas D. P., Tsitsiklis J. N. Parallel and distributed computation: numerical methods. Prentice hall Englewood Cliffs, NJ, 1989. Vol. 23.
29. Scaman K., Bach F., Bubeck S. et al. Optimal algorithms for smooth and strongly convex distributed optimization in networks // Proceedings of the 34th International Conference on Machine Learning-Volume 70 / JMLR. org. 2017. P. 3027–3036.
30. Khaled A., Mishchenko K., Richtárik P. Better Communication Complexity for Local SGD // arXiv preprint arXiv:1909.04746. 2019.
31. Khaled A., Mishchenko K., Richtárik P. First analysis of local gd on heterogeneous data // arXiv preprint arXiv:1909.04715. 2019.
32. Stich S. U. Local SGD converges fast and communicates little // arXiv preprint arXiv:1805.09767. 2018.
33. Yu H., Jin R., Yang S. On the linear speedup analysis of communication efficient momentum sgd for distributed non-convex optimization // arXiv preprint arXiv:1905.03817. 2019.
34. Karimireddy S. P., Rebjock Q., Stich S. U., Jaggi M. Error feedback fixes signsgd and other gradient compression schemes // arXiv preprint arXiv:1901.09847. 2019.
35. Stich S. U., Cordonnier J.-B., Jaggi M. Sparsified SGD with memory // Advances in Neural Information Processing Systems. 2018. P. 4447–4458.
36. Alistarh D., Grubic D., Li J. et al. QSGD: Communication-efficient SGD via gradient quantization and encoding // Advances in Neural Information Processing Systems. 2017. P. 1709–1720.
37. Horvath S., Ho C.-Y., Horvath L. et al. Natural Compression for Distributed Deep Learning // arXiv preprint arXiv:1905.10988. 2019.

38. Horváth S., Kovalev D., Mishchenko K. et al. Stochastic distributed learning with gradient quantization and variance reduction // arXiv preprint arXiv:1904.05115. 2019.
39. Mishchenko K., Gorbunov E., Takáč M., Richtárik P. Distributed Learning with Compressed Gradient Differences // arXiv preprint arXiv:1901.09269. 2019.
40. Wen W., Xu C., Yan F. et al. Terngrad: Ternary gradients to reduce communication in distributed deep learning // Advances in Neural Information Processing Systems. 2017. P. 1509–1519.
41. Basu D., Data D., Karakus C., Diggavi S. Qsparse-local-SGD: Distributed SGD with Quantization, Sparsification, and Local Computations // arXiv preprint arXiv:1906.02367. 2019.
42. Liu X., Li Y., Tang J., Yan M. A Double Residual Compression Algorithm for Efficient Distributed Learning // arXiv preprint arXiv:1910.07561. 2019.
43. Kibardin V. Decomposition into functions in the minimization problem // Avtomatika i Telemekhanika. 1979. no. 9. P. 66–79.
44. Rogozin A., Gasnikov A. Projected Gradient Method for Decentralized Optimization over Time-Varying Networks // arXiv preprint arXiv:1911.08527. 2019.
45. Dvinskikh D., Gasnikov A. Decentralized and Parallelized Primal and Dual Accelerated Methods for Stochastic Convex Programming Problems // arXiv preprint arXiv:1904.09015. 2019.
46. Dvinskikh D., Gorbunov E., Gasnikov A. et al. On Dual Approach for Distributed Stochastic Convex Optimization over Networks // arXiv preprint arXiv:1903.09844. 2019.
47. Gorbunov E., Dvurechensky P., Gasnikov A. An Accelerated Method for Derivative-Free Smooth Stochastic Convex Optimization // arXiv preprint arXiv:1802.09022. 2018.
48. Stonyakin F. S., Dvinskikh D., Dvurechensky P. et al. Gradient methods for problems with inexact model of the objective // International Conference on Mathematical Optimization Theory and Operations Research / Springer. 2019. P. 97–114.
49. Gasnikov A. Universal gradient descent // MIPT. 2018.
50. Nesterov Y. Primal-dual subgradient methods for convex problems // Mathematical Programming. 2009. — Aug. Vol. 120, no. 1. P. 221–259. First appeared

in 2005 as CORE discussion paper 2005/67. URL: https://doi.org/10.1007/s10107-007-0149-x.

