Weierstraß-Institut
für Angewandte Analysis und Stochastik

Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

Optimal decentralized distributed algorithms for stochastic convex optimization

Eduard Gorbunov[1], Darina Dvinskikh[2], Alexander Gasnikov[1,3]

submitted: February 13, 2020

[1] Moscow Institute of Physics and Technology
Institutskiy Pereulok, 9
141701 Dolgoprudny
Moscow Region
Russian Federation
and
Institute for Information Transmission Problems RAS
Bolshoy Karetny per. 19, build.1
127051 Moscow
Russian Federation
E-Mail: eduard.gorbunov@phystech.edu
gasnikov@yandex.ru

[2] Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: darina.dvinskikh@wias-berlin.de

[3] Higher School of Economics
Myasnitskaya Ulitsa 20
101000 Moscow
Russian Federation
E-Mail: gasnikov@yandex.ru

No. 2691
Berlin 2020

2010 *Mathematics Subject Classification.* 90C25, 90C06, 90C90.

Key words and phrases. Convex optimization, stochastic optimization, primal and dual methods, distributed methods, decentralized algorithms, first-order methods, optimal complexity bounds, mini-batch.

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

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Optimal decentralized distributed algorithms for stochastic convex optimization

Eduard Gorbunov, Darina Dvinskikh, Alexander Gasnikov

Abstract

We consider stochastic convex optimization problems with affine constraints and develop several methods using either primal or dual approach to solve it. In the primal case we use special penalization technique to make the initial problem more convenient for using optimization methods. We propose algorithms to solve it based on Similar Triangles Method [25, 59] with Inexact Proximal Step for the convex smooth and strongly convex smooth objective functions and methods based on Gradient Sliding algorithm [47] to solve the same problems in the non-smooth case. We prove the convergence guarantees in smooth convex case with deterministic first-order oracle.

We propose and analyze three novel methods to handle stochastic convex optimization problems with affine constraints: `SPDSTM`, `R-RRMA-AC-SA`^{2} and `SSTM_sc`. All methods use stochastic dual oracle. `SPDSTM` is the stochastic primal-dual modification of `STM` and it is applied for the dual problem when the primal functional is strongly convex and Lipschitz continuous on some ball. We extend the result from [15] for this method to the case when only biased stochastic oracle is available. `R-RRMA-AC-SA`^{2} is an accelerated stochastic method based on the restarts of `RRMA-AC-SA`^{2} from [21] and `SSTM_sc` is just stochastic `STM` for strongly convex problems. Both methods are applied to the dual problem when the primal functional is strongly convex, smooth and Lipschitz continuous on some ball and use stochastic dual first-order oracle. We develop convergence analysis for these methods for the unbiased and biased oracles respectively.

Finally, we apply all aforementioned results and approaches to solve decentralized distributed optimization problem and discuss optimality of the obtained results in terms of communication rounds and number of oracle calls per node.

1 Introduction

In this paper we are interested in the convex optimization problem

$$f(x) \to \min_{x \in Q \subseteq \mathbb{R}^n}, \tag{1}$$

where f is a convex function and Q is closed and convex subset of $\mathbb{R}^n$. More precisely, we study particular case of (1) when the objective function f could be represented as a mathematical expectation

$$f(x) = \mathbb{E}_\xi \left[f(x, \xi) \right], \tag{2}$$

where ξ is a random variable. Problems of this type play central role in a bunch of applications of machine learning [70, 72] and mathematical statistics [73]. 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 (2) is to solve empirical risk minimization or finite-sum minimization

DOI 10.20347/WIAS.PREPRINT.2691 Berlin 2020

problem instead, i.e. solve (1) with the objective

$$f(x) = \frac{1}{m}\sum_{i=1}^{m} f(x, \xi_i), \tag{3}$$

where m should be sufficiently large to approximate the initial problem (see Section 3 for the details).

Stochastic first-order methods such as Stochastic Gradient Descent (SGD) [30, 54, 60, 63, 79] or its accelerated variants like AC−SA [47] or Similar Triangles Method (STM) [19, 25, 59] are very popular choice to solve either (1)+(2) or (1)+(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 [11, 29, 35, 69] and its accelerated variants [2, 83, 84]. Unfortunately, this technique is not applicable in general for the problems of type (1)+(2) and 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 [12, 18, 25, 27] and it is one of the examples where centralized distributed optimization appears naturally [9].

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. Then, we can rewrite the objective function in the following form

$$f(x) = \frac{1}{q}\sum_{i=1}^{q} f_i(x), \quad f_i(x) = \mathbb{E}_{\xi_i}\left[f(x, \xi_i)\right] \text{ or } f_i(x) = \frac{1}{s_i}\sum_{j=1}^{s_i} f(x, \xi_{ij}). \tag{4}$$

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 [66]. The big line of works is aimed to solve this issue via periodical synchronization [40, 41, 74, 82], error-compensation [39, 75], quantization [1, 32, 33, 53, 80] or combination of these techniques [7, 52].

However, in this paper we mainly focus on another approach to deal with aforementioned drawbacks — decentralized distributed optimization [9, 42]. 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 [65].

1.1 Contributions

One can consider this paper as a continuation of work [15] where authors mentioned the key ideas that form a basis of this work. However, in this paper we provide formal proofs of some results announced in [15] 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 [16] 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 analyses, but we handle it via advanced technique for estimating recurrences (see also [16, 28]).

- **Two accelerated methods with biased 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 [16], but in this paper 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 paper there were no analysis for STM in this case. The closest work to ours in this topic is [76], but in [76] authors considered optimization problems on bounded sets.

1.2 Outline of the Paper

After introducing main notation and definitions in Section 2 we provide a short overview of the state-of-the-art results for the problem (1)+(2) that use deterministic and stochastic first-order oracles. After that, we focus on the stochastic optimization problems with affine constraints and present the state-of-the-art methods that solves specially penalized unconstrained problem instead of the original one in Section 4 together with the novel approach which we call STP_IPS that aims to solve convex smooth unconstrained minimization problems with smooth convex composite term and inexact proximal step. Next, we consider the same type of problems but using dual approach and develop three different accelerated methods for this case together with the convergence analysis for each of them. The first one is Stochastic Primal-Dual STM (SPDSTM) and it uses biased stochastic dual oracle to solve primal and dual problems simultaneously for the case when the primal problem is μ-strongly convex and Lipschitz continuous on some ball centered at zero. Next two methods are R-RRMA-AC-SA^2 and SSTM_sc and they solve the same problem when the primal functional is additionally L-smooth using stochastic dual oracle. The difference between them is that R-RRMA-AC-SA^2 uses tricky restarts technique and works with unbiased stochastic oracle, while SSTM_sc is directly accelerated and able to work with biased stochastic gradients. Then we show how to apply established in the previous sections results to the decentralized distributed optimization problems and derive the bounds for the proposed methods in Section 6. Finally, in Section 7 we compare bounds for the convergence rate in parallel and decentralized optimization, discuss the optimality of the obtained results and present possible directions for the future work. We leave long proofs, auxiliary and technical results and the whole section about STP_IPS in the appendix.

2 Notation 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 $\|x\|_2 \stackrel{\text{def}}{=} \sqrt{\langle x, x\rangle}$. 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. 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. Conditional mathematical expectation with respect to all randomness coming from random variable ξ is denoted in our paper by $\mathbb{E}_\xi[\cdot]$. We use $B_r(y) \subseteq \mathbb{R}^n$ to denote Euclidean ball centered at $y \in \mathbb{R}^n$ with radius r: $B_r(y) \stackrel{\text{def}}{=} \{x \in \mathbb{R}^n \mid \|x - y\|_2 \le r\}$. 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

$$C = \begin{bmatrix} A_{11}B & A_{12}B & A_{13}B & \ldots & A_{1m}B \\ A_{21}B & A_{22}B & A_{23}B & \ldots & A_{2m}B \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ A_{m1}B & A_{m2}B & A_{m3}B & \ldots & A_{mm}B \end{bmatrix}. \tag{5}$$

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, [55] for the details).

Definition 1 (L-smoothness)**.** Function f is called L-smooth in $Q \subseteq \mathbb{R}^n$ with $L > 0$ when it is differentiable and its gradient is L-Lipschitz continuous in Q, i.e.

$$\|\nabla f(x) - \nabla f(y)\|_2 \le L\|x - y\|_2, \quad \forall x, y \in Q. \tag{6}$$

Definition 2 (μ-strong convexity)**.** Differentiable function f is called μ-strongly convex in $Q \subseteq \mathbb{R}^n$ with $\mu \ge 0$ if

$$f(x) \ge f(y) + \langle \nabla f(y), x - y\rangle + \frac{\mu}{2}\|x - y\|_2^2, \quad \forall x, y \in Q. \tag{7}$$

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^*.

3 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 (1)+(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

Assumptions on f	Method	Citation	# of oracle calls
μ-strongly convex, L-smooth	R-STM	[25, 59]	$O\left(\sqrt{\frac{L}{\mu}}\ln\left(\frac{\mu R^2}{\varepsilon}\right)\right)$
L-smooth	STM	[25, 59]	$O\left(\sqrt{\frac{LR^2}{\varepsilon}}\right)$
μ-strongly convex, $\|\nabla f(x)\|_2 \leq M$	MD	[8, 36]	$O\left(\frac{M^2}{\mu\varepsilon}\right)$
$\|\nabla f(x)\|_2 \leq M$	MD	[8, 36]	$O\left(\frac{M^2R^2}{\varepsilon^2}\right)$

Table 1: Optimal number N of deterministic first-order oracle calls in order to get such a point x^N that $f(x^N) - f(x^*) \leq \varepsilon$. First column contains assumptions on f in addition to the convexity. MD states for Mirror Descent.

arbitrary $x \in Q$ one can get such stochastic gradient $\nabla f(x,\xi)$ that

$$\|\mathbb{E}_\xi\left[\nabla f(x,\xi)\right] - \nabla f(x)\|_2 \leq \delta, \tag{8}$$

$$\mathbb{E}_\xi\left[\exp\left(\frac{\|\nabla f(x,\xi) - \mathbb{E}_\xi\left[\nabla f(x,\xi)\right]\|_2^2}{\sigma^2}\right)\right] \leq \exp(1), \tag{9}$$

where $\delta \geq 0$ and $\sigma \geq 0$. If $\sigma = 0$, let us suppose that $\nabla f(x,\xi) = \mathbb{E}_\xi\left[\nabla f(x,\xi)\right]$ 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 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 [23, 57, 59].

In this paper 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 (1) with μ-strongly convex and L-smooth

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

Input: $\tilde{x}^0 = z^0 = x^0$, number of iterations N, $\alpha_0 = A_0 = 0$

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

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

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

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

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

6: **end for**

Output: x^N

objective, it will return x^N such that $f(x^N) - f(x^*) \leq \varepsilon$ after $N = O\left(\sqrt{{}^{L}/_{\mu}}\ln\left({}^{LR^2}/_{\varepsilon}\right)\right)$ iterations which is not optimal, see Table 1. To match the optimal bound in this case one should use classical restart of STM which is run with $\mu = 0$ [25].

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

$$f(x) + h(x) \to \min_{x \in Q}, \tag{10}$$

where h is a convex proximable function. For this case STM can be generalized via modifying the update rule in the following way [25, 59]:

$$z^{k+1} = \operatorname*{argmin}_{z \in Q} \left\{ \frac{1}{2}\|z - z^0\|_2^2 + \sum_{l=0}^{k+1} \alpha_l \left(\left\langle \nabla f(\tilde{x}^l), z - \tilde{x}^l \right\rangle + h(z) + \frac{\mu}{2}\|z - \tilde{x}^l\|_2^2 \right) \right\}. \quad (11)$$

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

Next, we go back to the problem (1)+(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 $\nabla f(x, \xi_i)$:

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

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

$$N = O\left(\min\left\{\sqrt{\frac{LR^2}{\varepsilon}}, \sqrt{\frac{L}{\mu}}\ln\left(\frac{LR^2}{\varepsilon}\right)\right\}\right) \quad (12)$$

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

$$r_{k+1} = \Theta\left(\max\left\{1, \frac{\sigma^2 \alpha_{k+1} \ln\frac{N}{\beta}}{(1 + A_{k+1}\mu)\varepsilon}\right\}\right). \quad (13)$$

The total number of oracle calls is

$$\sum_{k=1}^N r_k = O\left(N + \min\left\{\frac{\sigma^2 R^2}{\varepsilon^2}\ln\left(\frac{\sqrt{LR^2/\varepsilon}}{\beta}\right), \frac{\sigma^2}{\mu\varepsilon}\ln\left(\frac{LR^2}{\varepsilon}\right)\ln\left(\frac{\sqrt{L/\mu}}{\beta}\right)\right\}\right) \quad (14)$$

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 2.

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:

$$f(x) \to \min_{Ax=0, x \in Q}, \quad (15)$$

Assumptions on f	Method	Citation	# of iterations	# of oracle calls
μ-strongly convex, L-smooth	R-SSTM	[25, 46, 59]	$O\left(\sqrt{\frac{L}{\mu}}\ln\left(\frac{\mu R^2}{\varepsilon}\right)\right)$	$\widetilde{O}\left(\max\left\{\sqrt{\frac{L}{\mu}}\ln\left(\frac{\mu R^2}{\varepsilon}\right), \frac{\sigma^2}{\mu\varepsilon}\right\}\right)$
L-smooth	SSTM	[25, 46, 59]	$O\left(\sqrt{\frac{LR^2}{\varepsilon}}\right)$	$\widetilde{O}\left(\max\left\{\sqrt{\frac{LR^2}{\varepsilon}}, \frac{\sigma^2R^2}{\varepsilon^2}\right\}\right)$
μ-strongly convex, $\mathbb{E}_\xi\left[\|\nabla f(x,\xi)\|_2^2\right] \leq M^2$	MD	[8, 36]	$O\left(\frac{M^2}{\mu\varepsilon}\right)$	$O\left(\frac{M^2}{\mu\varepsilon}\right)$
$\mathbb{E}_\xi\left[\|\nabla f(x,\xi)\|_2^2\right] \leq M^2$	MD	[8, 36]	$O\left(\frac{M^2R^2}{\varepsilon^2}\right)$	$O\left(\frac{M^2R^2}{\varepsilon^2}\right)$

Table 2: Optimal (up to logarithmic factors) number of iterations and stochastic unbiased first-order oracle calls in order to get such a point x^N that $f(x^N) - f(x^*) \leq \varepsilon$ with probability at least $1-\beta$, $\beta \in (0,1)$. First column contains assumptions on f in addition to the convexity.

where $A \succeq 0$ and Ker$A \neq \{0\}$. Up to a sign we can define the dual problem in the following way

$$\begin{aligned}
\psi(y) &\to \min_y, \quad \text{where} && (16)\\
\varphi(y) &= \max_{x\in Q}\{\langle y, x\rangle - f(x)\}, && (17)\\
\psi(y) &= \varphi(A^\top y) = \max_{x\in Q}\{\langle y, Ax\rangle - f(x)\} = \langle y, Ax(A^\top y)\rangle - f(x(A^\top y)) && (18)\\
&= \langle A^\top y, x(A^\top y)\rangle - f(x(A^\top y)),
\end{aligned}$$

where $x(y) \stackrel{\text{def}}{=} \operatorname{argmax}_{x\in Q}\{\langle y, x\rangle - f(x)\}$. Since Ker$A \neq \{0\}$ the solution of the dual problem is not unique (16). We use y^* to denote the solution of (16) with the smallest ℓ_2-norm. This norm $R_y \stackrel{\text{def}}{=} \|y^*\|_2$ can be bounded as follows [49]:

$$R_y^2 \leq \frac{\|\nabla f(x^*)\|_2^2}{\lambda_{\min}^+(A^\top A)}. \tag{19}$$

The following lemma provides one of the key relations that we use in our analysis.

Lemma 1. Consider the function $f(x)$ defined on a closed convex set $Q \subseteq R^n$ and linear operator A such that 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

$$\psi(y^*) = -f(x^*) \geq \langle y^*, A\hat{x}\rangle - f(\hat{x}) \quad \forall \hat{x} \in Q. \tag{20}$$

However, in this section we are interested only in primal approaches to solve (15) 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 (15) is chosen as a square root of Laplacian matrix W of communication network [66] (see Section 6 for the details). In asynchronous case the square root $\sqrt{W}$ is replaced by incidence matrix M [31] ($W = M^\top M$). Then in asynchronized case instead of accelerate methods for (16) one should use accelerated block-coordinate descent method [19, 22, 31, 71].

To solve problem (15) we use the following trick [15, 23]: instead of (15) we consider penalized problem

$$F(x) = f(x) + \frac{R_y^2}{\varepsilon}\|Ax\|_2^2 \to \min_{x\in Q}, \tag{21}$$

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 1 (See also Remark 4.3 from [23]). Assume that $x^N \in Q$ is such that

$$F(x^N) - \min_{x \in Q} F(x) \leq \varepsilon. \tag{22}$$

Then

$$f(x^N) - \min_{Ax=0, x \in Q} f(x) \leq \varepsilon, \quad \|Ax^N\|_2 \leq \frac{2\varepsilon}{R_y}. \tag{23}$$

Next, we introduce $h(x) \overset{\text{def}}{=} {}^{R_y^2\|Ax\|_2^2}/_{\varepsilon}$ and notice that problem (21) is a special case of the problem (10). First of all, we consider the case when f is convex and L-smooth, $Q = \mathbb{R}^n$ and full gradients of f and h are available, i.e. we consider deterministic first-order oracle without noise. 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

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

for all $x, y \in \mathbb{R}^n$. We can apply `STM` with inexact proximal step (`STP_IPS`) which is presented in Section C as Algorithm 8 to solve problem (21). Corollary 6 (see Section C in the Appendix) states that in order to get such x^N that satisfy (22) we should run `STP_IPS` for $N = O\left(\sqrt{{}^{LR^2}/_{\varepsilon}}\right)$ iterations with $\delta = O\left({}^{\varepsilon^{3/2}}/_{((L_h+L)\sqrt{L}R^3)}\right)$, 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 $g_{k+1}(z^{k+1}) - g_{k+1}(\hat{z}^{k+1}) \leq \delta\|z^k - \hat{z}^{k+1}\|_2^2$ where $g_{k+1}(z)$ is defined as (see also (105))

$$g_{k+1}(z) = \frac{1}{2}\|z^k - z\|_2^2 + \alpha_{k+1}\left(f(\tilde{x}^{k+1}) + \langle \nabla f(\tilde{x}^{k+1}), z - \tilde{x}^{k+1}\rangle + h(z)\right), \quad k = 0, 1, \ldots$$

and $\hat{z}^{k+1} = \operatorname{argmin}_{z \in \mathbb{R}^n} g_{k+1}(z)$. 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 1.

However, in order to solve the auxiliary problem $g_{k+1}(z) \to \min_{z \in \mathbb{R}^n}$ one should run another optimization method as a subroutine, e.g. `STM`. Note that $\text{Im}A = \text{Im}A^\top = (\text{Ker}A)^\perp$ and the iterates of `STM_IPS` with `STM` as a subroutine lie in $x^0 + (\text{Ker}A)^\perp$ (one can prove the last statement using simple induction, see Theorem 8 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 $x^0 + (\text{Ker}A)^\perp$ 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 $x^0 + (\text{Ker}A)^\perp$:

$$\frac{1 + {}^{2\alpha_{k+1}R_y^2\lambda_{\max}(A^\top A)}/_{\varepsilon}}{1 + {}^{2\alpha_{k+1}R_y^2\lambda^+_{\min}(A^\top A)}/_{\varepsilon}} \leq \frac{\lambda_{\max}(A^\top A)}{\lambda^+_{\min}(A^\top A)} \overset{\text{def}}{=} \chi(A^\top A). \tag{24}$$

It means that to achieve $g_{k+1}(z^{k+1}) - g_{k+1}(\hat{z}^{k+1}) \leq \delta\|z^k - \hat{z}^{k+1}\|_2^2$ with $\delta = O\left({}^{\varepsilon^{3/2}}/_{((L_h+L)\sqrt{L}R^3)}\right)$ 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 where

$$\begin{aligned} T &= O\left(\sqrt{\chi(A^\top A)}\ln\left(\frac{L_{g_N}\sqrt{L}(\lambda_{\max}(A^\top A) + L)R^3}{\varepsilon^{3/2}}\right)\right), \\ L_{g_N} &= 1 + \frac{2\alpha_{k+1}R_y^2\lambda_{\max}(A^\top A)}{\varepsilon} \overset{(116)+(196)}{=} O\left(\frac{R_y^2 R\lambda_{\max}(A^\top A)}{\sqrt{L}\varepsilon^{3/2}}\right) \end{aligned}$$

or, equivalently,

$$T = O\left(\sqrt{\chi(A^\top A)}\ln\left(\frac{\lambda_{\max}(A^\top A)(\lambda_{\max}(A^\top A)+L)R_y^2R^4}{\varepsilon^3}\right)\right). \tag{25}$$

That is, number of $A^\top Ax$ calculations equals NT and it matches the optimal bound for deterministic convex and L-smooth problems of type (1) multiplied by $\sqrt{\chi(A^\top A)}$ up to logarithmic factors (see Table 1).