51. Ben-Tal A., Nemirovski A. Lectures on Modern Convex Optimization. Society for Industrial and Applied Mathematics, 2001. http://epubs.siam.org/doi/pdf/10.1137/1.9780898718829. URL: http://epubs.siam.org/doi/abs/10.1137/1.9780898718829.
52. Juditsky A., Nemirovski A. First Order Methods for Non-smooth Convex Large-scale Optimization, I: General purpose methods // Optimization for Machine Learning / Ed. by S. W. Suvrit Sra, Sebastian Nowozin. Cambridge, MA: MIT Press, 2012. P. 121–184.
53. Lan G. Lectures on Optimization Methods for Machine Learning // e-print. 2019.
54. Dvurechenskii P., Dvinskikh D., Gasnikov A. et al. Decentralize and randomize: Faster algorithm for wasserstein barycenters // Advances in Neural Information Processing Systems. 2018. P. 10760–10770.
55. Hendrikx H., Bach F., Massoulié L. Accelerated decentralized optimization with local updates for smooth and strongly convex objectives // arXiv preprint arXiv:1810.02660. 2018.
56. Gasnikov A. Universal gradient descent // arXiv preprint arXiv:1711.00394. 2017.
57. Shalev-Shwartz S., Zhang T. Accelerated Proximal Stochastic Dual Coordinate Ascent for Regularized Loss Minimization // Proceedings of the 31st International Conference on Machine Learning / Ed. by E. P. Xing, T. Jebara. Vol. 32 of Proceedings of Machine Learning Research. Bejing, China: PMLR, 2014. — 22–24 Jun. P. 64–72. First appeared in arXiv:1309.2375. URL: http://proceedings.mlr.press/v32/shalev-shwartz14.html.
58. Fallah A., Gurbuzbalaban M., Ozdaglar A. et al. Robust Distributed Accelerated Stochastic Gradient Methods for Multi-Agent Networks // arXiv preprint arXiv:1910.08701. 2019.
59. Aybat N. S., Fallah A., Gurbuzbalaban M., Ozdaglar A. A universally optimal multistage accelerated stochastic gradient method // arXiv preprint arXiv:1901.08022. 2019.
60. Lan G. Gradient sliding for composite optimization // Mathematical Programming. 2016. — Sep. Vol. 159, no. 1. P. 201–235. URL: https://doi.org/10.1007/

s10107-015-0955-5.

61. Uribe C. A., Lee S., Gasnikov A., Nedić A. Optimal algorithms for distributed optimization // arXiv preprint arXiv:1712.00232. 2017.
62. Kakade S., Shalev-Shwartz S., Tewari A. On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization // Unpublished Manuscript, http://ttic. uchicago. edu/shai/papers/KakadeShalevTewari09.pdf. 2009. Vol. 2, no. 1.
63. Rockafellar R. T. Convex analysis. Princeton university press, 2015.
64. Allen-Zhu Z. How to make the gradients small stochastically: Even faster convex and nonconvex sgd // Advances in Neural Information Processing Systems. 2018. P. 1157–1167.
65. Anikin A. S., Gasnikov A. V., Dvurechensky P. E. et al. Dual approaches to the minimization of strongly convex functionals with a simple structure under affine constraints // Computational Mathematics and Mathematical Physics. 2017. — Aug. Vol. 57, no. 8. P. 1262–1276. URL: https://doi.org/10.1134/S0965542517080048.
66. Nesterov Y. How to make the gradients small // Optima. 2012. Vol. 88. P. 10–11.
67. Foster D., Sekhari A., Shamir O. et al. The Complexity of Making the Gradient Small in Stochastic Convex Optimization // arXiv preprint arXiv:1902.04686. 2019.
68. Ghadimi S., Lan G. Optimal Stochastic Approximation Algorithms for Strongly Convex Stochastic Composite Optimization I: A Generic Algorithmic Framework // SIAM Journal on Optimization. 2012. Vol. 22, no. 4. P. 1469–1492.
69. Arjevani Y., Shamir O. Communication complexity of distributed convex learning and optimization // Advances in neural information processing systems. 2015. P. 1756–1764.
70. Xu J., Tian Y., Sun Y., Scutari G. Accelerated Primal-Dual Algorithms for Distributed Smooth Convex Optimization over Networks // arXiv preprint arXiv:1910.10666. 2019.
71. Lan G., Lee S., Zhou Y. Communication-efficient algorithms for decentralized and stochastic optimization // Mathematical Programming. 2017. P. 1–48.
72. Lan G., Zhou Z. Algorithms for stochastic optimization with expectation constraints // arXiv:1604.03887. 2016.
73. Scaman K., Bach F., Bubeck S. et al. Optimal Convergence Rates for Convex

Distributed Optimization in Networks // Journal of Machine Learning Research. 2019. Vol. 20, no. 159. P. 1–31.