We believe that using the same recurrence technique that we use in Sections C and 5 one can generalize this result for the case when instead of $\nabla f(x)$ only stochastic gradient $\nabla f(x,\xi)$ (see inequalities (8)-(9)) 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 1 and 2 in the corresponding cases with additional factor $\sqrt{\chi(A^\top A)}$ and logarithmic factors. Recently such bounds were shown in [20] for the distributed version of Multistage Accelerated Stochastic Gradient method from [6]. 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: $\|\nabla f(x)\|_2 \leq M$ for all $x \in Q$. Let us start with the case $\mu = 0$. Then, to achieve (22) one can run `Sliding` method from [46, 48] considering $f(x)$ as a composite term. In this case `Sliding` requires

$$O\left(\sqrt{\frac{\lambda_{\max}(A^\top A)R_y^2R^2}{\varepsilon^2}}\right) \text{ calculations of } A^\top Ax \tag{26}$$

and

$$O\left(\frac{M^2R^2}{\varepsilon^2}\right) \text{ calculations of } \nabla f(x). \tag{27}$$

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 (8)-(9) with $\delta = 0$) one can show [46, 48] that Stochastic `Sliding` (`S-Sliding`) method can achieve (22) with probability at least $1-\beta$, $\beta \in (0,1)$, and it requires the same number of calculations of $A^\top Ax$ as in (26) up to logarithmic factors and

$$\widetilde{O}\left(\frac{(M^2+\sigma^2)R^2}{\varepsilon^2}\right) \text{ calculations of } \nabla f(x,\xi). \tag{28}$$

When $\mu > 0$ one can apply restarts technique on top of `S-Sliding` (`RS-Sliding`) [15, 78] and get that to guarantee (22) with probability at least $1-\beta$, $\beta \in (0,1)$ `RS-Sliding` requires

$$\widetilde{O}\left(\sqrt{\frac{\lambda_{\max}(A^\top A)R_y^2}{\mu\varepsilon}}\right) \text{ calculations of } A^\top Ax \tag{29}$$

and

$$\widetilde{O}\left(\frac{M^2+\sigma^2}{\mu\varepsilon}\right) \text{ calculations of } \nabla f(x,\xi). \tag{30}$$

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.

5 Stochastic Convex Optimization with Affine Constraints: Dual Approach

In this section we assume that one can construct a dual problem for (15). If f is μ-strongly convex in ℓ_2-norm, then ψ and φ have L_ψ–Lipschitz continuous and L_φ–Lipschitz continuous in ℓ_2-norm gradients respectively [38, 64], where $L_\psi = {\lambda_{\max}(A^\top A)}/{\mu}$ and $L_\varphi = {1}/{\mu}$. In our proofs we often use Demyanov–Danskin theorem [64] which states that

$$\nabla\psi(y) = Ax(A^\top y), \quad \nabla\varphi(y) = x(y). \tag{31}$$

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

Below we propose one 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 all subsection below we assume that $Q = \mathbb{R}^n$.

5.1 Convex Dual Function

In this section we assume that the dual function $\varphi(y)$ could be rewritten as an expectation, i.e. $\varphi(y) = \mathbb{E}_\xi\left[\varphi(y,\xi)\right]$, 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: $\psi(y) = \mathbb{E}_\xi\left[\psi(y,\xi)\right]$. Consider the stochastic function $f(x,\xi)$ which is defined implicitly as follows:

$$\varphi(y,\xi) = \max_{x\in\mathbb{R}^n}\left\{\langle y, x\rangle - f(x,\xi)\right\}. \tag{32}$$

Similarly to the deterministic case we introduce $x(y,\xi) \stackrel{\text{def}}{=} \operatorname{argmax}_{x\in\mathbb{R}^n}\left\{\langle y,x\rangle - f(x,\xi)\right\}$ 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 $x(y) = \mathbb{E}_\xi[x(y,\xi)]$ and $\nabla\psi(y) = \mathbb{E}_\xi[\nabla\psi(y,\xi)]$.

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

$$\left\|\mathbb{E}_\xi[\delta(y,\xi)]\right\|_2 \leq \delta_y, \quad \forall y \in \mathbb{R}^n. \tag{33}$$

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

$$\mathbb{E}_\xi\left[\exp\left(\frac{\left\|\tilde{x}(y,\xi) - \mathbb{E}_\xi\left[\tilde{x}(y,\xi)\right]\right\|_2^2}{\sigma_x^2}\right)\right] \leq \exp(1), \quad \forall y \in \mathbb{R}^n, \tag{34}$$

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

$$\left\|\mathbb{E}_\xi\left[\tilde{\nabla}\psi(y,\xi)\right] - \nabla\psi(y)\right\|_2 \leq \delta, \quad \forall y \in \mathbb{R}^n, \tag{35}$$

$$\mathbb{E}_\xi\left[\exp\left(\frac{\left\|\tilde{\nabla}\psi(y,\xi) - \mathbb{E}_\xi\left[\tilde{\nabla}\psi(y,\xi)\right]\right\|_2^2}{\sigma_\psi^2}\right)\right] \leq \exp(1), \quad \forall y \in \mathbb{R}^d, \tag{36}$$

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

$$\tilde{\nabla}\Psi(y,\boldsymbol{\xi}^k) = \frac{1}{r_k}\sum_{l=1}^{r_k}\tilde{\nabla}\psi(y,\xi^l), \quad \tilde{x}(y,\boldsymbol{\xi}^k) = \frac{1}{r_k}\sum_{l=1}^{r_k}\tilde{x}(y,\xi^l) \tag{37}$$

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

$$\left\|\mathbb{E}\left[\tilde{\nabla}\Psi(x,\boldsymbol{\xi}^k)\right] - \nabla\psi(x)\right\|_2 \leq \delta, \quad \forall x \in \mathbb{R}^n, \tag{38}$$

$$\mathbb{E}\left[\exp\left(\frac{\left\|\tilde{\nabla}\Psi(x,\boldsymbol{\xi}^k) - \mathbb{E}\left[\tilde{\nabla}\Psi(x,\boldsymbol{\xi}^k)\right]\right\|_2^2}{O(\sigma_\psi^2/r_k^2)}\right)\right] \leq \exp(1), \quad \forall x \in \mathbb{R}^n, \tag{39}$$

where in the last inequality we used combination of Lemmas 7 and 9 (see two inequalities after (132) for the details). We call this approach `SPDSTM` (Stochastic Primal-Dual Similar Triangles Method, see Algorithm 2). Note that Algorithm 4 from [16] 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 2 `SPDSTM`

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

1: **for** $k = 0, \ldots, N$ **do**
2: Set $\tilde{L} = 2L_\psi$
3: Set $A_{k+1} = A_k + \alpha_{k+1}$, where $2\tilde{L}\alpha_{k+1}^2 = A_k + \alpha_{k+1}$
4: $\tilde{y}^{k+1} = {}^{(A_k y^k + \alpha_{k+1} z^k)}/_{A_{k+1}}$
5: $z^{k+1} = z^k - \alpha_{k+1}\tilde{\nabla}\Psi(\tilde{y}^{k+1},\boldsymbol{\xi}^k)$
6: $y^{k+1} = {}^{(A_k y^k + \alpha_{k+1} z^{k+1})}/_{A_{k+1}}$
7: **end for**

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

The following lemma is rather technical and provides useful inequalities that show how biasedness of $\tilde{\nabla}\Psi(y,\boldsymbol{\xi}^k)$ interacts with convexity and L_ψ-smoothness of ψ.

Lemma 2. Assume that function $\psi(y)$ is convex and L_ψ-smooth on $\mathbb{R}^n$. Then for all $x, y \in \mathbb{R}^n$

$$\psi(y) \geq \psi(x) + \left\langle\mathbb{E}\left[\tilde{\nabla}\Psi(x,\boldsymbol{\xi}^k)\right], y - x\right\rangle - \delta\|y - x\|_2, \tag{40}$$

$$\psi(y) \leq \psi(x) + \left\langle\mathbb{E}\left[\tilde{\nabla}\Psi(x,\boldsymbol{\xi}^k)\right], y - x\right\rangle + L_\psi\|y - x\|_2^2 + \frac{\delta^2}{2L_\psi}. \tag{41}$$

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

Lemma 3 (see also Theorem 1 from [17])**.** For each iteration of Algorithm 2 we have

$$\begin{aligned} A_N\psi(y^N) \;\leq\; & \frac{1}{2}\|z-z^0\|_2^2 - \frac{1}{2}\|z-z^N\|_2^2 \\ & + \sum_{k=0}^{N-1}\alpha_{k+1}\left(\psi(\tilde{y}^{k+1}) + \langle\tilde{\nabla}\Psi(\tilde{y}^{k+1},\boldsymbol{\xi}^{k+1}), z-\tilde{y}^{k+1}\rangle\right) \\ & + \sum_{k=0}^{N-1} A_k\left\langle\tilde{\nabla}\Psi(\tilde{y}^{k+1},\boldsymbol{\xi}^{k+1}) - \mathbb{E}_k\left[\tilde{\nabla}\Psi(\tilde{y}^{k+1},\boldsymbol{\xi}^{k+1})\right], y^k-\tilde{y}^{k+1}\right\rangle \\ & + \sum_{k=0}^{N-1}\frac{A_{k+1}}{2\tilde{L}}\left\|\mathbb{E}_k\left[\tilde{\nabla}\Psi(\tilde{y}^{k+1},\boldsymbol{\xi}^{k+1})\right] - \tilde{\nabla}\Psi(\tilde{y}^{k+1},\boldsymbol{\xi}^{k+1})\right\|_2^2 \\ & + \delta\sum_{k=0}^{N-1}A_k\|y^k-\tilde{y}^{k+1}\|_2 + \delta^2\sum_{k=0}^{N-1}\frac{A_{k+1}}{\tilde{L}}, \end{aligned} \tag{42}$$

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

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 (see also Lemma 7 from [16])**.** Let the sequences of non-negative numbers $\{\alpha_k\}_{k\geq 0}$, random non-negative variables $\{R_k\}_{k\geq 0}$ and random vectors $\{\eta^k\}_{k\geq 0}$, $\{a^k\}_{k\geq 0}$ satisfy inequality

$$\frac{1}{2}R_l^2 \leq A + h\delta\sum_{k=0}^{l-1}\alpha_{k+1}\widetilde{R}_k + u\sum_{k=0}^{l-1}\alpha_{k+1}\langle\eta^k,a^k\rangle + c\sum_{k=0}^{l-1}\alpha_{k+1}^2\|\eta^k\|_2^2, \quad \forall l=1,\ldots,N, \tag{43}$$

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

$$\mathbb{E}\left[\eta^k \mid \eta^0,\ldots,\eta^{k-1}\right] = 0, \qquad \mathbb{E}\left[\exp\left(\frac{\|\eta^k\|_2^2}{\sigma_k^2}\right) \mid \eta^0,\ldots,\eta^{k-1}\right] \leq \exp(1), \qquad \forall k\geq 0, \tag{44}$$

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

$$\widetilde{R}_l \leq JR_0 \tag{45}$$

and

$$u\sum_{k=0}^{l-1}\alpha_{k+1}\langle\eta^k,a^k\rangle + c\sum_{k=0}^{l-1}\alpha_{k+1}^2\|\eta^k\|_2^2 \leq \left(24cCDH + hGDJ + udC_1\sqrt{CDHJg(N)}\right)R_0^2 \tag{46}$$

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

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

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

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

Finally, we state the main result of this section.

Theorem 2 (see also Theorem 2 from [16])**.** 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 $R_f = \tilde{\Omega}\left(\max\left\{\frac{R_y}{A_N\sqrt{\lambda_{\max}(A^\top A)}}, \frac{\sqrt{\lambda_{\max}(A^\top A)}R_y}{\mu}, R_x\right\}\right)$, where R_y is such that $\|y^*\|_2 \leq R_y$, y^* is the solution of the dual problem (16), and $R_x = \|x(A^\top y^*)\|_2$. Assume that at iteration k of Algorithm 2 batch size is chosen according to the formula $r_k \geq \max\left\{1, \frac{\sigma_\psi^2\widetilde{\alpha}_k\ln(N/\beta)}{\hat{C}\varepsilon}\right\}$, where $\widetilde{\alpha}_k = \frac{k+1}{2\tilde{L}}$, $0 < \varepsilon \leq \frac{H\tilde{L}R_0^2}{N^2}$, $0 \leq \delta \leq \frac{G\tilde{L}R_0}{(N+1)^2}$ and $N \geq 1$ for some numeric constant $H > 0$, $G > 0$ and $\hat{C} > 0$. Then with probability $\geq 1 - 4\beta$

$$\begin{aligned}\psi(y^N) + f(\tilde{x}^N) + 2R_y\|A\tilde{x}^N\|_2 \;\leq\; & \frac{R_y^2}{A_N}\left(8\sqrt{HC_2} + 2 + 12CH + \frac{G(6J+4)}{2}\right.\\ & + \frac{L_f\left(\sqrt{96C_2H} + G\right)}{2R_y\sqrt{\lambda_{\max}(A^\top A)}} + \frac{G^2}{2(N+1)}\\ & \left. + C_1\sqrt{\frac{CHJg(N)}{2}} + \sqrt{96C_2H} + G\right), \quad (47)\end{aligned}$$

where $\beta \in (0, 1/4)$ is such that $\frac{1+\sqrt{\ln\frac{1}{\beta}}}{\sqrt{\ln\frac{N}{\beta}}} \leq 2$, 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 $\chi(A^\top A) = \frac{\lambda_{\max}(A^\top A)}{\lambda^+_{\min}(A^\top A)}$, the outputs $\tilde{x}^N$ and y^N of Algorithm 2 satisfy the following condition

$$f(\tilde{x}^N) - f(x^*) \leq f(\tilde{x}^N) + \psi(y^N) \leq \varepsilon, \quad \|A\tilde{x}^N\|_2 \leq \frac{\varepsilon}{R_y} \quad (48)$$

with probability at least $1 - 4\beta$. What is more, to guarantee (48) with probability at least $1 - 4\beta$ Algorithm 2 requires

$$\widetilde{O}\left(\max\left\{\frac{\sigma_\psi^2M_f^2}{\varepsilon^2\lambda^+_{\min}(A^\top A)}\ln\left(\frac{1}{\beta}\sqrt{\frac{M_f}{\mu\varepsilon}\chi(A^\top A)}\right), \sqrt{\frac{M_f}{\mu\varepsilon}\chi(A^\top A)}\right\}\right) \quad (49)$$

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

5.2 Strongly Convex Dual Function and Restarts Technique

In this section we assume that primal functional f is additionally L-smooth. It implies that the dual function ψ in (16) is additionally μ_ψ-strongly convex in $y^0 + (\text{Ker}A^\top)^\perp$ where $\mu_\psi = {}^{\lambda^+_{\min}(A^\top A)}/_{L}$ [38, 64] and $\lambda^+_{\min}(A^\top A)$ is the minimal positive eigenvalue of $A^\top A$.

From weak duality $-f(x^*) \leq \psi(y^*)$ and (18) we get the key relation of this section (see also [3, 4, 58])

$$f(x(A^\top y)) - f(x^*) \leq \langle \nabla\psi(y), y\rangle = \langle Ax(A^\top y), y\rangle \tag{50}$$

This inequality implies the following theorem.

Theorem 3. Consider function f and its dual function ψ defined in (18) such that problems (15) and (16) 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:

$$f(x^N) - f(x^*) \leq 2\varepsilon, \quad \|Ax^N\|_2 \leq \frac{\varepsilon}{R_y}, \tag{51}$$

where x^* is any minimizer of f.

Proof. Applying Cauchy-Schwarz inequality to (50) we get

$$f(x^N) - f(x^*) \overset{(50)}{\leq} \|\nabla\psi(y^N)\|_2 \cdot \|y^N\|_2 \leq \frac{\varepsilon}{R_y} \cdot 2R_y = 2\varepsilon.$$

The second part (51) immediately follows from $\|\nabla\psi(y^N)\|_2 \leq {}^{\varepsilon}/_{R_y}$ and Demyanov-Danskin theorem which implies $\nabla\psi(y^N) = Ax^N$. □

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 3) [21] with $\texttt{AC-SA}^2$ (see Algorithm 5) as a subroutine (i.e. $\texttt{RRMA-AC-SA}^2$) which is based on $\texttt{AC-SA}$ algorithm (see Algorithm 4) from [26]. We notice that $\texttt{RRMA-AC-SA}^2$ is applied for a regularized dual function

$$\tilde{\psi}(y) = \psi(y) + \frac{\lambda}{2}\|y - y^0\|_2^2, \tag{52}$$

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

$$\nabla\Psi(y, \boldsymbol{\xi}^t, r_t) = \frac{1}{r_t}\sum_{l=1}^{r_t} \nabla\psi(y, \xi^l), \quad x(y, \boldsymbol{\xi}^t, r_t) = \frac{1}{r_t}\sum_{l=1}^{r_t} x(y, \xi^l). \tag{53}$$

As before in the cases when the batch-size r_t can be restored from the context, we will use simplified notation $\nabla\Psi(y, \boldsymbol{\xi}^t)$ and $x(y, \boldsymbol{\xi}^t)$. In the $\texttt{AC-SA}$ algorithm we use batched stochastic gradients of functions ψ_k which are defined as follows:

$$\nabla\Psi_k(y, \boldsymbol{\xi}^t) = \frac{1}{r_t}\sum_{l=1}^{r_t} \nabla\psi_k(y, \xi^l), \quad \nabla\psi_k(y, \xi) = \nabla\psi(y, \xi) + \lambda(y - y^0) + \lambda\sum_{l=1}^{k} 2^l(y - \hat{y}^l). \tag{54}$$

Algorithm 3 RRMA−AC−SA^2 [21]

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

1: $\psi_0 \leftarrow \tilde{\psi}$, $\hat{y}^0 \leftarrow y^0$, $T \leftarrow \left\lfloor \log_2 \frac{\tilde{L}_\psi}{\lambda} \right\rfloor$
2: **for** $k = 1, \ldots, T$ **do**
3: Run AC−SA^2 for m/T iterations to optimize ψ_{k-1} with $\hat{y}^{k-1}$ as a starting point and get the output $\hat{y}^k$
4: $\psi_k(y) \leftarrow \tilde{\psi}(y) + \lambda \sum_{l=1}^{k} 2^{l-1} \|y - \hat{y}^l\|_2^2$
5: **end for**

Output: $\hat{y}^T$.

Algorithm 4 AC−SA [26]

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

1: $y_{ag}^0 \leftarrow z^0$, $y_{md}^0 \leftarrow z^0$
2: **for** $t = 1, \ldots, m$ **do**
3: $\alpha_t \leftarrow \frac{2}{t+1}$, $\gamma_t \leftarrow \frac{4\tilde{L}_\psi}{t(t+1)}$
4: $y_{md}^t \leftarrow \frac{(1-\alpha_t)(\lambda+\gamma_t)}{\gamma_t+(1-\alpha_t^2)\lambda} y_{ag}^{t-1} + \frac{\alpha_t((1-\alpha_t)\lambda+\gamma_t)}{\gamma_t+(1-\alpha_t^2)\lambda} z^{t-1}$
5: $z^t \leftarrow \frac{\alpha_t\lambda}{\lambda+\gamma_t} y_{md}^t + \frac{(1-\alpha_t)\lambda+\gamma_t}{\lambda+\gamma_t} z^{t-1} - \frac{\alpha_t}{\lambda+\gamma_t} \nabla\Psi_k(y_{md}^t, \boldsymbol{\xi}^t)$
6: $y_{ag}^t \leftarrow \alpha_t z^t + (1-\alpha_t) x_{ag}^{t-1}$
7: **end for**

Output: y_{ag}^m.

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

Theorem 4 (Corollary 1 from [21])**.** Let ψ be L_ψ-smooth and μ_ψ-strongly convex function and $\lambda = \Theta\left(\frac{L_\psi \ln^2 N}{N^2}\right)$ for some $N > 1$. If the Algorithm 3 performs N iterations in total[1] with batch size r for all iterations, then it will provide such a point $\hat{y}$ that

$$\mathbb{E}\left[\|\nabla\psi(\hat{y})\|_2^2 \mid y^0, r\right] \leq C\left(\frac{L_\psi^2 \|y^0 - y^*\|_2^2 \ln^4 N}{N^4} + \frac{\sigma_\psi^2 \ln^6 N}{rN}\right), \tag{55}$$

where $C > 0$ is some positive constant and y^* is a solution of the dual problem (16).

Let us show that w.l.o.g. we can assume in this section that function ψ defined in (18) is μ_ψ-strongly convex everywhere with $\mu_\psi = \lambda_{\min}^+(A^\top A)/L$. In fact, from L-smoothness of f we have only that ψ is μ_ψ-strongly convex in $y^0 + \left(\text{Ker}(A^\top)\right)^\perp$ (see [38, 64] 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 $y^0 + \left(\text{Ker}(A^\top)\right)^\perp$.

Theorem 5. 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 z^0 as a starting point, where $z^0, \hat{y}^1, \ldots, \hat{y}^k$ are some points from $y^0 + \left(\text{Ker}(A^\top)\right)^\perp$ and $y^0 \in \mathbb{R}^n$. Then for all $t \geq 0$ we have $y_{md}^t, z^t, y_{ag}^t \in y^0 + \left(\text{Ker}(A^\top)\right)^\perp$.

Proof. We prove the statement of the theorem by induction. For $t = 0$ the statement is trivial, since $y_{md}^0 = y_{ag}^0 = z^0 \in y_0 + \left(\text{Ker}(A^\top)\right)^\perp$. Assume that $y_{md}^t, z^t, y_{ag}^t \in y^0 + \left(\text{Ker}(A^\top)\right)^\perp$ for some

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

Algorithm 5 $\texttt{AC-SA}^2$ [21]

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

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

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

Output: y^2.

$t \geq 0$ and prove it for $t+1$. Since $y_0 + \left(\text{Ker}(A^\top)\right)^\perp$ is a convex set and y_{md}^{t+1} is a convex combination of y_{ag}^t and z^t we have $y_{md}^{t+1} \in y^0 + \left(\text{Ker}(A^\top)\right)^\perp$. Next, the point $\frac{\alpha_t\lambda}{\lambda+\gamma_t} y_{md}^{t+1} + \frac{(1-\alpha_t)\lambda+\gamma_t}{\lambda+\gamma_t} z^t$ also lies in $y^0 + \left(\text{Ker}(A^\top)\right)^\perp$ since it is convex combination of the points lying in this set. Due to (52), (53) and (54) we have that $\nabla\Psi_k(y_{md}^{t+1}, \boldsymbol{\xi}^t) = Ax(A^\top y_{md}^{t+1}, \boldsymbol{\xi}^t) + \lambda(y_{md}^{t+1} - y^0) + \lambda\sum_{l=1}^k 2^l(y_{md}^{t+1} - \hat{y}^l)$. The first term lies in $\left(\text{Ker}(A^\top)\right)^\perp$ since $\text{Im}(A) = \left(\text{Ker}(A^\top)\right)^\perp$ and the second and the third terms also lie in $\left(\text{Ker}(A^\top)\right)^\perp$ since $y_{md}^{t+1}, y^0, \hat{y}^1, \ldots, \hat{y}^k \in y^0 + \left(\text{Ker}(A^\top)\right)^\perp$. Putting all together we get $z^{t+1} \in y^0 + \left(\text{Ker}(A^\top)\right)^\perp$. Finally, y_{ag}^{t+1} lies in $y^0 + \left(\text{Ker}(A^\top)\right)^\perp$ as a convex combination of points from this set. □

Corollary 1. Assume that Algorithm 3 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 $\hat{y}^k \in y^0 + \left(\text{Ker}(A^\top)\right)^\perp$.

Proof. We prove this result by induction. For $t = 0$ the statement is trivial since $\hat{y}^0 = y^0$. Next, assume that $\hat{y}^0, \hat{y}^1, \ldots, \hat{y}^k \in y^0 + \left(\text{Ker}(A^\top)\right)^\perp$ and prove that $\hat{y}^{k+1} \in y^0 + \left(\text{Ker}(A^\top)\right)^\perp$. Our assumption implies that the assumptions from Theorem 5 and applying the result of the theorem we get that y^1 and y^2 from the method $\texttt{AC-SA}^2$ applied to the ψ_k also lie in $y^0 + \left(\text{Ker}(A^\top)\right)^\perp$. That is, the output of $\texttt{AC-SA}^2$ applied for ψ_k lies in $y^0 + \left(\text{Ker}(A^\top)\right)^\perp$. □

Now we are ready to present our approach which was sketched in [15] of constructing an accelerated method for the strongly convex dual problem using restarts of $\texttt{RRMA-AC-SA}^2$. To explain the main idea we start with the simplest case: $\sigma_\psi^2 = 0$, $r = 0$. It means that there is no stochasticity in the method and the bound (55) can be rewritten in the following form:

$$\|\nabla\psi(\hat{y})\|_2 \leq \frac{\sqrt{C}L_\psi\|y^0 - y^*\|_2 \ln^2 N}{N^2} \leq \frac{\sqrt{C}L_\psi\|\nabla\psi(y^0)\|_2 \ln^2 N}{\mu_\psi N^2}, \tag{56}$$

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

In the case when $\sigma_\psi^2 \neq 0$ 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 $\Omega\left(\sigma_\psi^2/(\bar{N}\|\nabla\psi(\bar{y}^{l-1})\|_2^2)\right)$. However, in the stochastic case we do not have an access to the $\nabla\psi(\bar{y}^{l-1})$, 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 $\nabla\Psi(\bar{y}^{l-1}, \boldsymbol{\xi}^{l-1}, \hat{r}_l)$ of $\nabla\psi(\bar{y}^{l-1})$ 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:

$$\mathbb{E}\left[\|\nabla\psi(\bar{y}^l)\|_2^2 \mid \bar{y}^{l-1}, r_l, \hat{r}_l\right] \leq A_l \overset{\text{def}}{=} \frac{\|\nabla\psi(\bar{y}^{l-1})\|_2^2}{8} + \frac{\|\nabla\Psi(\bar{y}^{l-1}, \boldsymbol{\xi}^{l-1}, \hat{r}_l) - \nabla\psi(\bar{y}^{l-1})\|_2^2}{32}.$$

The third ingredient is the amplification trick: we run $p_l = \Omega(\ln(1/\beta))$ independent trajectories of RRMA−AC−SA^2, get points $\bar{y}^{l,1}, \ldots, \bar{y}^{l,p_l}$ and choose such $\bar{y}^{l,p(l)}$ among of them that $\|\nabla\psi(\bar{y}^{l,p(l)})\|_2$ is *close enough* to $\min_{p=1,\ldots,p_l}\|\nabla\psi(\bar{y}^{l,p})\|_2$ with high probability, i.e.

$$\|\nabla\psi(\bar{y}^{l,p(l)})\|_2^2 \leq 2\min_{p=1,\ldots,p_l}\|\nabla\psi(\bar{y}^{l,p})\|_2^2 + \varepsilon^2/8R_y^2$$

with probability at least $1-\beta$ for fixed $\nabla\Psi(\bar{y}^{l-1}, \boldsymbol{\xi}^{l-1}, \hat{r}_l)$. We achieve it due to additional sampling of $\bar{r}_l \sim R_y^2/\varepsilon^2$ stochastic gradients at $\bar{y}^{l,p}$ 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 $p = 1, \ldots, p_l$

$$\mathbb{P}\left\{\|\nabla\psi(\bar{y}^{l,p})\|_2^2 \geq 2A_l \mid \bar{y}^{l-1}, r_l, \bar{r}_l\right\} \leq \frac{1}{2},$$

hence

$$\mathbb{P}\left\{\min_{p=1,\ldots,p_l}\|\nabla\psi(\bar{y}^{l,p})\|_2^2 \geq 2A_l \mid \bar{y}^{l-1}, r_l, \bar{r}_l\right\} \leq \frac{1}{2^{p_l}}.$$

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

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

for fixed $\nabla\Psi(\bar{y}^{l-1}, \boldsymbol{\xi}^{l-1}, \hat{r}_l)$ which means that

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

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

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

The approach informally described above is stated as Algorithm 6.

Theorem 6. Assume that ψ is μ_ψ-strongly convex and L_ψ-smooth. If Algorithm 6 is run with

$$l = \max\left\{1, \log_2\frac{2R_y^2\|\nabla\psi(y^0)\|_2^2}{\varepsilon^2}\right\}, \hat{r}_k = \max\left\{1, \frac{4\sigma_\psi^2\left(1+\sqrt{3\ln\frac{l}{\beta}}\right)^2R_y^2}{\varepsilon^2}\right\},$$

$$r_k = \max\left\{1, \frac{64C\sigma_\psi^2\ln^6\bar{N}}{\bar{N}\|\nabla\Psi(\bar{y}^{k-1,p(k-1)}, \boldsymbol{\xi}^{k-1,p(k-1)}, \hat{r}_k)\|_2^2}\right\},$$

$$p_k = \max\left\{1, \log_2\frac{l}{\beta}\right\}, \bar{r}_k = \max\left\{1, \frac{128\sigma_\psi^2\left(1+\sqrt{3\ln\frac{lp_k}{\beta}}\right)^2R_y^2}{\varepsilon^2}\right\}, k = 1, \ldots, l, \quad (57)$$

Algorithm 6 `Restarted-RRMA-AC-SA`^{2}

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

1: Choose the smallest integer $\bar{N} > 1$ such that $\frac{CL_\psi^2 \ln^4 \bar{N}}{\mu_\psi^2 \bar{N}^4} \leq \frac{1}{32}$
2: $\bar{y}^{0,p(0)} \leftarrow y^0$
3: **for** $k = 1, \ldots, l$ **do**
4: Compute $\nabla\Psi(\bar{y}^{k-1,p(k-1)}, \boldsymbol{\xi}^{k-1,p(k-1)}, \hat{r}_k)$
5: $r_k \leftarrow \max\left\{1, \frac{64C\sigma_\psi^2 \ln^6 \bar{N}}{\bar{N}\|\nabla\Psi(\bar{y}^{k-1,p(k-1)}, \boldsymbol{\xi}^{k-1,p(k-1)}, \hat{r}_k)\|_2^2}\right\}$
6: Run p_k independent trajectories of `RRMA-AC-SA`^{2} for $\bar{N}$ iterations with batch-size r_k with $\bar{y}^{k-1,p(k-1)}$ as a starting point and get outputs $\bar{y}^{k,1}, \ldots, \bar{y}^{k,p_k}$
7: Compute $\nabla\Psi(\bar{y}^{k,1}, \boldsymbol{\xi}^{k,1}, \bar{r}_k), \ldots, \nabla\Psi(\bar{y}^{k,p_k}, \boldsymbol{\xi}^{k,p_k}, \bar{r}_k)$
8: $p(k) \leftarrow \operatorname{argmin}_{p=1,\ldots,p_k} \|\nabla\Psi(\bar{y}^{k,p}, \boldsymbol{\xi}^{k,p}, \bar{r}_k)\|_2$
9: **end for**

Output: $\bar{y}^{l,p(l)}$.

where $\bar{N} > 1$ is such that $\frac{CL_\psi^2 \ln^4 \bar{N}}{\mu_\psi^2 \bar{N}^4} \leq \frac{1}{32}$, $\beta \in (0, 1/3)$ and $\varepsilon > 0$, then with probability at least $1 - 3\beta$

$$\|\nabla\psi(\bar{y}^{l,p(l)})\|_2 \leq \frac{\varepsilon}{R_y} \tag{58}$$

and the total number of the oracle calls equals

$$\sum_{k=1}^{l}(\hat{r}_k + \bar{N}p_k r_k + p_k\bar{r}_k) = \widetilde{O}\left(\max\left\{\sqrt{\frac{L_\psi}{\mu_\psi}}, \frac{\sigma_\psi^2 R_y^2}{\varepsilon^2}\right\}\right). \tag{59}$$

Corollary 2. Under assumptions of Theorem 6 we get that with probability at least $1 - 3\beta$

$$\|\bar{y}^{l,p(l)} - y^*\|_2 \leq \frac{\varepsilon}{\mu_\psi R_y}, \tag{60}$$

where $\beta \in (0, 1/3)$ the total number of the oracle calls is defined in (59).

Proof. Inequalities (58) and $\mu_\psi\|y - y^*\|_2 \leq \|\nabla\psi(y)\|_2$ which follows from μ_ψ-strong convexity of ψ imply that

$$\|\bar{y}^{l,p(l)} - y^*\|_2 \leq \frac{\|\nabla\psi(\bar{y}^{l,p(l)})\|_2}{\mu_\psi} \overset{(58)}{\leq} \frac{\varepsilon}{\mu_\psi R_y}.$$

□

Corollary 3. Let the assumptions of Theorem 6 hold. Assume that f is L_f-Lipschitz continuous on $B_{R_f}(0)$ where $R_f = \left(\frac{\mu_\psi}{8\sqrt{\lambda_{\max}(A^\top A)}} + \frac{\sqrt{\lambda_{\max}(A^\top A)}}{\mu} + \frac{R_x}{R_y}\right) R_y$ and $R_x = \|x(A^\top y^*)\|_2$. Then, with probability at least $1 - 4\beta$

$$f(x^l) - f(x^*) \leq \left(2 + \frac{L_f}{8R_y\sqrt{\lambda_{\max}(A^\top A)}}\right)\varepsilon, \quad \|Ax^l\| \leq \frac{9\varepsilon}{8R_y}, \tag{61}$$

where $\beta \in (0, 1/4)$, $\varepsilon \in (0, \mu_\psi R_y^2)$ $x^l \overset{\text{def}}{=} x(A^\top\bar{y}^{l,p(l)}, \boldsymbol{\xi}^{l,p(l)}, \bar{r}_l)$ and to achieve it we need the total number of oracle calls as in (59).

5.3 Direct Acceleration for Strongly Convex Dual Function

We consider first the following minimization problem:

$$\psi(y) \to \min_{y \in \mathbb{R}^n}, \tag{62}$$

where $\psi(y)$ is μ_ψ-strongly convex and L_ψ-smooth. We use the same notation to define the objective in (62) as for the dual function from (16) because later in the section we apply the algorithm introduced below to the (16), but for now it is not important that ψ is a dual function for (15) and we prefer to consider more general situation. As in Section 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 $\tilde{\nabla}\psi(y,\xi)$ satisfying inequalities (35) and (36) with $\delta \geq 0$ and $\sigma_\psi \geq 0$. In the main method of this section batched version of the stochastic gradient is used:

$$\tilde{\nabla}\Psi(y, \boldsymbol{\xi}^k) = \frac{1}{r_k}\sum_{l=1}^{r_k} \tilde{\nabla}\psi(y, \xi^l), \tag{63}$$

where r_k is the batch-size that we leave unspecified for now. Note that $\tilde{\nabla}\Psi(y, \boldsymbol{\xi}^k)$ satisfies inequalities (38) and (39).

We use Stochastic Similar Triangles Method which is stated in this section as Algorithm 7 to solve problem (62). To define the iterate z^{k+1} we use the following sequence of functions:

$$\begin{aligned}
\tilde{g}_0(z) &\overset{\text{def}}{=} \frac{1}{2}\|z - z^0\|_2^2 + \alpha_0\left(\psi(y^0) + \langle\tilde{\nabla}\Psi(y^0, \boldsymbol{\xi}^0), z - y^0\rangle + \frac{\mu_\psi}{2}\|z - y^0\|_2^2\right), \\
\tilde{g}_{k+1}(z) &\overset{\text{def}}{=} \tilde{g}_k(z) + \alpha_{k+1}(\psi(\tilde{y}^{k+1}) + \langle\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}), z - \tilde{y}^{k+1}\rangle + \frac{\mu_\psi}{2}\|z - \tilde{y}^{k+1}\|_2^2) \\
&= \frac{1}{2}\|z - z^0\|_2^2 + \sum_{l=0}^{k+1}\alpha_l\left(\psi(\tilde{y}^l) + \langle\tilde{\nabla}\Psi(\tilde{y}^l, \boldsymbol{\xi}^l), z - \tilde{y}^l\rangle + \frac{\mu_\psi}{2}\|z - \tilde{y}^l\|_2^2\right)
\end{aligned} \tag{64}$$

We notice that $\tilde{g}_k(z)$ is $(1 + A_k\mu_\psi)$-strongly convex.

Algorithm 7 Stochastic Similar Triangles Methods for strongly convex problems (SSTM_sc)

Input: $\tilde{y}^0 = z^0 = y^0$ — starting point, N — number of iterations
1: Set $\alpha_0 = A_0 = {}^1/_{L_\psi}$
2: Get $\tilde{\nabla}\Psi(y^0, \boldsymbol{\xi}^0)$ to define $\tilde{g}_0(z)$
3: **for** $k = 0, 1, \ldots, N-1$ **do**
4: Choose α_{k+1} such that $A_{k+1} = A_k + \alpha_{k+1}$, $A_{k+1}(1 + A_k\mu_\psi) = \alpha_{k+1}^2 L_\psi$
5: $\tilde{y}^{k+1} = {}^{(A_k y^k + \alpha_{k+1} z^k)}/_{A_{k+1}}$
6: $z^{k+1} = \operatorname{argmin}_{z \in \mathbb{R}^n} \tilde{g}_{k+1}(z)$, where $\tilde{g}_{k+1}(z)$ is defined in (64)
7: $y^{k+1} = {}^{(A_k y^k + \alpha_{k+1} z^{k+1})}/_{A_{k+1}}$
8: **end for**
Output: x^N

Lemma 5. Assume that Algorithm 7 is run to solve problem (62) with $\psi(y)$ being μ_ψ-strongly convex and L_ψ-smooth. Then, for all $k \geq 0$ we have

$$A_k\psi(y^k) \leq \tilde{g}_k(z^k) - \sum_{l=0}^{k-1}\frac{A_l\mu_\psi}{2}\|y^l - \tilde{y}^{l+1}\|_2^2 + \sum_{l=0}^{k}\frac{\alpha_l}{2\mu_\psi}\left\|\tilde{\nabla}\Psi(\tilde{y}^l, \boldsymbol{\xi}^l) - \nabla\psi(\tilde{y}^l)\right\|_2^2. \tag{65}$$

Lemma 6. 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 $\{\eta^k\}_{k\geq 0}$, $\{a^k\}_{k\geq 0}$, $\{\tilde{a}^k\}_{k\geq 0}$ satisfy inequality

$$A_l R_l^2 + \sum_{k=0}^{l-1} A_k \widetilde{R}_k^2 \leq A + h\delta \sum_{k=0}^{l} \alpha_k (R_{k-1} + \widetilde{R}_k) + u \sum_{k=0}^{l-1} \alpha_{k+1} \langle \eta^k, a^k + \tilde{a}^k \rangle + c \sum_{k=0}^{l-1} \alpha_{k+1} \|\eta^k\|_2^2, \tag{66}$$

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

$$\mathbb{E}\left[\eta^k \mid \eta^0, \ldots, \eta^{k-1}\right] = 0, \qquad \mathbb{E}\left[\exp\left(\frac{\|\eta^k\|_2^2}{\sigma_k^2}\right) \mid \eta^0, \ldots, \eta^{k-1}\right] \leq \exp(1), \qquad \forall k \geq 0, \tag{67}$$

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

$$R_l \leq \frac{JR_0}{\sqrt{A_l}}, \quad \widetilde{R}_{l-1} \leq \frac{JR_0}{\sqrt{A_{l-1}}} \tag{68}$$

and

$$h\delta \sum_{k=0}^{l-1} \alpha_{k+1}(R_k + \widetilde{R}_k) + u \sum_{k=0}^{l-1} \alpha_{k+1} \langle \eta^k, a^k + \tilde{a}^k \rangle + c \sum_{k=0}^{l-1} \alpha_{k+1} \|\eta^k\|_2^2 \leq \left(2cHC + 2JD\left(hG + uC_1\sqrt{2HCg(N)}\right)\right) R_0^2 \tag{69}$$

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 7. Assume that the function ψ is μ_ψ-strongly convex and L_ψ-smooth,

$$r_k = \Theta\left(\max\left\{1, \left(\frac{\mu_\psi}{L_\psi}\right)^{3/2} \frac{N^2\sigma_\psi^2 \ln\frac{N}{\beta}}{\varepsilon}\right\}\right),$$

i.e. $r_k \geq \frac{1}{C}\max\left\{1, \left(\frac{\mu_\psi}{L_\psi}\right)^{3/2} \frac{N^2\sigma_\psi^2\left(1+\sqrt{3\ln\frac{N}{\beta}}\right)^2}{\varepsilon}\right\}$ with positive constants $C > 0, \varepsilon > 0$ and $N \geq 1$.