74. Scaman K., Bach F., Bubeck S. et al. Optimal algorithms for non-smooth distributed optimization in networks // Advances in Neural Information Processing Systems. 2018. P. 2745–2754.
75. Kulunchakov A., Mairal J. Estimate sequences for stochastic composite optimization: Variance reduction, acceleration, and robustness to noise // arXiv preprint arXiv:1901.08788. 2019.
76. Kulunchakov A., Mairal J. Estimate Sequences for Variance-Reduced Stochastic Composite Optimization // arXiv preprint arXiv:1905.02374. 2019.
77. Kulunchakov A., Mairal J. A Generic Acceleration Framework for Stochastic Composite Optimization // arXiv preprint arXiv:1906.01164. 2019.
78. Lan G., Zhou Y. Random gradient extrapolation for distributed and stochastic optimization // SIAM Journal on Optimization. 2018. Vol. 28, no. 4. P. 2753–2782.
79. Olshevsky A., Paschalidis I. C., Pu S. Asymptotic Network Independence in Distributed Optimization for Machine Learning // arXiv preprint arXiv:1906.12345. 2019.
80. Olshevsky A., Paschalidis I. C., Pu S. A Non-Asymptotic Analysis of Network Independence for Distributed Stochastic Gradient Descent // arXiv preprint arXiv:1906.02702. 2019.
81. Devolder O., Glineur F., Nesterov Y. Double smoothing technique for large-scale linearly constrained convex optimization // SIAM Journal on Optimization. 2012. Vol. 22, no. 2. P. 702–727.
82. Nesterov Y. Smooth minimization of non-smooth functions // Mathematical Programming. 2005. Vol. 103, no. 1. P. 127–152. URL: http://dx.doi.org/10.1007/s10107-004-0552-5.
83. Tang J., Egiazarian K., Golbabaee M., Davies M. The Practicality of Stochastic Optimization in Imaging Inverse Problems // arXiv preprint arXiv:1910.10100. 2019.
84. Dvinskikh D. SA vs SAA for population Wasserstein barycenter calculation // arXiv preprint arXiv:2001.07697. 2020.
85. Peyré G., Cuturi M. et al. Computational optimal transport // Foundations and Trends® in Machine Learning. 2019. Vol. 11, no. 5-6. P. 355–607.