If additionally $\delta \leq \frac{GR_0}{N\sqrt{A_N}}$ and $\varepsilon \leq \frac{HR_0^2}{A_N}$ where $R_0 = \|y^* - y^0\|_2$ and Algorithm 7 is run for N iterations, then with probability at least $1 - 3\beta$

$$\|y^N - y^*\|_2^2 \leq \frac{\hat{J}^2R_0^2}{A_N}, \tag{70}$$

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

$$\hat{g}(N) = \frac{\ln\left(\frac{N}{\beta}\right) + \ln\ln\left(\frac{\hat{B}}{b}\right)}{\left(1 + \sqrt{3\ln\left(\frac{N}{\beta}\right)}\right)^2},\ b = \frac{2\sigma_1^2\alpha_1^2 R_0^2}{r_1},\ D \overset{(200)}{=} 1 + \frac{\mu_\psi}{L_\psi} + \sqrt{1 + \frac{\mu_\psi}{L_\psi}},$$

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

$$h = u = \frac{2}{\mu_\psi},\ c = \frac{2}{\mu_\psi^2},$$

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

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

$\hat{B}_1 = hG + uC_1\sqrt{2HC\left(\frac{L_\psi}{\mu_\psi}\right)^{3/2}\hat{g}(N)}$ 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 7 needs $N = \widetilde{O}\left(\sqrt{\frac{L_\psi}{\mu_\psi}}\right)$ iterations and $\widetilde{O}\left(\max\left\{\sqrt{\frac{L_\psi}{\mu_\psi}}, \frac{\sigma_\psi^2}{\varepsilon}\right\}\right)$ oracle calls where $\widetilde{O}(\cdot)$ hides polylogarithmic factors depending on $L_\psi, \mu_\psi, R_0, \varepsilon$ and β.

Next, we apply the `SSTM_sc` for the problem (16) when the objective of the primal problem (15) 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 5 but we additionally assume that the primal functional f has L-Lipschitz continuous gradient. As in Section 5 we also consider the case when the gradient of the dual functional is known only through biased stochastic estimators, see (32)–(39) and the paragraphs containing these formulas.

In Section 5 and 5.2 we mentioned that in the considered case dual function ψ is L_ψ-smooth on $\mathbb{R}^n$ and μ_ψ-strongly convex on $y^0 + (\text{Ker}A^\top)^\perp$ where $L_\psi = \lambda_{\max}(A^\top A)/\mu$ and $\mu_\psi = \lambda^+_{\min}(A^\top A)/L$. Using the same technique as in the proof of Theorem 5 we show next that w.l.o.g. one can assume that ψ is μ_ψ-strongly convex on $\mathbb{R}^n$ since $\tilde{\nabla}\Psi(y, \boldsymbol{\xi}^k)$ lies in $\text{Im}A = (\text{Ker}A^\top)^\perp$ by definition of $\tilde{\nabla}\Psi(y, \boldsymbol{\xi}^k)$. For this purposes we need the explicit formula for z^{k+1} which follows from the equation $\nabla\tilde{g}_{k+1}(z^{k+1}) = 0$:

$$z^{k+1} = \frac{z^0}{1 + A_{k+1}\mu_\psi} + \sum_{l=0}^{k+1}\frac{\alpha_l\mu_\psi}{1 + A_{k+1}\mu_\psi}\tilde{y}^l - \frac{1}{1 + A_{k+1}\mu_\psi}\sum_{l=0}^{k+1}\alpha_l\tilde{\nabla}\Psi(\tilde{y}^l, \boldsymbol{\xi}^l). \tag{71}$$

Theorem 8. For all $k \ge 0$ we have that the iterates of Algorithm 7 $\tilde{y}^k, z^k, y^k$ lie in $y^0 + \left(\text{Ker}(A^\top)\right)^\perp$.

Proof. We prove the statement of the theorem by induction. For $k = 0$ the statement is trivial, since $\tilde{y}^0 = z^0 = y^0$. Assume that for some $k \ge 0$ we have $\tilde{y}^t, z^t, y^t \in y^0 + \left(\text{Ker}(A^\top)\right)^\perp$ for all $0 \le t \le k$ and prove it for $k + 1$. Since $y_0 + \left(\text{Ker}(A^\top)\right)^\perp$ is a convex set and $\tilde{y}^{k+1}$ is a convex combination of y^k and z^k we have $\tilde{y}^{k+1} \in y^0 + \left(\text{Ker}(A^\top)\right)^\perp$. Next, the point $\frac{z^0}{1+A_{k+1}\mu_\psi} + \sum\limits_{l=0}^{k+1}\frac{\alpha_l\mu_\psi}{1+A_{k+1}\mu_\psi}\tilde{y}^l$ also lies

in $y^0 + \big(\text{Ker}(A^\top)\big)^\perp$ 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 $\tilde{\nabla}\Psi(\tilde{y}^l, \boldsymbol{\xi}^l)$ of we have that $\tilde{\nabla}\Psi(\tilde{y}^l, \boldsymbol{\xi}^l)$ lies in $\text{Im}A = (\text{Ker}A^\top)^\perp$ for all $\tilde{y}^l$. Putting all together and using (71) we get $z^{k+1} \in y^0 + \big(\text{Ker}(A^\top)\big)^\perp$. Finally, y^{k+1} lies in $y^0 + \big(\text{Ker}(A^\top)\big)^\perp$ as a convex combination of points from this set. □

This theorem makes it possible to apply the result from Theorem 7 for `SSTM_sc` which is run on the problem (16).

Corollary 4. Under assumptions of Theorem 7 we get that after $N = \widetilde{O}\left(\sqrt{\frac{L_\psi}{\mu_\psi}} \ln \frac{1}{\varepsilon}\right)$ iterations of Algorithm 7 which is run on the problem (16) with probability at least $1 - 3\beta$

$$\|\nabla \psi(y^N)\|_2 \leq \frac{\varepsilon}{R_y}, \tag{72}$$

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

$$\widetilde{O}\left(\max\left\{\sqrt{\frac{L_\psi}{\mu_\psi}}, \frac{\sigma_\psi^2 R_y^2}{\varepsilon^2}\right\}\right). \tag{73}$$

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

$$\|y^N - y^*\|_2 \leq \frac{\varepsilon}{\mu_\psi R_y}, \tag{74}$$

$$\|y^N\|_2 \leq 2R_y \tag{75}$$

Proof. Theorem 7 implies that with probability at least $1 - 3\beta$ we have

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

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

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

Since $A \overset{(199)}{\geq} \frac{1}{L_\psi}\left(1 + \frac{1}{2}\sqrt{\frac{\mu_\psi}{L_\psi}}\right)^{2k}$, it implies that after $N = \widetilde{O}\left(\sqrt{\frac{L_\psi}{\mu_\psi}} \ln \frac{1}{\varepsilon}\right)$ iterations of `SSTM_sc` we will get (72) with probability at least $1 - 3\beta$ and the number of oracle calls will be

$$\sum_{k=0}^{N} r_k = \widetilde{O}\left(\max\left\{\sqrt{\frac{L_\psi}{\mu_\psi}}, \frac{\sigma_\psi^2 R_y^2}{\varepsilon^2}\right\}\right).$$

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

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

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

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

□

Corollary 5. Let the assumptions of Theorem 7 hold. Assume that f is L_f-Lipschitz continuous on $B_{R_f}(0)$ where $R_f = \left(\sqrt{\frac{2C}{\lambda_{\max}(A^\top A)}} + G_1 + \frac{\sqrt{\lambda_{\max}(A^\top A)}}{\mu}\right)\frac{\varepsilon}{R_y} + R_x$, $R_x = \|x(A^\top y^*)\|_2$, $\varepsilon \leq \mu_\psi R_y^2$ and $\delta_y \leq \frac{G_1\varepsilon}{NR_y}$ for some positive constant G_1. Assume additionally that the last batch-size r_N is slightly bigger than other batch-sizes, i.e.

$$r_N \geq \frac{1}{C}\max\left\{1, \left(\frac{\mu_\psi}{L_\psi}\right)^{3/2}\frac{N^2\sigma_\psi^2\left(1+\sqrt{3\ln\frac{N}{\beta}}\right)^2 R_y^2}{\varepsilon^2}, \frac{\sigma_\psi^2\left(1+\sqrt{3\ln\frac{N}{\beta}}\right)^2 R_y^2}{\varepsilon^2}\right\}. \tag{76}$$

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

$$f(\tilde{x}^N) - f(x^*) \leq \left(2 + \left(\sqrt{\frac{2C}{\lambda_{\max}(A^\top A)}} + G_1\right)\frac{L_f}{R_y}\right)\varepsilon, \tag{77}$$

$$\|A\tilde{x}^N\|_2 \leq \left(1 + \sqrt{2C} + G_1\sqrt{\lambda_{\max}(A^\top A)}\right)\frac{\varepsilon}{R_y}, \tag{78}$$

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

$$\widetilde{O}\left(\max\left\{\sqrt{\frac{L_\psi}{\mu_\psi}}, \frac{\sigma_\psi^2 R_y^2}{\varepsilon^2}\right\}\right). \tag{79}$$

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 [12, 18, 25, 27] in order to accelerate the work of the method and (12)-(14) imply that

$$O\left(\frac{\sigma^2 R^2/\varepsilon^2}{\sqrt{LR^2/\varepsilon}}\right) \text{ or } O\left(\frac{\sigma^2/\mu\varepsilon}{\sqrt{L/\mu}\ln\left(\mu R^2/\varepsilon\right)}\right) \tag{80}$$

number of workers in such a parallel scheme gives the method with working time proportional to the number of iterations defined in (12). However, number of workers defined in (80) could be too big in order to use such an approach in practice. But still computing the batches in parallel even with much smaller could reduce the working time of the method if the communication is fast enough and it follows from (14).

Besides the computation of batches in parallel for the general type of problem (1)+(2), parallel optimization is often applied to the finite-sum minimization problems (1)+(3) or (1)+(4) that we rewrite here in the following form:

$$f(x) = \frac{1}{m}\sum_{k=1}^{m} f_k(x) \to \min_{x\in Q\subseteq\mathbb{R}^n}. \tag{81}$$

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 [66]. 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 (12).

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 (81). 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:

$$\overline{W}_{ij} = \begin{cases} -1, & \text{if } (i,j) \in E, \\ \deg(i), & \text{if } i = j, \\ 0 & \text{otherwise,} \end{cases} \tag{82}$$

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 $\mathbf{1}_m \overset{\text{def}}{=} (1,\ldots,1)^\top \in \mathbb{R}^m$ 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:

$$a_1 = \ldots = a_m \iff Wa = 0. \tag{83}$$

Now let us think about a_i as a number that i-th node stores. Then, using (83) 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 $W \overset{\text{def}}{=} \overline{W} \otimes I_n$ where $\otimes$ represents the Kronecker product (see (5)). 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 (83) implies

$$x_1 = \ldots = x_m \iff W\mathbf{x} = 0. \tag{84}$$

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 [66] and, as a consequence, $\sqrt{W}$ exists. Moreover, we can replace W by $\sqrt{W}$ in (84) and get the equivalent statement:

$$x_1 = \ldots = x_m \iff \sqrt{W}\mathbf{x} = 0. \tag{85}$$

Using this we can rewrite the problem (81) in the following way:

$$f(\mathbf{x}) = \frac{1}{m}\sum_{k=1}^m f_k(x_k) \to \min_{\substack{\sqrt{W}\mathbf{x}=0,\\ x_1,\ldots,x_m\in Q\subseteq\mathbb{R}^n}}. \tag{86}$$

We are interested in the general case when $f_k(x_k) = \mathbb{E}_{\xi_k}\left[f_k(x_k,\xi_k)\right]$ where $\{\xi_k\}_{k=1}^m$ are independent. This type of objective can be considered as a special case of (4). Then, as it was mentioned in the introduction it is natural to use stochastic gradients $\nabla f_k(x_k,\xi_k)$ that satisfy

$$\left\|\mathbb{E}_{\xi_k}\left[\nabla f_k(x_k,\xi_k)\right] - \nabla f_k(x_k)\right\|_2 \leq \delta, \tag{87}$$

$$\mathbb{E}_{\xi_k}\left[\exp\left(\frac{\left\|\nabla f_k(x_k,\xi_k) - \mathbb{E}_{\xi_k}\left[\nabla f_k(x_k,\xi_k)\right]\right\|_2^2}{\sigma^2}\right)\right] \leq \exp(1). \tag{88}$$

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 (39))

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

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

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 (86) has separable structure, it implies that f is L/m-smooth, μ/m-strongly convex and satisfies $\|\nabla f(\mathbf{x})\|_2\leq M/\sqrt{m}$ 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 (86) as (15) 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

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

Now it should become clear why in Section 4 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. For convenience, we summarize them in Tables 3 and 4 which are obtained via plugging the parameters that we obtained above in the bounds from Section 4. Note that the results presented in this match the lower bounds obtained in [5] in terms of the number of communication rounds up to logarithmic factors and and there is a conjecture [15] 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 [81].

Assumptions on f	Method	Citation	# of communication rounds	# of $\nabla_k f_k(x_k)$ oracle calls per node
μ-strongly convex, L-smooth	D-MASG, $Q = \mathbb{R}^n$	[20]	$\widetilde{O}\left(\sqrt{\frac{L}{\mu}}\chi\right)$	$\widetilde{O}\left(\sqrt{\frac{L}{\mu}}\right)$
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, $\|\nabla_k f_k(x_k)\|_2 \leq M$	R-Sliding	[15, 46, 48, 49]	$\widetilde{O}\left(\sqrt{\frac{M^2}{\mu\varepsilon}}\chi\right)$	$\widetilde{O}\left(\frac{M^2}{\mu\varepsilon}\right)$
$\|\nabla_k f_k(x_k)\|_2 \leq M$	Sliding	[46, 48, 49]	$O\left(\sqrt{\frac{M^2R^2}{\varepsilon^2}}\chi\right)$	$O\left(\frac{M^2R^2}{\varepsilon^2}\right)$

Table 3: Summary of the covered results in this paper for solving (86) using primal deterministic approach from Section 4. First column contains assumptions on f in addition to the convexity, $\chi \overset{\text{def}}{=} \chi(W)$. All methods except D-MASG should be applied to solve (21).

Assumptions on f	Method	Citation	# of communication rounds	# of $\nabla_k f_k(x_k, \xi_k)$ oracle calls per node
μ-strongly convex, L-smooth	D-MASG, in expectation, $Q = \mathbb{R}^n$	[20]	$\widetilde{O}\left(\sqrt{\frac{L}{\mu}}\chi\right)$	$\widetilde{O}\left(\max\left\{\sqrt{\frac{L}{\mu}}, \frac{\sigma^2}{\mu\varepsilon}\right\}\right)$
L-smooth	SSTP_IPS with STP as a subroutine, $Q = \mathbb{R}^n$	conjecture, [This paper] [15]	$\widetilde{O}\left(\sqrt{\frac{LR^2}{\varepsilon}}\chi\right)$	$\widetilde{O}\left(\max\left\{\sqrt{\frac{LR^2}{\varepsilon}}, \frac{\sigma^2R^2}{\varepsilon^2}\right\}\right)$
μ-strongly convex, $\|\nabla_k f_k(x_k)\|_2 \leq M$	RS-Sliding Q is bounded	[15, 46, 48, 49]	$\widetilde{O}\left(\sqrt{\frac{M^2}{\mu\varepsilon}}\chi\right)$	$\widetilde{O}\left(\frac{M^2+\sigma^2}{\mu\varepsilon}\right)$
$\|\nabla_k f_k(x_k)\|_2 \leq M$	S-Sliding Q is bounded	[46, 48, 49]	$\widetilde{O}\left(\sqrt{\frac{M^2R^2}{\varepsilon^2}}\chi\right)$	$\widetilde{O}\left(\frac{(M^2+\sigma^2)R^2}{\varepsilon^2}\right)$

Table 4: Summary of the covered results in this paper for solving (86) using primal stochastic approach from Section 4 with the stochastic oracle satisfying (87)-(88) with $\delta = 0$. First column contains assumptions on f in addition to the convexity, $\chi \overset{\text{def}}{=} \chi(W)$. All methods except D-MASG should be applied to solve (21). We notice it was also shown in [51] that the bounds from the last two rows hold even in the case when Q is unbounded, but in the expectation.

Finally, consider the situation when $Q = \mathbb{R}^n$ and each f_k from (86) is dual-friendly, i.e. one can construct dual problem for (86)

$$\Psi(\mathbf{y}) \quad \to \quad \min_{\mathbf{y}\in\mathbb{R}^{nm}}, \quad \text{where } \mathbf{y} = (y_1^\top, \ldots, y_m^\top)^\top \in \mathbb{R}^{nm},\ y_1, \ldots, y_m \in \mathbb{R}^n, \tag{89}$$

$$\varphi_k(y_k) \quad = \quad \max_{x_k\in\mathbb{R}^n} \left\{\langle y_k, x_k\rangle - f_k(x_k)\right\}, \tag{90}$$

$$\Phi(\mathbf{y}) \quad = \quad \frac{1}{m}\sum_{k=1}^m \varphi_k(my_k), \quad \Psi(\mathbf{y}) = \Phi(\sqrt{W}\mathbf{y}) = \frac{1}{m}\sum_{k=1}^m \varphi_k(m[\sqrt{W}\mathbf{x}]_k), \tag{91}$$

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

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

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

$$\|\mathbb{E}_{\xi_k}[\nabla\varphi_k(y_k,\xi_k)] - \nabla\varphi_k(y_k)\|_2 \le \delta_\varphi, \tag{92}$$

$$\mathbb{E}_{\xi_k}\left[\exp\left(\frac{\|\nabla\varphi_k(y_k,\xi_k) - \mathbb{E}_{\xi_k}[\nabla\varphi_k(y_k,\xi_k)]\|_2^2}{\sigma^2}\right)\right] \le \exp(1). \tag{93}$$

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

$$\varphi_k(y_k,\xi_k) = \max_{x_k\in\mathbb{R}^n}\{\langle y_k,x_k\rangle - f(x_k,\xi_k)\}. \tag{94}$$

Since

$$\nabla\Phi(\mathbf{y}) = \sum_{k=1}^m\nabla\varphi_k(my_k) \overset{(31)}{=} \sum_{k=1}^m x_k(my_k) \overset{\text{def}}{=} \mathbf{x}(\mathbf{y}), \quad x_k(y_k) \overset{\text{def}}{=} \operatorname*{argmax}_{x_k\in\mathbb{R}^n}\{\langle y_k,x_k\rangle - f_k(x_k)\}$$

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

$$\begin{aligned}\nabla\Phi(\mathbf{y},\xi) &\overset{\text{def}}{=} \nabla\Phi(\mathbf{y},\{\xi_k\}_{k=1}^m) \overset{\text{def}}{=} \sum_{k=1}^m\nabla\varphi_k(my_k,\xi_k) \overset{(31)}{=} \sum_{k=1}^m x_k(my_k,\xi_k) \overset{\text{def}}{=} \mathbf{x}(\mathbf{y},\xi),\\ x_k(y_k,\xi_k) &\overset{\text{def}}{=} \operatorname*{argmax}_{x_k\in\mathbb{R}^n}\{\langle y_k,x_k\rangle - f_k(x_k,\xi_k)\}.\end{aligned}$$

It satisfies (see also (39))

$$\begin{aligned}\|\mathbb{E}_\xi[\nabla\Phi(\mathbf{y},\xi)] - \nabla\Phi(\mathbf{y})\|_2 &\le \delta_\Phi,\\ \mathbb{E}_\xi\left[\exp\left(\frac{\|\nabla\Phi(\mathbf{y},\xi) - \mathbb{E}_\xi[\nabla\Phi(\mathbf{y},\xi)]\|_2^2}{\sigma_\Phi^2}\right)\right] &\le \exp(1)\end{aligned}$$

with $\delta_\Phi = m\delta_\varphi$ and $\sigma_\Phi^2 = O(m\sigma^2)$. Using this, we define the stochastic gradient of $\Psi(\mathbf{y})$ as $\nabla\Psi(\mathbf{y},\xi) \overset{\text{def}}{=} \sqrt{W}\nabla\Phi(\sqrt{W}\mathbf{y},\xi) = \sqrt{W}\mathbf{x}(\sqrt{W}\mathbf{y},\xi)$ and, as a consequence, we get

$$\begin{aligned}\|\mathbb{E}_\xi[\nabla\Psi(\mathbf{y},\xi)] - \nabla\Psi(\mathbf{y})\|_2 &\le \delta_\Psi,\\ \mathbb{E}_\xi\left[\exp\left(\frac{\|\nabla\Psi(\mathbf{y},\xi) - \mathbb{E}_\xi[\nabla\Psi(\mathbf{y},\xi)]\|_2^2}{\sigma_\Psi^2}\right)\right] &\le \exp(1)\end{aligned}$$

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

Assumptions on f	Method	Citation	# of communication rounds	# of $\nabla\varphi_k(y_k,\xi_k)$ oracle calls per node
μ-strongly convex, L-smooth, $\|\nabla f_k(x_k)\|_2 \leq M$ on $B_{R_M}(x^*)$	R-RRMA-AC-SA^2 (Algorithm 6), SSTM_SC (Algorithm 7)	Corollaries 3, [15], Corollary 5	$\widetilde{O}\left(\sqrt{\frac{L}{\mu}\chi}\right)$	$\widetilde{O}\left(\max\left\{\sqrt{\frac{L}{\mu}\chi}, \frac{\sigma_\Phi^2 M^2}{\varepsilon^2}\chi\right\}\right)$
μ-strongly convex, $\|\nabla f_k(x_k)\|_2 \leq M$ on $B_{R_M}(x^*)$	SPDSTM (Algorithm 2)	Theorem 2, [15, 16]	$\widetilde{O}\left(\sqrt{\frac{M^2}{\mu\varepsilon}\chi}\right)$	$\widetilde{O}\left(\max\left\{\sqrt{\frac{M^2}{\mu\varepsilon}\chi}, \frac{\sigma_\Phi^2 M^2}{\varepsilon^2}\chi\right\}\right)$

Table 5: Summary of the covered results in this paper for solving (89) using dual stochastic approach from Section 5 with the stochastic oracle satisfying (87)-(88) with $\delta = 0$. First column contains assumptions on f in addition to the convexity, $\chi \overset{\text{def}}{=} \chi(W)$.