86. Rigollet P., Weed J. Entropic optimal transport is maximum-likelihood deconvolution // Comptes Rendus Mathematique. 2018. Vol. 356, no. 11-12. P. 1228–1235.
87. Cuturi M., Peyré G. A smoothed dual approach for variational Wasserstein problems // SIAM Journal on Imaging Sciences. 2016. Vol. 9, no. 1. P. 320–343.
88. Juditsky A., Nesterov Y. Deterministic and Stochastic Primal-Dual Subgradient Algorithms for Uniformly Convex Minimization // Stochastic Systems. 2014. Vol. 4, no. 1. P. 44–80. URL: https://doi.org/10.1287/10-SSY010.
89. Gasnikov A. V., Lagunovskaya A. A., Usmanova I. N., Fedorenko F. A. Gradient-free proximal methods with inexact oracle for convex stochastic nonsmooth optimization problems on the simplex // Automation and Remote Control. 2016. — Nov. Vol. 77, no. 11. P. 2018–2034. arXiv:1412.3890. URL: http://dx.doi.org/10.1134/S0005117916110114.
90. Kroshnin A., Dvinskikh D., Dvurechensky P. et al. On the Complexity of Approximating Wasserstein Barycenter // arXiv preprint arXiv:1901.08686. 2019.
91. Guminov S., Dvurechensky P., Gasnikov A. Accelerated alternating minimization // arXiv preprint arXiv:1906.03622. 2019.
92. Jin C., Netrapalli P., Ge R. et al. A Short Note on Concentration Inequalities for Random Vectors with SubGaussian Norm // arXiv preprint arXiv:1902.03736. 2019.
93. Juditsky A., Nemirovski A. S. Large deviations of vector-valued martingales in 2-smooth normed spaces // arXiv preprint arXiv:0809.0813. 2008.
94. Chernov A., Dvurechensky P., Gasnikov A. Fast Primal-Dual Gradient Method for Strongly Convex Minimization Problems with Linear Constraints // Discrete Optimization and Operations Research: 9th International Conference, DOOR 2016, Vladivostok, Russia, September 19-23, 2016, Proceedings / Ed. by Y. Kochetov, M. Khachay, V. Beresnev et al. Springer International Publishing, 2016. P. 391–403.
95. Gasnikov A., Nesterov Y. Universal fast gradient method for stochastic composit optimization problems // arXiv:1604.05275. 2016.
96. Devolder O., Glineur F., Nesterov Y. et al. First-order methods with inexact oracle: the strongly convex case. 2013.
97. Gorbunov E., Bibi A., Sener O. et al. A stochastic derivative free optimization method with momentum // arXiv preprint arXiv:1905.13278. 2019.
98. Conn A. R., Scheinberg K., Vicente L. N. Introduction to Derivative-Free Optimiza-

tion. Philadelphia, PA, USA: SIAM, 2009.

99. Kolda T. G., Lewis R. M., Torczon V. J. Optimization by direct search: New perspectives on some classical and modern methods // SIAM Review. 2003. Vol. 45. P. 385–482.
100. Chen R. Stochastic Derivative-Free Optimization of Noisy Functions // PhD thesis at Lehigh University. 2015.
101. Todorov E., Erez T., Tassa Y. Mujoco: A physics engine for model-based control // Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on / IEEE. 2012. P. 5026–5033.
102. Marsden A. L., Feinstein J. A., Taylor C. A. A computational framework for derivative-free optimization of cardiovascular geometries // Computer Methods in Applied Mechanics and Engineering. 2008. Vol. 197. P. 1890–1905.
103. Allaire G. Shape Optimization by the Homogenization Method. New York, USA: Springer, 2001.
104. Haslinger J., Mäckinen R. Introduction to Shape Optimization: Theory, Approximation, and Computation. Philadelphia, PA, USA: SIAM, 2003.
105. Mohammadi B., Pironneau O. Applied Shape Optimization for Fluids. Clarendon Press, Oxford, 2001.
106. Marsden A. L., Wang M., Dennis J. E., Moin P. Optimal aeroacustic shape design using the surrogate management framework // Optimization and Engineering. 2004. Vol. 5. P. 235–262.
107. Marsden A. L., Wang M., Dennis J. E., Moin P. Trailing-edge noise reduction using derivative-free optimization and large-eddy simulation // Journal of Fluid Mechanics. 2007. Vol. 5. P. 235–262.
108. Mania H., Guy A., Recht B. Simple random search provides a competitive approach to reinforcement learning // arXiv preprint arXiv:1803.07055. 2018.
109. Salimans T., Ho J., Chen X. et al. Evolution strategies as a scalable alternative to reinforcement learning // arXiv preprint arXiv:1703.03864. 2017.
110. Hooke R., Jeeves T. Direct search solution of numerical and statistical problems // J. Assoc. Comput. Mach. 1961. Vol. 8. P. 212–229.
111. Su Y. W. Positive basis and a class of direct search techniques // Scientia Sinica (in Chinese). 1979. Vol. 9, no. S1. P. 53–67.

112. Torczon V. On the convergence of pattern search algorithms // SIAM Journal on optimization. 1997. Vol. 7, no. 1. P. 1–25.

113. Vicente L. N. Worst case complexity of direct search // EURO Journal on Computational Optimization. 2013. Vol. 1, no. 1-2. P. 143–153.

114. Dodangeh M., Vicente L. N. Worst case complexity of direct search under convexity // Mathematical Programming. 2016. Vol. 155, no. 1-2. P. 307–332.