Taking all of this into account we conclude that problem (89) is a special case of (16) with $A = \sqrt{W}$. To make the algorithms from Section 5 distributed we should change the variables in those methods via multiplying them by $\sqrt{W}$ from the left [15, 16, 78], e.g. for the iterates of SPDSTM we will get

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

which means that it is needed to multiply lines 4-6 of Algorithm 2 by $\sqrt{W}$ from the left. After such a change of variables all methods from Section 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 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 5 and 6. Note that the proposed bounds are optimal in terms of the number of communication rounds up to polylogarithmic factors [5, 66, 67, 68]. Note that the lower bounds from [66, 67, 68] 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 [15] 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.

7 Discussion

In this section we want to discuss some aspects of the proposed results that were not covered in previous sections. 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 [15]. From the line of works [43, 44,

Assumptions on f	Method	Citation	# of communication rounds	# of $\nabla\varphi_k(y_k,\xi_k)$ oracle calls per node
μ-strongly convex, L-smooth, $\|\nabla f_k(x_k)\|_2 \leq M$ on $B_{R_M}(x^*)$	SSTM_sc (Algorithm 7)	Corollary 5	$\widetilde{O}\left(\sqrt{\frac{L}{\mu}\chi}\right)$	$\widetilde{O}\left(\max\left\{\sqrt{\frac{L}{\mu}\chi}, \frac{\sigma_\Phi^2 M^2}{\varepsilon^2}\chi\right\}\right)$
μ-strongly convex, $\|\nabla f_k(x_k)\|_2 \leq M$ on $B_{R_M}(x^*)$	SPDSTM (Algorithm 2)	Theorem 2	$\widetilde{O}\left(\sqrt{\frac{M^2}{\mu\varepsilon}\chi}\right)$	$\widetilde{O}\left(\max\left\{\sqrt{\frac{M^2}{\mu\varepsilon}\chi}, \frac{\sigma_\Phi^2 M^2}{\varepsilon^2}\chi\right\}\right)$

Table 6: Summary of the covered results in this paper for solving (89) using **biased** dual stochastic approach from Section 5 with the stochastic oracle satisfying (87)-(88) with $\delta_\varphi > 0$. First column contains assumptions on f in addition to the convexity, $\chi \overset{\text{def}}{=} \chi(W)$. For both cases the noise level should satisfy $\delta_\varphi = \widetilde{O}\left(\varepsilon/M\sqrt{m\chi}\right)$.

45, 50] it is known that for the problem (1)+(4) (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

$$\widetilde{O}\left(m + \sqrt{m\frac{L}{\mu}} + \frac{\sigma^2}{\mu\varepsilon}\right). \tag{95}$$

The bad news is that (95) does not work with full parallelization trick and the best possible way to parallelize it is described in [50]. However, standard accelerated scheme using mini-batched versions of the stochastic gradients without variance-reduction technique and incremental oracles which gives the bound

$$\widetilde{O}\left(m\sqrt{\frac{L}{\mu}} + \frac{\sigma^2}{\mu\varepsilon}\right) \tag{96}$$

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

$$\widetilde{O}\left(\sqrt{\frac{L}{\mu}} + \frac{\sigma^2}{m\mu\varepsilon}\right) = \widetilde{O}\left(\max\left\{\sqrt{\frac{L}{\mu}}, \frac{\sigma^2}{m\mu\varepsilon}\right\}\right) \tag{97}$$

and the number of communication steps is

$$\widetilde{O}\left(d\sqrt{\frac{L}{\mu}}\right). \tag{98}$$

However, for the decentralized setup the second row of Table 4 states that the number of communication rounds is the same as in (98) up to substitution of d by $\sqrt{\chi(W)}$ and the number of oracle calls per node is

$$\widetilde{O}\left(\max\left\{\sqrt{\frac{L}{\mu}}, \frac{\sigma^2}{\mu\varepsilon}\right\}\right) \tag{99}$$

which has m times bigger statistical term under the maximum than in (97). 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 [61, 62], but it is not optimal in terms of the number of communications. Moreover, there is a hypothesis [15] that in the smooth case the bounds from Tables 3 and 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 5 was also presented in [15] as a hypothesis and in this paper we propose the same hypothesis for the result stated Table 6 up to polylogarithmic and additionally we hypothesise that the noise level that we obtained is also unimprovable up to polylogarithmic factors.

7.1 Possible Extensions

- As it was mentioned in Section 4, the recurrence technique that we use in Sections C and 5 can be very useful in the generalization of the results for `STM` from Section 4 for the case when instead of $\nabla f(x)$ only stochastic gradient $\nabla f(x, \xi)$ (see inequalities (8)-(9)) 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 [20]. For the convex but non-strongly convex case one can also try to combine Nesterov's smoothing technique [13, 56, 78] with `D-MASG` from [20].

- We believe that the technique presented in the proofs of Lemmas 4 and 6 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 (86) 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 $f(x) = {\|x\|_2^2}/{2m} = {1}/{m} \sum_{i=1}^m f_i(x)$, $f_i(x) = {x_i^2}/{2}$ where $L_i = 1$ for all i but f is ${1}/{d}$-smooth [77]. It was shown [66, 78] that instead of worst-case μ and L one can use $\bar{\mu} = {1}/{m} \sum_{i=1}^m \mu_i$ 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 [6, 30, 60, 79].

Acknowledgments. We would like to thank F. Bach, P. Dvurechensky, M. Gürbüzbalaban, A. Nemirovski, A. Olshevsky, N. Srebro, A. Taylor and C. Uribe for useful discussions. The work of E. Gorbunov was supported by RFBR 18-31-20005 mol-a-ved. The work of D. Dvinskikh was supported by Russian Science Foundation (project 18-71-10108). The work of A. Gasnikov was supported by RFBR 18-29-03071 mk and by Yahoo! Research Faculty Engagement Program.

Bibliography

[1] Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-efficient SGD via gradient quantization and encoding. In *Advances in Neural Information Processing Systems*, pages 1709–1720, 2017.

[2] Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2017, pages 1200–1205, New York, NY, USA, 2017. ACM. arXiv:1603.05953.

[3] Zeyuan Allen-Zhu. How to make the gradients small stochastically: Even faster convex and nonconvex sgd. In *Advances in Neural Information Processing Systems*, pages 1157–1167, 2018.

[4] A. S. Anikin, A. V. Gasnikov, P. E. Dvurechensky, A. I. Tyurin, and A. V. Chernov. Dual approaches to the minimization of strongly convex functionals with a simple structure under affine constraints. *Computational Mathematics and Mathematical Physics*, 57(8):1262–1276, Aug 2017.

[5] Yossi Arjevani and Ohad Shamir. Communication complexity of distributed convex learning and optimization. In *Advances in neural information processing systems*, pages 1756–1764, 2015.

[6] Necdet Serhat Aybat, Alireza Fallah, Mert Gurbuzbalaban, and Asuman Ozdaglar. A universally optimal multistage accelerated stochastic gradient method. *arXiv preprint arXiv:1901.08022*, 2019.

[7] Debraj Basu, Deepesh Data, Can Karakus, and Suhas Diggavi. Qsparse-local-sgd: Distributed sgd with quantization, sparsification, and local computations. *arXiv preprint arXiv:1906.02367*, 2019.

[8] Aaron Ben-Tal and Arkadi Nemirovski. *Lectures on Modern Convex Optimization.* Society for Industrial and Applied Mathematics, 2001.

[9] Dimitri P Bertsekas and John N Tsitsiklis. *Parallel and distributed computation: numerical methods*, volume 23. Prentice hall Englewood Cliffs, NJ, 1989.

[10] Alexey Chernov, Pavel Dvurechensky, and Alexander Gasnikov. Fast primal-dual gradient method for strongly convex minimization problems with linear constraints. In Yury Kochetov, Michael Khachay, Vladimir Beresnev, Evgeni Nurminski, and Panos Pardalos, editors, *Discrete Optimization and Operations Research: 9th International Conference, DOOR 2016, Vladivostok, Russia, September 19-23, 2016, Proceedings*, pages 391–403. Springer International Publishing, 2016.

[11] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, NIPS'14, pages 1646–1654, Cambridge, MA, USA, 2014. MIT Press.

[12] Olivier Devolder. *Exactness, inexactness and stochasticity in first-order methods for large-scale convex optimization*. PhD thesis, ICTEAM and CORE, Université Catholique de Louvain, 2013.

[13] Olivier Devolder, François Glineur, and Yurii Nesterov. Double smoothing technique for large-scale linearly constrained convex optimization. *SIAM Journal on Optimization*, 22(2):702–727, 2012.

[14] Olivier Devolder, François Glineur, Yurii Nesterov, et al. First-order methods with inexact oracle: the strongly convex case. 2013.

[15] Darina Dvinskikh and Alexander Gasnikov. Decentralized and parallelized primal and dual accelerated methods for stochastic convex programming problems. *arXiv preprint arXiv:1904.09015*, 2019.

[16] Darina Dvinskikh, Eduard Gorbunov, Alexander Gasnikov, Pavel Dvurechensky, and Cesar A Uribe. On dual approach for distributed stochastic convex optimization over networks. *arXiv preprint arXiv:1903.09844*, 2019.

[17] Pavel Dvurechenskii, Darina Dvinskikh, Alexander Gasnikov, Cesar Uribe, and Angelia Nedich. Decentralize and randomize: Faster algorithm for wasserstein barycenters. In *Advances in Neural Information Processing Systems*, pages 10760–10770, 2018.

[18] Pavel Dvurechensky and Alexander Gasnikov. Stochastic intermediate gradient method for convex problems with stochastic inexact oracle. *Journal of Optimization Theory and Applications*, 171(1):121–145, 2016.

[19] Pavel Dvurechensky, Alexander Gasnikov, and Alexander Tiurin. Randomized similar triangles method: A unifying framework for accelerated randomized optimization methods (coordinate descent, directional search, derivative-free method). *arXiv:1707.08486*, 2017.

[20] Alireza Fallah, Mert Gurbuzbalaban, Asu Ozdaglar, Umut Simsekli, and Lingjiong Zhu. Robust distributed accelerated stochastic gradient methods for multi-agent networks. *arXiv preprint arXiv:1910.08701*, 2019.

[21] Dylan Foster, Ayush Sekhari, Ohad Shamir, Nathan Srebro, Karthik Sridharan, and Blake Woodworth. The complexity of making the gradient small in stochastic convex optimization. *arXiv preprint arXiv:1902.04686*, 2019.

[22] Alexander Gasnikov. Universal gradient descent. *arXiv preprint arXiv:1711.00394*, 2017.

[23] Alexander Gasnikov. Universal gradient descent. *MIPT*, 2018.

[24] Alexander Gasnikov and Yurii Nesterov. Universal fast gradient method for stochastic composit optimization problems. *arXiv:1604.05275*, 2016.

[25] Alexander Vladimirovich Gasnikov and Yu E Nesterov. Universal method for stochastic composite optimization problems. *Computational Mathematics and Mathematical Physics*, 58(1):48–64, 2018.

[26] S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

[27] Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013. arXiv:1309.5549.

[28] Eduard Gorbunov, Pavel Dvurechensky, and Alexander Gasnikov. An accelerated method for derivative-free smooth stochastic convex optimization. *arXiv preprint arXiv:1802.09022*, 2018.

[29] Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. A unified theory of sgd: Variance reduction, sampling, quantization and coordinate descent. *arXiv preprint arXiv:1905.11261*, 2019.

[30] Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtarik. Sgd: General analysis and improved rates. *arXiv preprint arXiv:1901.09401*, 2019.

[31] Hadrien Hendrikx, Francis Bach, and Laurent Massoulié. Accelerated decentralized optimization with local updates for smooth and strongly convex objectives. *arXiv preprint arXiv:1810.02660*, 2018.

[32] Samuel Horvath, Chen-Yu Ho, Ludovit Horvath, Atal Narayan Sahu, Marco Canini, and Peter Richtarik. Natural compression for distributed deep learning. *arXiv preprint arXiv:1905.10988*, 2019.

[33] Samuel Horváth, Dmitry Kovalev, Konstantin Mishchenko, Sebastian Stich, and Peter Richtárik. Stochastic distributed learning with gradient quantization and variance reduction. *arXiv preprint arXiv:1904.05115*, 2019.

[34] Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan. A short note on concentration inequalities for random vectors with subgaussian norm. *arXiv preprint arXiv:1902.03736*, 2019.

[35] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323, 2013.

[36] A. Juditsky and A. Nemirovski. First order methods for non-smooth convex large-scale optimization, i: General purpose methods. In Stephen Wright Suvrit Sra, Sebastian Nowozin, editor, *Optimization for Machine Learning*, pages 121–184. Cambridge, MA: MIT Press, 2012.

[37] Anatoli Juditsky and Arkadii S Nemirovski. Large deviations of vector-valued martingales in 2-smooth normed spaces. *arXiv preprint arXiv:0809.0813*, 2008.

[38] Sham Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization. *Unpublished Manuscript, http://ttic. uchicago. edu/shai/papers/KakadeShalevTewari09.pdf*, 2(1), 2009.

[39] Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian U Stich, and Martin Jaggi. Error feedback fixes signsgd and other gradient compression schemes. *arXiv preprint arXiv:1901.09847*, 2019.

[40] Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Better communication complexity for local sgd. *arXiv preprint arXiv:1909.04746*, 2019.

[41] Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. First analysis of local gd on heterogeneous data. *arXiv preprint arXiv:1909.04715*, 2019.

[42] VM Kibardin. Decomposition into functions in the minimization problem. *Avtomatika i Telemekhanika*, (9):66–79, 1979.

[43] Andrei Kulunchakov and Julien Mairal. Estimate sequences for stochastic composite optimization: Variance reduction, acceleration, and robustness to noise. *arXiv preprint arXiv:1901.08788*, 2019.

[44] Andrei Kulunchakov and Julien Mairal. Estimate sequences for variance-reduced stochastic composite optimization. *arXiv preprint arXiv:1905.02374*, 2019.

[45] Andrei Kulunchakov and Julien Mairal. A generic acceleration framework for stochastic composite optimization. *arXiv preprint arXiv:1906.01164*, 2019.

[46] George Lan. Lectures on optimization methods for machine learning. *e-print*, 2019.

[47] Guanghui Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1):365–397, Jun 2012. Firs appeared in June 2008.

[48] Guanghui Lan. Gradient sliding for composite optimization. *Mathematical Programming*, 159(1):201–235, Sep 2016.

[49] Guanghui Lan, Soomin Lee, and Yi Zhou. Communication-efficient algorithms for decentralized and stochastic optimization. *Mathematical Programming*, pages 1–48, 2017.

[50] Guanghui Lan and Yi Zhou. Random gradient extrapolation for distributed and stochastic optimization. *SIAM Journal on Optimization*, 28(4):2753–2782, 2018.

[51] Guanghui Lan and Zhiqiang Zhou. Algorithms for stochastic optimization with expectation constraints. *arXiv:1604.03887*, 2016.

[52] Xiaorui Liu, Yao Li, Jiliang Tang, and Ming Yan. A double residual compression algorithm for efficient distributed learning. *arXiv preprint arXiv:1910.07561*, 2019.

[53] Konstantin Mishchenko, Eduard Gorbunov, Martin Takáč, and Peter Richtárik. Distributed learning with compressed gradient differences. *arXiv preprint arXiv:1901.09269*, 2019.

[54] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[55] Yurii Nesterov. *Introductory Lectures on Convex Optimization: a basic course*. Kluwer Academic Publishers, Massachusetts, 2004.

[56] Yurii Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.

[57] Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259, Aug 2009. First appeared in 2005 as CORE discussion paper 2005/67.

[58] Yurii Nesterov. How to make the gradients small. *Optima*, 88:10–11, 2012.

[59] Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

[60] Lam M Nguyen, Phuong Ha Nguyen, Marten van Dijk, Peter Richtárik, Katya Scheinberg, and Martin Takáč. Sgd and hogwild! convergence without the bounded gradients assumption. *arXiv preprint arXiv:1802.03801*, 2018.

[61] Alex Olshevsky, Ioannis Ch Paschalidis, and Shi Pu. Asymptotic network independence in distributed optimization for machine learning. *arXiv preprint arXiv:1906.12345*, 2019.

[62] Alex Olshevsky, Ioannis Ch Paschalidis, and Shi Pu. A non-asymptotic analysis of network independence for distributed stochastic gradient descent. *arXiv preprint arXiv:1906.02702*, 2019.

[63] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

[64] Ralph Tyrell Rockafellar. *Convex analysis*. Princeton university press, 2015.

[65] Alexander Rogozin, César A Uribe, Alexander Gasnikov, Nikolay Malkovsky, and Angelia Nedić. Optimal distributed optimization on slowly time-varying graphs. *arXiv preprint arXiv:1805.06045*, 2018.

[66] Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3027–3036. JMLR. org, 2017.

[67] Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal convergence rates for convex distributed optimization in networks. *Journal of Machine Learning Research*, 20(159):1–31, 2019.

[68] Kevin Scaman, Francis Bach, Sébastien Bubeck, Laurent Massoulié, and Yin Tat Lee. Optimal algorithms for non-smooth distributed optimization in networks. In *Advances in Neural Information Processing Systems*, pages 2745–2754, 2018.

[69] Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017.

[70] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[71] Shai Shalev-Shwartz and Tong Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 64–72, Bejing, China, 22–24 Jun 2014. PMLR. First appeared in arXiv:1309.2375.

[72] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming*. Society for Industrial and Applied Mathematics, 2009.

[73] Vladimir Spokoiny et al. Parametric estimation. finite sample theory. *The Annals of Statistics*, 40(6):2877–2909, 2012.

[74] Sebastian U Stich. Local sgd converges fast and communicates little. *arXiv preprint arXiv:1805.09767*, 2018.

[75] Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified sgd with memory. In *Advances in Neural Information Processing Systems*, pages 4447–4458, 2018.

[76] Fedor S Stonyakin, Darina Dvinskikh, Pavel Dvurechensky, Alexey Kroshnin, Olesya Kuznetsova, Artem Agafonov, Alexander Gasnikov, Alexander Tyurin, César A Uribe, Dmitry Pasechnyuk, et al. Gradient methods for problems with inexact model of the objective. In *International Conference on Mathematical Optimization Theory and Operations Research*, pages 97–114. Springer, 2019.

[77] Junqi Tang, Karen Egiazarian, Mohammad Golbabaee, and Mike Davies. The practicality of stochastic optimization in imaging inverse problems. *arXiv preprint arXiv:1910.10100*, 2019.

[78] César A Uribe, Soomin Lee, Alexander Gasnikov, and Angelia Nedić. Optimal algorithms for distributed optimization. *arXiv preprint arXiv:1712.00232*, 2017.

[79] Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1195–1204, 2019.

[80] Wei Wen, Cong Xu, Feng Yan, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Terngrad: Ternary gradients to reduce communication in distributed deep learning. In *Advances in Neural Information Processing Systems*, pages 1509–1519, 2017.

[81] Jinming Xu, Ye Tian, Ying Sun, and Gesualdo Scutari. Accelerated primal-dual algorithms for distributed smooth convex optimization over networks. *arXiv preprint arXiv:1910.10666*, 2019.

[82] Hao Yu, Rong Jin, and Sen Yang. On the linear speedup analysis of communication efficient momentum sgd for distributed non-convex optimization. *arXiv preprint arXiv:1905.03817*, 2019.

[83] Kaiwen Zhou. Direct acceleration of saga using sampled negative momentum. *arXiv preprint arXiv:1806.11048*, 2018.

[84] Kaiwen Zhou, Fanhua Shang, and James Cheng. A simple stochastic variance reduced algorithm with fast convergence rates. *arXiv preprint arXiv:1806.11027*, 2018.

A 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 $\lambda > 0$

$$|\langle a, b\rangle| \leq \frac{\|a\|_2^2}{2\lambda} + \frac{\lambda\|b\|_2^2}{2}. \tag{100}$$

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

$$\|a+b\|_2^2 \leq 2\|a\|_2^2 + 2\|b\|_2^2. \tag{101}$$

B Auxiliary Results

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

Lemma 7 (Lemma 2 from [34])**.** For random vector $\xi \in \mathbb{R}^n$ following statements are equivalent up to absolute constant difference in σ.

1 Tails: $\mathbb{P}\{\|\xi\|_2 \geq \gamma\} \leq 2\exp\left(-\frac{\gamma^2}{2\sigma^2}\right) \forall \gamma \geq 0.$

2 Moments: $\left(\mathbb{E}\left[\xi^p\right]\right)^{\frac{1}{p}} \leq \sigma\sqrt{p}$ for any positive integer p.

3 Super-exponential moment: $\mathbb{E}\left[\exp\left(\frac{\|\xi\|_2^2}{\sigma^2}\right)\right] \leq \exp(1).$

Lemma 8 (Corollary 8 from [34])**.** Let $\{\xi_k\}_{k=1}^N$ be a sequence of random vectors with values in $\mathbb{R}^n$ such that for $k = 1, \ldots, N$ and for all $\gamma \geq 0$

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

where σ_k^2 belongs to the filtration $\sigma(\xi_1, \ldots, \xi_{k-1})$ for all $k = 1, \ldots, N$. Let $S_N = \sum\limits_{k=1}^{N} \xi_k$. 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 9 (corollary of Theorem 2.1, item (ii) from [37])**.** Let $\{\xi_k\}_{k=1}^N$ be a sequence of random vectors with values in $\mathbb{R}^n$ such that

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

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

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

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

$$\mathbb{P}\left\{\|S_N\|_2 \geq \left(\sqrt{2} + \sqrt{2\gamma}\right) \sqrt{\sum_{k=1}^{N} \sigma_k^2}\right\} \leq \exp\left(-\frac{\gamma^2}{3}\right). \tag{102}$$

C Similar Triangles Method with Inexact Proximal Step

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

$$F(x) = f(x) + h(x) \to \min_{x \in \mathbb{R}^n}, \tag{103}$$

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 3. 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 $\hat{x} = \operatorname{argmin}^{\delta}_{x \in \mathbb{R}^n} g(x)$ if $g(\hat{x}) - g(x^*) \leq \delta$.

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 10 (See also Theorem 9 from [76])**.** Let $g(x)$ be convex, L-smooth, x^* is such that $g(x^*) = \min_{x \in \mathbb{R}^n} g(x)$ and $\hat{x} = \operatorname{argmin}^{\delta}_{x \in \mathbb{R}^n} g(x)$ for some $\delta > 0$. Then for all $x \in \mathbb{R}^n$

$$\langle \nabla g(\hat{x}), \hat{x} - x \rangle \leq \sqrt{2L\delta} \|\hat{x} - x\|_2. \tag{104}$$

Proof. Since x^* is a minimizer of $g(x)$ on $\mathbb{R}^n$, we have $\nabla g(x^*) = 0$ and [55]

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

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. □

Algorithm 8 Similar Triangles Methods with Inexact Proximal Step (STM_IPS)

Input: $\tilde{x}^0 = z^0 = x^0$ — 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: $\tilde{x}^{k+1} = {}^{(A_k x^k + \alpha_{k+1} z^k)}/_{A_{k+1}}$
5: $z^{k+1} = \operatorname{argmin}_{z\in\mathbb{R}^n}^{\delta_{k+1}} g_{k+1}(z)$, where $g_{k+1}(z)$ is defined in (105) 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**

Output: x^N

The main method of this section is stated as Algorithm 8. In the STM_IPS we use functions $g_{k+1}(z)$ which are defined as follows:

$$g_{k+1}(z) = \frac{1}{2}\|z^k - z\|_2^2 + \alpha_{k+1}\left(f(\tilde{x}^{k+1}) + \langle \nabla f(\tilde{x}^{k+1}), z - \tilde{x}^{k+1}\rangle + h(z)\right), \quad k = 0, 1, \ldots. \tag{105}$$

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 $T = O\left(\sqrt{{L_{k+1}}/{\mu_{k+1}}}\ln {L_{k+1}}/{\delta}\right)$ 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 $z^{k+1} = \operatorname{argmin}_{z\in\mathbb{R}^n}^{\delta_{k+1}} g_{k+1}(z)$. 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 11 (see also Theorem 1 from [17])**.** 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 8 we have

$$A_N\left(F(x^N) - F(x^*)\right) \;\leq\; \frac{1}{2}R_0^2 - \frac{1}{2}R_N^2 + \hat{\delta}\sum_{k=0}^{N-1}\sqrt{k+2}\widetilde{R}_{k+1}^2, \tag{106}$$

where x^* is the solution of (103) closest to the starting point z^0, $R_{k+1} \stackrel{\text{def}}{=} \|x^* - z^{k+1}\|_2$, $\widetilde{R}_0 \stackrel{\text{def}}{=} R_0 \stackrel{\text{def}}{=} \|x^* - z^0\|_2$, $\widetilde{R}_{k+1} \stackrel{\text{def}}{=} \max\{\widetilde{R}_k, R_{k+1}\}$ for $k = 0, 1, \ldots, N-1$ and $\hat{\delta} \stackrel{\text{def}}{=} \sqrt{\frac{(L_h+2L)\delta}{(1-\sqrt{2\delta})^2 L}}$.

Proof. First of all, we prove by induction that $\tilde{x}^{k+1}, x^k, z^k \in B_{\widetilde{R}_k}(x^*)$ 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 $\tilde{x}^1 = {}^{(A_0 x^0 + \alpha_{k+1} z^0)}/_{A_1} = z^0$, since $A_0 = \alpha_0 = 0$ and $A_1 = \alpha_1$. Next, assume that $\tilde{x}^{k+1}, x^k, z^k \in B_{\widetilde{R}_k}(x^*)$ for some $k \geq 0$. By definition of R_{k+1}

and $\widetilde{R}_{k+1}$ we have $z^{k+1} \in B_{R_{k+1}}(x^*) \subseteq B_{\widetilde{R}_{k+1}}(x^*)$. Due to the assumption that $x^k \in B_{R_k}(x^*) \subseteq B_{R_{k+1}}(x^*) \subseteq B_{\widetilde{R}_{k+1}}(x^*)$ 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, $\tilde{x}^{k+2}$ 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 $\tilde{x}^{k+1}, x^k, z^k \in B_{\widetilde{R}_k}(x^*)$ for all non-negative integers k.

Since $z^{k+1} = \operatorname{argmin}_{z \in \mathbb{R}^n}^{\delta_{k+1}} g_{k+1}(z)$ and $g_{k+1}(z)$ is 1-strongly convex and $(\alpha_{k+1}L_h + 1)$-smooth we can apply Lemma 10 and get

$$\langle \nabla g_{k+1}(z^{k+1}), z^{k+1} - x^* \rangle \leq \sqrt{2(\alpha_{k+1}L_h + 1)\delta \|z^k - \hat{z}^{k+1}\|_2^2} \cdot \|z^{k+1} - x^*\|_2. \tag{107}$$

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,

$$\|z^k - \hat{z}^{k+1}\|_2 \leq \frac{2}{1 - \sqrt{2\delta}} \widetilde{R}_{k+1}. \tag{108}$$

Applying inequality above and (196) for the r.h.s. of (107) we obtain

$$\langle z^{k+1} - z^k + \alpha_{k+1}\nabla f(\tilde{x}^{k+1}) + \alpha_{k+1}\nabla h(z^{k+1}), z^{k+1} - x^* \rangle \leq \hat{\delta}\sqrt{k+2}\widetilde{R}_{k+1}^2, \tag{109}$$

where we used

$$2\sqrt{\frac{2(\alpha_{k+1}L_h + 1)\delta}{(1 - \sqrt{2\delta})^2}} \overset{(196)}{\leq} 2\sqrt{\frac{2\left((k+2)L_h + 2(k+2)L\right)\delta}{2(1 - \sqrt{2\delta})^2 L}} \leq 2\sqrt{\frac{(L_h + 2L)\,\delta}{(1 - \sqrt{2\delta})^2 L}}\sqrt{k+2}$$

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

$$\begin{aligned}
\alpha_{k+1}\langle \nabla f(\tilde{x}^{k+1}), z^k - x^* \rangle &= \alpha_{k+1}\langle \nabla f(\tilde{x}^{k+1}), z^k - z^{k+1} \rangle + \alpha_{k+1}\langle \nabla f(\tilde{x}^{k+1}), z^{k+1} - x^* \rangle \\
&\overset{(109)}{\leq} \alpha_{k+1}\langle \nabla f(\tilde{x}^{k+1}), z^k - z^{k+1} \rangle + \langle z^{k+1} - z^k, x^* - z^{k+1} \rangle \\
&\quad + \alpha_{k+1}\langle \nabla h(z^{k+1}), x^* - z^{k+1} \rangle + \hat{\delta}\sqrt{k+2}\widetilde{R}_{k+1}^2.
\end{aligned}$$

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

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

By definition of x^{k+1} and $\tilde{x}^{k+1}$

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

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

$$\begin{aligned}
&\alpha_{k+1}\langle \nabla f(\tilde{x}^{k+1}), z^k - x^*\rangle \leq A_{k+1}\langle \nabla f(\tilde{x}^{k+1}), \tilde{x}^{k+1} - x^{k+1}\rangle \\
&-\frac{A_{k+1}^2}{2\alpha_{k+1}^2}\|\tilde{x}^{k+1} - x^{k+1}\|_2^2 + \frac{1}{2}\|z^k - x^*\|_2^2 \\
&\quad -\frac{1}{2}\|z^{k+1} - x^*\|_2^2 + \alpha_{k+1}\left(h(x^*) - h(z^{k+1})\right) + \hat{\delta}\sqrt{k+2}\widetilde{R}_{k+1}^2 \\
&\leq A_{k+1}\left(\langle \nabla f(\tilde{x}^{k+1}), \tilde{x}^{k+1} - x^{k+1}\rangle - \frac{2L}{2}\|\tilde{x}^{k+1} - x^{k+1}\|_2^2\right) + \frac{1}{2}\|z^k - x^*\|_2^2 \\
&\quad -\frac{1}{2}\|z^{k+1} - x^*\|_2^2 + \alpha_{k+1}\left(h(x^*) - h(z^{k+1})\right) + \hat{\delta}\sqrt{k+2}\widetilde{R}_{k+1}^2 \\
&\leq A_{k+1}(f(\tilde{x}^{k+1}) - f(x^{k+1})) + \frac{1}{2}\|z^k - x^*\|_2^2 - \frac{1}{2}\|z^{k+1} - x^*\|_2^2 \\
&\quad +\alpha_{k+1}\left(h(x^*) - h(z^{k+1})\right) + \hat{\delta}\sqrt{k+2}\widetilde{R}_{k+1}^2
\end{aligned} \tag{110}$$

From the convexity of f we get

$$\langle \nabla f(\tilde{x}^{k+1}), x^k - \tilde{x}^{k+1}\rangle \quad \leq \quad f(x^k) - f(\tilde{x}^{k+1}). \tag{111}$$

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

$$\alpha_{k+1}\left(\tilde{x}^{k+1} - z^k\right) = A_k\left(x^k - \tilde{x}^{k+1}\right). \tag{112}$$

Putting all together, we get

$$\begin{aligned}
\alpha_{k+1}\langle \nabla f(\tilde{x}^{k+1}), \tilde{x}^{k+1} - x^*\rangle \quad &= \quad \alpha_{k+1}\langle \nabla f(\tilde{x}^{k+1}), \tilde{x}^{k+1} - z^k\rangle + \alpha_{k+1}\langle \nabla f(\tilde{x}^{k+1}), z^k - x^*\rangle \\
&\overset{(112)}{=} \quad A_k\langle \nabla f(\tilde{x}^{k+1}), x^k - \tilde{x}^{k+1}\rangle + \alpha_{k+1}\langle \nabla f(\tilde{x}^{k+1}), z^k - x^*\rangle \\
&\overset{(110),(111)}{\leq} \quad A_k\left(f(x^k) - f(\tilde{x}^{k+1})\right) + A_{k+1}\left(f(\tilde{x}^{k+1}) - f(x^{k+1})\right) \\
&\qquad +\frac{1}{2}\|z^k - x^*\|_2^2 - \frac{1}{2}\|z^{k+1} - x^*\|_2^2 \\
&\qquad +\alpha_{k+1}\left(h(x^*) - h(z^{k+1})\right) + \hat{\delta}\sqrt{k+2}\widetilde{R}_{k+1}^2.
\end{aligned}$$

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

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

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

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

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. □

Below we state our main result of this section.

Theorem 9. Let $f(x)$ be convex and L-smooth, $h(x)$ be convex and L_h-smooth and $\delta \leq \frac{1}{4}$. Assume that for a given number of iterations $N \geq 1$ the number $\hat{\delta} \overset{\text{def}}{=} 2\sqrt{\frac{(L_h+2L)\delta}{(1-\sqrt{2\delta})^2 L}}$ satisfies $\hat{\delta} \leq \frac{C}{(N+1)^{3/2}}$ with some positive constant $C \in (0, 1/4)$. Then after N iteration of Algorithm 8 we have

$$F(x^N) - F(x^*) \leq \frac{3R_0^2}{2A_N}. \tag{113}$$

Proof. Lemma 11 implies that

$$A_l \left(F(x^l) - F(x^*)\right) \leq \frac{1}{2}R_0^2 - \frac{1}{2}R_l^2 + \hat{\delta}\sum_{k=0}^{l-1}\sqrt{k+2}\widetilde{R}_{k+1}^2 \tag{114}$$

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,

$$\widetilde{R}_l^2 \leq \widetilde{R}_0^2 + \frac{2C}{(N+1)^{3/2}}\sum_{k=0}^{l-1}(k+2)^{1/2}\widetilde{R}_{k+1}^2, \quad \forall l = 1, \ldots, N. \tag{115}$$

Using Lemma 15 we get that $\widetilde{R}_l \leq 2R_0^2$ for all $l = 1, \ldots, N$. We plug this inequality together with $\delta \leq \frac{C}{(N+1)^{3/2}} \leq \frac{1}{4(N+1)^{3/2}}$ and $R_N^2 \geq 0$ in (114) and get

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

which concludes the proof. □

Corollary 6. Under assumptions of Theorem 9 we get that for an arbitrary $\varepsilon > 0$ after

$$N = O\left(\sqrt{\frac{LR_0^2}{\varepsilon}}\right) \tag{116}$$

iterations of Algorithm 8 we have $F(x^N) - F(x^*) \leq \varepsilon$. Moreover, we get that δ should satisfy

$$\delta = O\left(\frac{L}{(L_h + L)N^3}\right). \tag{117}$$

Proof. The first part of the corollary follows from (113) and Lemma 12. Relation (117) follows from the definition of $\hat{\delta}$ and $\hat{\delta} \leq \frac{C}{(N+1)^{3/2}}$. Indeed, since $\hat{\delta} \overset{\text{def}}{=} 2\sqrt{\frac{(L_h+2L)\delta}{(1-\sqrt{2\delta})^2 L}}$ and $C \leq \frac{1}{4}$ we get that

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

□

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 $F(x^N) - \min_{x\in\mathbb{R}^n} F(x) \leq \varepsilon$.

D Missing Proofs from Section 4

D.1 Proof of Lemma 1

We have

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

From Demyanov–Danskin theorem [64] 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,

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

D.2 Proof of Theorem 1

By definition of F

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

which implies

$$f(x^N) - f(x^*) + \frac{R_y^2}{\varepsilon}\|Ax^N\|_2^2 \overset{(22)}{\leq} \varepsilon, \tag{118}$$

where x^* is an arbitrary solution of (15). Taking inequality $\|Ax^N\|_2^2 \geq 0$ into account we get the first part of (23). From Cauchy-Schwarz inequality we obtain

$$-R_y\|Ax^N\|_2 \leq \|y^*\|_2 \cdot \|Ax^N\|_2 \leq \langle y^*, Ax^N\rangle \overset{(20)}{\leq} f(x^N) - f(x^*).$$

Together with (118) it gives us quadratic inequality on $R_y\|Ax^N\|_2$:

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

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

E Missing Proofs from Section 5.1

E.1 Proof of Lemma 2

From the convexity of ψ we have

$$\begin{aligned}
\psi(x) - \psi(y) &\leq \langle \nabla\psi(x), x - y\rangle = \left\langle \mathbb{E}\left[\tilde{\nabla}\Psi(x, \boldsymbol{\xi}^k)\right], x - y\right\rangle \\
&+ \left\langle \nabla\psi(x) - \mathbb{E}\left[\tilde{\nabla}\Psi(x, \boldsymbol{\xi}^k)\right], x - y\right\rangle \\
&\leq \left\langle \mathbb{E}\left[\tilde{\nabla}\Psi(x, \boldsymbol{\xi}^k)\right], x - y\right\rangle + \left\|\nabla\psi(x) - \mathbb{E}\left[\tilde{\nabla}\Psi(x, \boldsymbol{\xi}^k)\right]\right\|_2 \cdot \|x - y\|_2 \\
&\overset{(38)}{\leq} \left\langle \mathbb{E}\left[\tilde{\nabla}\Psi(x, \boldsymbol{\xi}^k)\right], x - y\right\rangle + \delta\|x - y\|_2,
\end{aligned}$$

which proves the inequality (40). Applying L-smoothness of $\psi(x)$ we get

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

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

$$\begin{aligned}
\left\langle \nabla\psi(x) - \mathbb{E}\left[\tilde{\nabla}\Psi(x, \boldsymbol{\xi}^k)\right], y - x\right\rangle &\leq \frac{1}{2L}\left\|\nabla\psi(x) - \mathbb{E}\left[\tilde{\nabla}\Psi(x, \boldsymbol{\xi}^k)\right]\right\|_2^2 + \frac{L}{2}\|y - x\|_2^2 \\
&\overset{(38)}{\leq} \frac{\delta^2}{2L} + \frac{L}{2}\|y - x\|_2^2.
\end{aligned}$$

Combining these two inequalities we get (41).