115. Matyas J. Random optimization // Automation and Remote Control. 1965. Vol. 26. P. 246–253.

116. Karmanov V. G. Convergence estimates for iterative minimization methods // USSR Computational Mathematics and Mathematical Physics. 1974. Vol. 14. P. 1–13.

117. Karmanov V. G. On convergence of a random search method in convex minimization problems // Theory of Probability and its applications. 1974. Vol. 19. P. 788–794.

118. Baba N. Convergence of a Random Optimization Method for Constrained Optimization Problems // Journal of Optimization Theory and Applications. 1981. Vol. 33. P. 1–11.

119. Dorea C. Expected number of steps of a random optimization method // Journal of Optimization Theory and Applications. 1983. Vol. 39. P. 165–171.

120. Sarma M. On the convergence of the Baba and Dorea random optimization methods // Journal of Optimization Theory and Applications. 1990. Vol. 66. P. 337–343.

121. Diniz-Ehrhardt M. A., Martinez J. M., Raydan M. A derivative-free nonmonotone line-search technique for unconstrained optimization // Journal of Optimization Theory and Applications. 2008. Vol. 219. P. 383–397.

122. Stich S. U., Muller C. L., Gartner B. Optimization of Convex Functions with Random Pursuit // arXiv preprint arXiv:1111.0194. 2011.

123. Ghadimi S., Lan G., Zhang H. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization // Mathematical Programming. 2016. Vol. 155, no. 1-2. P. 267–305.

124. Gratton S., Royer C. W., Vicente L. N., Zhang Z. DIRECT SEARCH BASED ON PROBABILISTIC DESCENT // SIAM Journal on Optimization. 2015. Vol. 25, no. 3. P. 1515–1541.

125. Nesterov Y., Spokoiny V. Random Gradient-Free Minimization of Convex Functions // Foundations of Computational Mathematics. 2017. Vol. 17. P. 527–566.

126. Gorbunov E., Dvurechensky P., Gasnikov A. An Accelerated Method for Derivative-Free Smooth Stochastic Convex Optimization // arXiv preprint arXiv:1802.09022. 2018.
127. Stich S. U. Convex optimization with random pursuit: Ph. D. thesis / ETH Zurich. 2014.
128. Stich S. U. On low complexity acceleration techniques for randomized optimization // International Conference on Parallel Problem Solving from Nature / Springer. 2014. P. 130–140.
129. Bergou E. H., Gorbunov E., Richtárik P. Stochastic Three Points Method for Unconstrained Smooth Minimization // arXiv preprint arXiv:1902.03591. 2019.
130. Polyak B. T. Some methods of speeding up the convergence of iteration methods // USSR Computational Mathematics and Mathematical Physics. 1964. Vol. 4, no. 5. P. 1–17.
131. Ghadimi E., Feyzmahdavian H. R., Johansson M. Global convergence of the heavy-ball method for convex optimization // 2015 European Control Conference (ECC) / IEEE. 2015. P. 310–315.
132. Lessard L., Recht B., Packard A. Analysis and design of optimization algorithms via integral quadratic constraints // SIAM Journal on Optimization. 2016. Vol. 26, no. 1. P. 57–95.
133. Loizou N., Richtárik P. Momentum and stochastic momentum for stochastic gradient, Newton, proximal point and subspace descent methods // arXiv preprint arXiv:1712.09677. 2017.
134. Zhao P., Zhang T. Stochastic optimization with importance sampling for regularized loss minimization // international conference on machine learning. 2015. P. 1–9.
135. Richtárik P., Takáč M. On optimal probabilities in stochastic coordinate descent methods // Optimization Letters. 2016. Vol. 10, no. 6. P. 1233–1243.
136. Bibi A., Bergou E. H., Sener O. et al. Stochastic Derivative-Free Optimization Method with Importance Sampling // arXiv preprint arXiv:1902.01272. 2019.
137. Yang T., Lin Q., Li Z. Unified convergence analysis of stochastic momentum methods for convex and non-convex optimization // arXiv preprint arXiv:1604.03257. 2016.
138. Schulman J., Levine S., Abbeel P. et al. Trust region policy optimization // International Conference on Machine Learning. 2015. P. 1889–1897.

139. Rajeswaran A., Lowrey K., Todorov E. V., Kakade S. M. Towards generalization and simplicity in continuous control // Advances in Neural Information Processing Systems. 2017. P. 6550–6561.