E.2 Proof of Lemma 3

The proof of this lemma follows a similar way as in the proof of Theorem 1 from [17]. We can rewrite the update rule for z^k in the equivalent way:

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

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

$$\langle z^{k+1} - z^k + \alpha_{k+1}\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}), z - z^{k+1}\rangle \geq 0. \tag{119}$$

Using this we get

$$\begin{aligned}
&\alpha_{k+1}\langle \tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}), z^k - z\rangle \\
&\qquad = \alpha_{k+1}\langle \tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}), z^k - z^{k+1}\rangle + \alpha_{k+1}\langle \tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}), z^{k+1} - z\rangle \\
&\qquad \overset{(119)}{\leq} \alpha_{k+1}\langle \tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}), z^k - z^{k+1}\rangle + \langle z^{k+1} - z^k, z - z^{k+1}\rangle.
\end{aligned}$$

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

$$\begin{aligned}
\alpha_{k+1}\langle \tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}), z^k - z\rangle \quad \leq \quad & \alpha_{k+1}\langle \tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}), z^k - z^{k+1}\rangle - \frac{1}{2}\|z^k - z^{k+1}\|_2^2 \\
& + \frac{1}{2}\|z^k - z\|_2^2 - \frac{1}{2}\|z^{k+1} - z\|_2^2.
\end{aligned}$$

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

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

Together with previous inequality, it implies

$$\begin{aligned}
&\alpha_{k+1}\langle \tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}), z^k - z\rangle \leq A_{k+1}\langle \tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}), \tilde{y}^{k+1} - y^{k+1}\rangle \\
&- \frac{A_{k+1}^2}{2\alpha_{k+1}^2}\|\tilde{y}^{k+1} - y^{k+1}\|_2^2 \\
&\quad + \frac{1}{2}\|z^k - z\|_2^2 - \frac{1}{2}\|z^{k+1} - z\|_2^2 \\
\leq \quad & A_{k+1}\left(\langle \tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}), \tilde{y}^{k+1} - y^{k+1}\rangle - \frac{2\tilde{L}}{2}\|\tilde{y}^{k+1} - y^{k+1}\|_2^2\right) \\
&\quad + \frac{1}{2}\|z^k - z\|_2^2 - \frac{1}{2}\|z^{k+1} - z\|_2^2 \\
&= A_{k+1}\left(\left\langle \mathbb{E}_k\left[\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1})\right], \tilde{y}^{k+1} - y^{k+1}\right\rangle - \frac{2\tilde{L}}{2}\|\tilde{y}^{k+1} - y^{k+1}\|_2^2\right) \\
&\quad + A_{k+1}\left\langle \tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}) - \mathbb{E}_k\left[\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1})\right], \tilde{y}^{k+1} - y^{k+1}\right\rangle \\
&\quad + \frac{1}{2}\|z^k - z\|_2^2 - \frac{1}{2}\|z^{k+1} - z\|_2^2.
\end{aligned}$$

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

$$\left\langle \tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}) - \mathbb{E}_k\left[\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1})\right], \tilde{y}^{k+1} - y^{k+1}\right\rangle \\ \leq \frac{1}{2\tilde{L}}\left\|\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}) - \mathbb{E}_k\left[\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1})\right]\right\|_2^2 + \frac{\tilde{L}}{2}\|\tilde{y}^{k+1} - y^{k+1}\|_2^2.$$

Using this, we get

$$\begin{aligned}
&\alpha_{k+1}\langle \tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}), z^k - z\rangle \\
&\leq A_{k+1}\left(\left\langle \mathbb{E}_k\left[\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1})\right], \tilde{y}^{k+1} - y^{k+1}\right\rangle - \frac{\tilde{L}}{2}\|\tilde{y}^{k+1} - y^{k+1}\|_2^2\right) \\
&+\frac{A_{k+1}}{2\tilde{L}}\left\|\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}) - \mathbb{E}_k\left[\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1})\right]\right\|_2^2 \\
&+\frac{1}{2}\|z^k - z\|_2^2 - \frac{1}{2}\|z^{k+1} - z\|_2^2 \\
&\overset{(41)}{\leq} A_{k+1}\left(\psi(\tilde{y}^{k+1}) - \psi(y^{k+1}) + \frac{\delta^2}{\tilde{L}}\right) + \frac{1}{2}\|z^k - z\|_2^2 - \frac{1}{2}\|z^{k+1} - z\|_2^2 \\
&+\frac{A_{k+1}}{2\tilde{L}}\left\|\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}) - \mathbb{E}_k\left[\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1})\right]\right\|_2^2.
\end{aligned} \tag{120}$$

With Lemma 2 in hand, we have

$$\begin{aligned}
&\langle \tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}), y^k - \tilde{y}^{k+1}\rangle = \left\langle \mathbb{E}_k\left[\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1})\right], y^k - \tilde{y}^{k+1}\right\rangle \\
&\quad + \left\langle \tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}) - \mathbb{E}_k\left[\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1})\right], y^k - \tilde{y}^{k+1}\right\rangle \\
&\overset{(40)}{\leq} \psi(y^k) - \psi(\tilde{y}^{k+1}) + \delta\|y^k - \tilde{y}^{k+1}\|_2 \\
&\quad + \left\langle \tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}) - \mathbb{E}_k\left[\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1})\right], y^k - \tilde{y}^{k+1}\right\rangle.
\end{aligned} \tag{121}$$

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

$$\alpha_{k+1}\left(\tilde{y}^{k+1} - z^k\right) = A_k\left(y^k - \tilde{y}^{k+1}\right). \tag{122}$$

Putting all together, we get

$$\begin{aligned}
&\alpha_{k+1}\langle \tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}), \tilde{y}^{k+1} - z\rangle \\
&\quad = \alpha_{k+1}\langle \tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}), \tilde{y}^{k+1} - z^k\rangle + \alpha_{k+1}\langle \tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}), z^k - z\rangle \\
&\quad \overset{(122)}{=} A_k\langle \tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}), y^k - \tilde{y}^{k+1}\rangle + \alpha_{k+1}\langle \tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}), z^k - z\rangle \\
&\quad \overset{(120),(121)}{\leq} A_k\left(\psi(y^k) - \psi(\tilde{y}^{k+1}) + \delta\|y^k - \tilde{y}^{k+1}\|_2\right) \\
&\qquad + A_k\left\langle \tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}) - \mathbb{E}_k\left[\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1})\right], y^k - \tilde{y}^{k+1}\right\rangle \\
&\qquad + A_{k+1}\left(\psi(\tilde{y}^{k+1}) - \psi(y^{k+1}) + \tfrac{\delta^2}{\tilde{L}}\right) + \tfrac{1}{2}\|z^k - z\|_2^2 - \tfrac{1}{2}\|z^{k+1} - z\|_2^2 \\
&\qquad + \tfrac{A_{k+1}}{2\tilde{L}}\left\|\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}) - \mathbb{E}_k\left[\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1})\right]\right\|_2^2.
\end{aligned}$$

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

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

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

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

where we use that $A_0 = 0$.

E.3 Proof of Lemma 4

We start with applying Cauchy-Schwarz inequality to the second and the third terms in the right-hand side of (43):

$$
\begin{aligned}
\frac{1}{2}R_l^2 &\leq A + h\delta\sum_{k=0}^{l-1}\alpha_{k+1}\widetilde{R}_k + ud\sum_{k=0}^{l-1}\alpha_{k+1}\|\eta^k\|_2\widetilde{R}_k + c\sum_{k=0}^{l-1}\alpha_{k+1}^2\|\eta^k\|_2^2, \\
&\leq A + \frac{h^2\delta^2}{2}\sum_{k=0}^{l-1}\alpha_{k+1}^2 + \frac{ud+1}{2}\sum_{k=0}^{l-1}\widetilde{R}_k^2 + \left(c + \frac{ud}{2}\right)\sum_{k=0}^{l-1}\widetilde{\alpha}_{k+1}^2\|\eta^k\|_2^2. \qquad (123)
\end{aligned}
$$

The idea of the proof is as following: estimate R_N^2 roughly, then apply Lemma 8 in order to estimate second term in the last row of (43) and after that use the obtained recurrence to estimate right-hand side of (43).

Using Lemma 9 we get that with probability at least $1 - \frac{\beta}{N}$

$$\begin{aligned}\|\eta^k\|_2 &\leq \sqrt{2}\left(1+\sqrt{3\ln\frac{N}{\beta}}\right)\sigma_k \leq \sqrt{2}\left(1+\sqrt{3\ln\frac{N}{\beta}}\right)\frac{\sqrt{C\varepsilon}}{\sqrt{\widetilde{\alpha}_{k+1}\ln\left(\frac{N}{\beta}\right)}} \\ &= \left(\frac{1}{\sqrt{\widetilde{\alpha}_{k+1}\ln\left(\frac{N}{\beta}\right)}}+\sqrt{\frac{3}{\widetilde{\alpha}_{k+1}}}\right)\sqrt{2C\varepsilon} \leq 2\sqrt{\frac{3}{\widetilde{\alpha}_{k+1}}}\sqrt{2C\varepsilon}, \qquad (124)\end{aligned}$$

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

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

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 $\hat{l} \leq l$ 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 $\geq 1-\beta$

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

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

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

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

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

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 $\geq 1 - \beta$

$$\sum_{k=0}^{l-1}\widetilde{R}_k^2 \leq l\left(2A + (1+ud)\widetilde{R}_0^2 + 12CD\varepsilon\,(2c+ud)\,l(l+3) + h^2\delta^2D^2l(l+1)^2\right)(2(1+ud))^l, \tag{125}$$

for all $l = 1, \ldots, N$. Next we apply delicate result from [34] which is presented in Section B as Lemma 8. We consider random variables $\xi^k = \widetilde{\alpha}_{k+1}\langle \eta^k, a^k\rangle$. Note that $\mathbb{E}\left[\xi^k \mid \xi^0, \ldots, \xi^{k-1}\right] = \widetilde{\alpha}_{k+1}\left\langle \mathbb{E}\left[\eta^k \mid \eta^0, \ldots, \eta^{k-1}\right], a^k\right\rangle = 0$ and

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

due to Cauchy-Schwarz inequality and assumptions of the lemma. If we denote $\hat{\sigma}_k^2 = \sigma_k^2\widetilde{\alpha}_{k+1}^2d^2\widetilde{R}_k^2$ and apply Lemma 8 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 $b = \hat{\sigma}_0^2$, we get that for all $l = 1, \ldots, N$ with probability $\geq 1 - \frac{\beta}{N}$

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

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

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

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

$$\begin{aligned}\sum_{k=0}^{l-1}\hat{\sigma}_k^2 &= d^2\sum_{k=0}^{l-1}\sigma_k^2\widetilde{\alpha}_{k+1}^2\widetilde{R}_k^2 \leq d^2\sum_{k=0}^{l-1}\frac{C\varepsilon}{\ln\frac{N}{\beta}}\widetilde{\alpha}_{k+1}\widetilde{R}_k^2\\ &\leq \frac{d^2CDHR_0^2}{N^2\ln\frac{N}{\beta}}\sum_{k=0}^{l-1}(k+2)\widetilde{R}_k^2 \leq \frac{d^2CDHR_0^2}{3N}\cdot\frac{N+1}{N}\sum_{k=0}^{l-1}\widetilde{R}_k^2\\ &\overset{(125)}{\leq} \frac{d^2CDHR_0^2}{N}l\left(2A + (1+ud)\widetilde{R}_0^2 + 12CD\varepsilon\,(2c+ud)\,l(l+3) + h^2\delta^2D^2l(l+1)^2\right)\\ &\cdot\ (2(1+ud))^l\\ &\leq d^2CDHR_0^2\left(2A + (1+ud)R_0^2 + 48CDHR_0^2\,(2c+ud) + h^2G^2R_0^2D^2\right)(2(1+ud))^N\\ &= \frac{B}{2}\end{aligned}$$

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

$$\left|\sum_{k=0}^{l-1} \xi^k\right| \leq C_1 \sqrt{\sum_{k=0}^{l-1} \hat{\sigma}_k^2 \left(\ln\left(\frac{N}{\beta}\right) + \ln\ln\left(\frac{B}{b}\right)\right)} \tag{126}$$

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

Note that we also proved that (124) is in the same event together with (126) and holds with probability $\geq 1 - 2\beta$. Putting all together in (43), we get that with probability at least $1 - 2\beta$ the inequality

$$\begin{aligned}
\frac{1}{2}\widetilde{R}_l^2 &\overset{(43)}{\leq} A + h\delta \sum_{k=0}^{l-1} \alpha_{k+1}\widetilde{R}_k + u\sum_{k=0}^{l-1} \alpha_{k+1}\langle \eta^k, a^k\rangle + c\sum_{k=0}^{l-1} \alpha_{k+1}^2 \|\eta^k\|_2^2 \\
&\overset{(126)}{\leq} A + h\delta \sum_{k=0}^{l-1} \alpha_{k+1}\widetilde{R}_k + uC_1 \sqrt{\sum_{k=0}^{l-1} \hat{\sigma}_k^2 \left(\ln\left(\frac{N}{\beta}\right) + \ln\ln\left(\frac{B}{b}\right)\right)} + 24cC\varepsilon \sum_{k=0}^{l-1} \widetilde{\alpha}_{k+1}
\end{aligned}$$

holds for all $l = 1, \ldots, N$ simultaneously. For brevity, we introduce new notation(neglecting constant factor):

$$g(N) = \frac{\ln\left(\frac{N}{\beta}\right) + \ln\ln\left(\frac{B}{b}\right)}{\ln\left(\frac{N}{\beta}\right)} \approx 1.$$

Using our assumption $\sigma_k^2 \leq \frac{C\varepsilon}{\widetilde{\alpha}_{k+1}\ln\left(\frac{N}{\beta}\right)}$ and definition $\hat{\sigma}_k^2 = \sigma_k^2 \widetilde{\alpha}_{k+1}^2 d^2 \widetilde{R}_k^2$ we obtain that with probability at least $1 - 2\beta$ the inequality

$$\begin{aligned}
\frac{1}{2}\widetilde{R}_l^2 &\leq A + h\delta \sum_{k=0}^{l-1} \alpha_{k+1}\widetilde{R}_k + u\sum_{k=0}^{l-1} \alpha_{k+1}\langle \eta^k, a^k\rangle + c\sum_{k=0}^{l-1} \alpha_{k+1}^2 \|\eta^k\|_2^2 \\
&\leq A + \frac{hGDR_0}{(N+1)^2}\sum_{k=0}^{l-1}(k+2)\widetilde{R}_k + udC_1\sqrt{C\varepsilon g(N)}\sqrt{\sum_{k=0}^{l-1}\widetilde{\alpha}_{k+1}\widetilde{R}_k^2} + 24cC\varepsilon\sum_{k=0}^{l-1}\widetilde{\alpha}_{k+1} \\
&\leq A + \frac{hGDR_0}{(N+1)^2}\sum_{k=0}^{l-1}(k+2)\widetilde{R}_k + udC_1\sqrt{CD\varepsilon g(N)}\sqrt{\sum_{k=0}^{l-1}(k+2)\widetilde{R}_k^2} \\
&+ 24cCD\varepsilon\sum_{k=0}^{l-1}(k+2) \\
&\leq A + 24cCD\frac{HR_0^2}{N^2}\frac{l(l+1)}{2} \\
&+ \frac{hGDR_0}{(N+1)^2}\sum_{k=0}^{l-1}(k+2)\widetilde{R}_k + udC_1\sqrt{CD\frac{HR_0^2}{N^2}g(N)}\sqrt{\sum_{k=0}^{l-1}(k+2)\widetilde{R}_k^2} \\
&\leq \left(\frac{A}{R_0^2} + 24cCDH\right)R_0^2 + \frac{hGDR_0}{(N+1)^2}\sum_{k=0}^{l-1}(k+2)\widetilde{R}_k \\
&\quad + \frac{udC_1R_0}{N}\sqrt{CDHg(N)}\sqrt{\sum_{k=0}^{l-1}(k+2)\widetilde{R}_k^2}
\end{aligned} \tag{127}$$

holds for all $l = 1, \ldots, N$ simultaneously. Next we apply Lemma 14 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

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

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

E.4 Proof of Theorem 2

Lemma 3 states that

$$\begin{aligned}
A_N\psi(y^N) \;\leq\; &\frac{1}{2}\|\tilde{y} - z^0\|_2^2 - \frac{1}{2}\|\tilde{y} - z^N\|_2^2 + \sum_{k=0}^{N-1}\alpha_{k+1}\left(\psi(\tilde{y}^{k+1}) + \langle\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}), \tilde{y} - \tilde{y}^{k+1}\rangle\right) \\
&+ \sum_{k=0}^{N-1}A_k\left\langle\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}) - \mathbb{E}_k\left[\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1})\right], y^k - \tilde{y}^{k+1}\right\rangle \\
&+ \sum_{k=0}^{N-1}\frac{A_{k+1}}{2\tilde{L}}\left\|\mathbb{E}_k\left[\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1})\right] - \tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1})\right\|_2^2 \\
&+ \delta\sum_{k=0}^{N-1}A_k\|y^k - \tilde{y}^{k+1}\|_2 + \delta^2\sum_{k=0}^{N-1}\frac{A_{k+1}}{\tilde{L}},
\end{aligned} \tag{128}$$

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

$$\alpha_{k+1}\left(\tilde{y}^{k+1} - z^k\right) = A_k\left(y^k - \tilde{y}^{k+1}\right). \tag{129}$$

Using this, we add and subtract $\sum_{k=0}^{N-1}\alpha_{k+1}\left\langle \mathbb{E}_k\left[\tilde{\nabla}\Psi(\tilde{y}^{k+1},\boldsymbol{\xi}^{k+1})\right],\tilde{y}^*-\tilde{y}^{k+1}\right\rangle$ in (128), and obtain the following inequality by choosing $\tilde{y}=\tilde{y}^*$ — the minimizer of $\psi(y)$:

$$\begin{aligned}A_N\psi(y^N) \;\leq\; & \frac{1}{2}\|\tilde{y}^*-z^0\|_2^2-\frac{1}{2}\|\tilde{y}^*-z^N\|_2^2\\ &+\sum_{k=0}^{N-1}\alpha_{k+1}\left(\psi(\tilde{y}^{k+1})+\left\langle \mathbb{E}_k\left[\tilde{\nabla}\Psi(\tilde{y}^{k+1},\boldsymbol{\xi}^{k+1})\right],\tilde{y}^*-\tilde{y}^{k+1}\right\rangle\right)\\ &+\sum_{k=0}^{N-1}\alpha_{k+1}\left\langle \tilde{\nabla}\Psi(\tilde{y}^{k+1},\boldsymbol{\xi}^{k+1})-\mathbb{E}_k\left[\tilde{\nabla}\Psi(\tilde{y}^{k+1},\boldsymbol{\xi}^{k+1})\right],\mathbf{a}^k\right\rangle\\ &+\sum_{k=0}^{N-1}\alpha_{k+1}^2\left\|\tilde{\nabla}\Psi(\tilde{y}^{k+1},\boldsymbol{\xi}^{k+1})-\mathbb{E}_k\left[\tilde{\nabla}\Psi(\tilde{y}^{k+1},\boldsymbol{\xi}^{k+1})\right]\right\|_2^2\\ &+\delta\sum_{k=0}^{N-1}\alpha_{k+1}\|\tilde{y}^{k+1}-z^k\|_2+\delta^2\sum_{k=0}^{N-1}\frac{A_{k+1}}{\tilde{L}},\end{aligned}\tag{130}$$

where $\mathbf{a}^k=\tilde{y}^*-z^k$. From (40) we have

$$\begin{aligned}&\sum_{k=0}^{N-1}\alpha_{k+1}\left(\psi(\tilde{y}^{k+1})+\left\langle \mathbb{E}_k\left[\tilde{\nabla}\Psi(\tilde{y}^{k+1},\boldsymbol{\xi}^{k+1})\right],\tilde{y}^*-\tilde{y}^{k+1}\right\rangle\right)\\ &\qquad\overset{(40)}{\leq}\sum_{k=0}^{N-1}\alpha_{k+1}\left(\psi(\tilde{y}^{k+1})+\psi(\tilde{y}^*)-\psi(\tilde{y}^{k+1})+\delta\|\tilde{y}^{k+1}-\tilde{y}^*\|_2\right)\\ &\qquad=\sum_{k=0}^{N-1}\alpha_{k+1}\left(\psi(\tilde{y}^*)+\delta\|\tilde{y}^{k+1}-\tilde{y}^*\|_2\right)\\ &\qquad=A_N\psi(\tilde{y}^*)+\delta\sum_{k=0}^{N-1}\alpha_{k+1}\|\tilde{y}^{k+1}-\tilde{y}^*\|_2\\ &\qquad\leq A_N\psi(y^N)+\delta\sum_{k=0}^{N-1}\alpha_{k+1}\|\tilde{y}^{k+1}-\tilde{y}^*\|_2\end{aligned}$$

From this and (130) we get

$$\begin{aligned}\frac{1}{2}\|\tilde{y}^*-z^N\|_2^2 \;\overset{(130)}{\leq}\; & \frac{1}{2}\|\tilde{y}^*-z^0\|_2^2+\delta^2\sum_{k=0}^{N-1}\frac{A_{k+1}}{\tilde{L}}+\delta\sum_{k=0}^{N-1}\alpha_{k+1}\left(\|\tilde{y}^{k+1}-z^k\|_2+\|\tilde{y}^{k+1}-\tilde{y}^*\|_2\right)\\ &+\sum_{k=0}^{N-1}\alpha_{k+1}\left\langle \tilde{\nabla}\Psi(\tilde{y}^{k+1},\boldsymbol{\xi}^{k+1})-\mathbb{E}_k\left[\tilde{\nabla}\Psi(\tilde{y}^{k+1},\boldsymbol{\xi}^{k+1})\right],\mathbf{a}^k\right\rangle\\ &+\sum_{k=0}^{N-1}\alpha_{k+1}^2\left\|\tilde{\nabla}\Psi(\tilde{y}^{k+1},\boldsymbol{\xi}^{k+1})-\mathbb{E}_k\left[\tilde{\nabla}\Psi(\tilde{y}^{k+1},\boldsymbol{\xi}^{k+1})\right]\right\|_2^2.\end{aligned}\tag{131}$$

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

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

Since in Algorithm 2 we choose $z^0=0$, then $R_0=R_y$. One can obtain by induction that $\forall k\geq 0$ we have $\tilde{y}^{k+1},y^k,z^k\in B_{\widetilde{R}_k}(\tilde{y}^*)$, where $B_{\widetilde{R}_k}(\tilde{y}^*)$ is Euclidean ball with radius $\widetilde{R}_k$ at centre $\tilde{y}^*$.

Indeed, since from lines 2 and 5 of Algorithm 2 y_{k+1} is a convex combination of $z_{k+1} \in B_{R_{k+1}}(\tilde{y}^*) \subseteq B_{\widetilde{R}_{k+1}}(\tilde{y}^*)$ and $y^k \in B_{\widetilde{R}_k}(\tilde{y}^*) \subseteq B_{\widetilde{R}_{k+1}}(\tilde{y}^*)$, where we use the fact that a ball is a convex set, we get $y^{k+1} \in B_{\widetilde{R}_{k+1}}(\tilde{y}^*)$. Analogously, since from lines 2 and 3 of Algorithm 2 $\tilde{y}^{k+1}$ is a convex combination of y^k and z^k we have $\tilde{y}^{k+1} \in B_{\widetilde{R}_k}(\tilde{y}^*)$. It implies that

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

Using new notation we can rewrite (131) as

$$\begin{aligned}\frac{1}{2}R_N^2 \quad \leq \quad & \frac{1}{2}R_0^2 + \delta^2 \sum_{k=0}^{N-1} \frac{A_{k+1}}{\tilde{L}} + 3\delta \sum_{k=0}^{N-1} \alpha_{k+1}\widetilde{R}_k \\ & + \sum_{k=0}^{N-1} \alpha_{k+1} \left\langle \tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}) - \mathbb{E}_k\left[\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1})\right], \mathbf{a}^k \right\rangle \\ & + \sum_{k=0}^{N-1} \alpha_{k+1}^2 \left\| \tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}) - \mathbb{E}_k\left[\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1})\right] \right\|_2^2, \end{aligned} \tag{132}$$

where $\|\mathbf{a}^k\|_2 = \|\tilde{y}^* - z^k\|_2 \leq \widetilde{R}_k$. Note that (132) holds for all $N \geq 1$.

Let us denote $\eta^k = \tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}) - \mathbb{E}_k\left[\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1})\right]$. Theorem 2.1 from [37] (see Lemma 9 in the Section B) says that

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

Using this and Lemma 2 from [34] (see Lemma 7 in the Section B) we get that

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

where $\sigma_k^2 \leq \frac{\widetilde{C}\sigma_\psi^2}{r_{k+1}} \leq \frac{C\varepsilon}{\widetilde{\alpha}_{k+1}\ln(\frac{N}{\delta})}$, $\widetilde{C}$ and $C = \widetilde{C} \cdot \hat{C}$ are some positive constants. From (196) we have that $\alpha_{k+1} \leq \widetilde{\alpha}_{k+1} = \frac{k+2}{2\tilde{L}}$. Moreover, $\mathbf{a}^k$ depends only on $\eta^0, \ldots, \eta^{k-1}$. Putting all together in (132) and changing the indices we get that for all $l = 1, \ldots, N$

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

Next we apply Lemma 4 with the constants $A = \frac{1}{2}R_0^2 + \delta^2 \sum\limits_{k=0}^{N-1} \frac{A_{k+1}}{\tilde{L}}, h = 3, u = 1, c = 1, D = \frac{1}{2\tilde{L}}, d = 1, \varepsilon \leq \frac{H\tilde{L}R_0^2}{N^2}$ and $\delta \leq \frac{G\tilde{L}R_0}{(N+1)^3}$, and get that with probability at least $1 - 2\beta$ the inequalities

$$\widetilde{R}_l \leq JR_0 \tag{133}$$

and

$$\sum_{k=0}^{l-1} \alpha_{k+1}\langle \eta^k, \mathbf{a}^k\rangle + \sum_{k=0}^{l-1} \alpha_{k+1}^2 \|\eta^k\|_2^2 \leq \left(12CH + \frac{3GJ}{2} + C_1\sqrt{\frac{CHJg(N)}{2}}\right) R_0^2 \tag{134}$$

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

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

To estimate the duality gap we need again refer to (128). Since $\tilde{y}$ is chosen arbitrary we can take the minimum in $\tilde{y}$ over the set $B_{2R_y}(0) = \{\tilde{y} : \|\tilde{y}\|_2 \leq 2R_y\}$:

$$\begin{aligned}
A_N\psi(y^N) \;\leq\; & \min_{\tilde{y}\in B_{2R_y}(0)}\left\{\frac{1}{2}\|\tilde{y}-z^0\|_2^2 + \sum_{k=0}^{N-1}\alpha_{k+1}\left(\psi(\tilde{y}^{k+1}) + \left\langle\tilde{\nabla}\Psi(\tilde{y}^{k+1},\boldsymbol{\xi}^{k+1}),\tilde{y}-\tilde{y}^{k+1}\right\rangle\right)\right\} \\
& + \sum_{k=0}^{N-1}A_k\left\langle\tilde{\nabla}\Psi(\tilde{y}^{k+1},\boldsymbol{\xi}^{k+1}) - \mathbb{E}_k\left[\tilde{\nabla}\Psi(\tilde{y}^{k+1},\boldsymbol{\xi}^{k+1})\right], y^k-\tilde{y}^{k+1}\right\rangle \\
& + \sum_{k=0}^{N-1}\frac{A_{k+1}}{2\tilde{L}}\left\|\mathbb{E}_k\left[\tilde{\nabla}\Psi(\tilde{y}^{k+1},\boldsymbol{\xi}^{k+1})\right] - \tilde{\nabla}\Psi(\tilde{y}^{k+1},\boldsymbol{\xi}^{k+1})\right\|_2^2 \\
& + \delta\sum_{k=0}^{N-1}A_k\|y^k-\tilde{y}^{k+1}\|_2 + \delta^2\sum_{k=0}^{N-1}\frac{A_{k+1}}{\tilde{L}} \\
\leq\; & 2R_y^2 + \min_{\tilde{y}\in B_{2R_y}(0)}\sum_{k=0}^{N-1}\alpha_{k+1}\left(\psi(\tilde{y}^{k+1}) + \left\langle\tilde{\nabla}\Psi(\tilde{y}^{k+1},\boldsymbol{\xi}^{k+1}),\tilde{y}-\tilde{y}^{k+1}\right\rangle\right) \\
& + \sum_{k=0}^{N-1}A_k\left\langle\tilde{\nabla}\Psi(\tilde{y}^{k+1},\boldsymbol{\xi}^{k+1}) - \mathbb{E}_k\left[\tilde{\nabla}\Psi(\tilde{y}^{k+1},\boldsymbol{\xi}^{k+1})\right], y^k-\tilde{y}^{k+1}\right\rangle \\
& + \sum_{k=0}^{N-1}\frac{A_{k+1}}{2\tilde{L}}\left\|\mathbb{E}_k\left[\tilde{\nabla}\Psi(\tilde{y}^{k+1},\boldsymbol{\xi}^{k+1})\right] - \tilde{\nabla}\Psi(\tilde{y}^{k+1},\boldsymbol{\xi}^{k+1})\right\|_2^2 \\
& + \delta\sum_{k=0}^{N-1}A_k\|y^k-\tilde{y}^{k+1}\|_2 + \delta^2\sum_{k=0}^{N-1}\frac{A_{k+1}}{\tilde{L}}, \qquad (135)
\end{aligned}$$

where we also used $\frac{1}{2}\|\tilde{y}-z^N\|_2^2 \geq 0$ and $z^0 = 0$. By adding and subtracting $\sum_{k=0}^{N-1}\alpha_{k+1}\left\langle\mathbb{E}_k\left[\tilde{\nabla}\Psi(\tilde{y}^{k+1},\boldsymbol{\xi}^{k+1})\right],\tilde{y}-\tilde{y}^{k+1}\right\rangle$ under the minimum in (135) we obtain

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

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

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

Putting all together in (135) and using (129) and line 2 from Algorithm 2 we get

$$\begin{aligned}
A_N \psi(y^N) \;\le\; & 2R_y^2 + \min_{\tilde{y} \in B_{2R_y}(0)} \sum_{k=0}^{N-1} \alpha_{k+1} \left(\psi(\tilde{y}^{k+1}) + \left\langle \mathbb{E}_k \left[\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}) \right], \tilde{y} - \tilde{y}^{k+1} \right\rangle \right) \\
& + 2 \max_{\tilde{y} \in B_{2R_y}(0)} \sum_{k=0}^{N-1} \alpha_{k+1} \left\langle \tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}) - \mathbb{E}_k \left[\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}) \right], \tilde{y} \right\rangle \\
& + \sum_{k=0}^{N-1} \alpha_{k+1} \left\langle \tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}) - \mathbb{E}_k \left[\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}) \right], \mathbf{a}^k \right\rangle \\
& + \sum_{k=0}^{N-1} \alpha_{k+1}^2 \left\| \tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}) - \mathbb{E}_k \left[\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}) \right] \right\|_2^2 \\
& + \delta \sum_{k=0}^{N-1} \alpha_{k+1} \|\tilde{y}^{k+1} - z^k\|_2 + \delta^2 \sum_{k=0}^{N-1} \frac{A_{k+1}}{\tilde{L}},
\end{aligned} \tag{136}$$

where $\mathbf{a}^k = \tilde{y}^* - z^k$. From (133) and (134) we have that with probability at least $1 - 2\beta$ the following inequality holds:

$$\begin{aligned}
A_N \psi(y^N) \;\le\; & \min_{\tilde{y} \in B_{2R_y}(0)} \sum_{k=0}^{N-1} \alpha_{k+1} \left(\psi(\tilde{y}^{k+1}) + \left\langle \mathbb{E}_k \left[\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}) \right], \tilde{y} - \tilde{y}^{k+1} \right\rangle \right) \\
& + 2 \max_{\tilde{y} \in B_{2R_y}(0)} \sum_{k=0}^{N-1} \alpha_{k+1} \left\langle \tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}) - \mathbb{E}_k \left[\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}) \right], \tilde{y} \right\rangle \\
& + 2R_y^2 + \left(12CH + \frac{5GJ}{2} + \frac{G^2}{2(N+1)} + C_1 \sqrt{\frac{CHJg(N)}{2}} \right) R_0^2,
\end{aligned} \tag{137}$$

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

$$\begin{aligned}
\delta \sum_{k=0}^{N-1} \alpha_{k+1} \|\tilde{y}^{k+1} - z^k\|_2 \;&\le\; 2\delta J R_0 \sum_{k=0}^{N-1} \alpha_{k+1} \le \frac{2G\tilde{L}R_0^2 J}{(N+1)^2} \frac{1}{2\tilde{L}} \sum_{k=0}^{N-1} (k+2) \le GJR_0^2, \\
\delta^2 \sum_{k=0}^{N-1} \frac{A_{k+1}}{\tilde{L}} \;&\le\; \frac{G^2 \tilde{L}^2 R_0^2}{(N+1)^4} \sum_{k=0}^{N-1} \frac{(k+2)^2}{2\tilde{L}^2} \le \frac{G^2 R_0^2}{2(N+1)}
\end{aligned}$$

By the definition of the norm we get

$$\max_{\tilde{y}\in B_{2R_y}(0)} \sum_{k=0}^{N-1} \alpha_{k+1} \left\langle \tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}) - \mathbb{E}_k\left[\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1})\right], \tilde{y}\right\rangle \\ \leq 2R_y \left\| \sum_{k=0}^{N-1} \alpha_{k+1} \left(\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}) - \mathbb{E}_k\left[\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1})\right]\right)\right\|_2. \quad (138)$$

Next we apply Lemma 9 to the right-hand side of the previous inequality and get

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

Since $N^2 \leq \frac{H\tilde{L}R_0^2}{\varepsilon}$ and $r_k = \Omega\left(\max\left\{1, \frac{\sigma_\psi^2 \alpha_k \ln(N/\beta)}{\varepsilon}\right\}\right)$ one can choose such $C_2 > 0$ that $\frac{\sigma_\psi^2}{r_k} \leq \frac{C_2\varepsilon}{\alpha_k \ln\left(\frac{N}{\beta}\right)} \leq \frac{H\tilde{L}C_2R_0^2}{\alpha_k N^2 \ln\left(\frac{N}{\beta}\right)}$. Moreover, let us choose γ such that $\exp\left(-\frac{\gamma^2}{3}\right) = \beta \implies \gamma = \sqrt{3\ln\frac{1}{\beta}}$. From this we get that with probability at least $1-\beta$

$$\left\| \sum_{k=0}^{N-1} \alpha_{k+1} \left(\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}) - \mathbb{E}_k\left[\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1})\right]\right)\right\|_2 \\ \leq \sqrt{2}\left(1 + \sqrt{\ln\frac{1}{\beta}}\right) R_y \sqrt{\frac{H\tilde{L}C_2}{\ln\left(\frac{N}{\beta}\right)}} \sqrt{\sum_{k=0}^{N-1} \frac{\alpha_{k+1}}{N^2}} \\ \overset{(196)}{\leq} 2\sqrt{2} R_y \sqrt{H\tilde{L}C_2} \sqrt{\sum_{k=0}^{N-1} \frac{k+2}{2\tilde{L}N^2}} = 2R_y\sqrt{HC_2}\sqrt{\frac{N(N+3)}{N^2}} \leq 4R_y\sqrt{HC_2}. \quad (139)$$

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 $1-3\beta$

$$A_N\psi(y^N) \overset{(137)+(138)+(139)}{\leq} \min_{\tilde{y}\in B_{2R_y}(0)} \sum_{k=0}^{N-1} \alpha_{k+1}\left(\psi(\tilde{y}^{k+1}) + \left\langle \mathbb{E}_k\left[\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1})\right], \tilde{y} - \tilde{y}^{k+1}\right\rangle\right) \\ + \left(8\sqrt{HC_2} + 2 + 12CH + \frac{5GJ}{2} + \frac{G^2}{2(N+1)^3} + C_1\sqrt{\frac{CHJg(N)}{2}}\right) R_y^2 \\ \leq \min_{\tilde{y}\in B_{2R_y}(0)} \sum_{k=0}^{N-1} \alpha_{k+1}\left(\psi(\tilde{y}^{k+1}) + \left\langle \nabla\psi(\tilde{y}^{k+1}), \tilde{y} - \tilde{y}^{k+1}\right\rangle\right) \\ + \max_{\tilde{y}\in B_{2R_y}(0)} \sum_{k=0}^{N-1} \alpha_{k+1}\left\langle \mathbb{E}_k\left[\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1})\right] - \nabla\psi(\tilde{y}^{k+1}), \tilde{y} - \tilde{y}^{k+1}\right\rangle \\ + \left(8\sqrt{HC_2} + 2 + 12CH + \frac{5GJ}{2} + \frac{G^2}{2(N+1)} + C_1\sqrt{\frac{CHJg(N)}{2}}\right) R_y^2 \quad (140)$$

First of all, we notice that in the same probabilistic event we have $\|\tilde{y}^{k+1} - \tilde{y}^*\|_2 \leq \widetilde{R}_k \overset{(133)}{\leq} JR_0$. Therefore, in the same probabilistic event we get that $\|\tilde{y}^{k+1} - \tilde{y}\|_2 \leq \|\tilde{y}^{k+1} - \tilde{y}^*\|_2 + \|\tilde{y}^* - \tilde{y}\|_2 \leq (J+4)R_y$ for all $\tilde{y} \in B_{2R_y}(0)$, where we used $R_0 = R_y$. It implies that in the same probabilistic event we have

$$\begin{aligned}
&\max_{\tilde{y}\in B_{2R_y}(0)} \sum_{k=0}^{N-1} \alpha_{k+1} \left\langle \mathbb{E}_k \left[\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1})\right] - \nabla\psi(\tilde{y}^{k+1}), \tilde{y} - \tilde{y}^{k+1}\right\rangle \\
&\qquad \leq \max_{\tilde{y}\in B_{2R_y}(0)} \sum_{k=0}^{N-1} \alpha_{k+1} \left\|\mathbb{E}_k \left[\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1})\right] - \nabla\psi(\tilde{y}^{k+1})\right\|_2 \cdot \left\|\tilde{y} - \tilde{y}^{k+1}\right\|_2 \\
&\qquad \overset{(35)}{\leq} \sum_{k=0}^{N-1} \alpha_{k+1}\delta(J+4)R_y \leq \sum_{k=0}^{N-1} \tfrac{k+2}{2\tilde{L}} \tfrac{G\tilde{L}R_0}{(N+1)^2}(J+4)R_y \leq \tfrac{G(J+4)R_y^2}{2}.
\end{aligned}$$

Secondly, using the same trick as in the proof of Theorem 1 from [10] we get that for arbitrary point y

$$\psi(y) - \langle \nabla\psi(y), y\rangle \overset{(18)+(31)}{=} \langle y, Ax(A^\top y)\rangle - f\left(x(A^\top y)\right) - \langle Ax(A^\top y), y\rangle = -f(x(A^\top y)).$$

Using these relations in (140) we obtain that with probability at least $1 - 3\beta$

$$\begin{aligned}
A_N\psi(y^N) \;\leq\; & -\sum_{k=0}^{N-1} \alpha_{k+1} f(x(A^\top\tilde{y}^{k+1})) + \min_{\tilde{y}\in B_{2R_y}(0)} \sum_{k=0}^{N-1} \alpha_{k+1}\langle\nabla\psi(\tilde{y}^{k+1}), \tilde{y}\rangle \\
& + \left(8\sqrt{HC_2} + 2 + 12CH + \frac{G(6J+4)}{2} + \frac{G^2}{2(N+1)} + \right. \\
& \left. + C_1\sqrt{\frac{CHJg(N)}{2}}\right) R_y^2. \qquad (141)
\end{aligned}$$

To bound the first term in (141) we apply convexity of f and introduce the virtual primal iterate $\hat{x}^N = \frac{1}{A_N}\sum\limits_{k=0}^{N-1} \alpha_{k+1} x(A^\top\tilde{y}^{k+1})$:

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

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

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

where we used equality (31). Putting all together we obtain that with probability at least $1 - 3\beta$

$$\begin{aligned}
\psi(y^N) + f(\hat{x}^N) + 2R_y\|A\hat{x}^N\|_2 \;\leq\; & \frac{R_y^2}{A_N}\left(8\sqrt{HC_2} + 2 + 12CH + \frac{G(6J+4)}{2}\right. \\
& \left. + \frac{G^2}{2(N+1)} + C_1\sqrt{\frac{CHJg(N)}{2}}\right). \qquad (142)
\end{aligned}$$

Lemma 9 implies that for all $\gamma > 0$

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

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

$$\begin{aligned}
\|\tilde{x}^N - \hat{x}^N\|_2 &= \frac{1}{A_N}\left\|\sum_{k=0}^{N-1}\alpha_{k+1}\left(\tilde{x}(A^\top\tilde{y}^{k+1},\boldsymbol{\xi}^{k+1}) - x(A^\top\tilde{y}^{k+1})\right)\right\|_2 \\
&\leq \frac{1}{A_N}\left\|\sum_{k=0}^{N-1}\alpha_{k+1}\left(\tilde{x}(A^\top\tilde{y}^{k+1},\boldsymbol{\xi}^{k+1}) - \mathbb{E}\left[\tilde{x}(A^\top\tilde{y}^{k+1},\boldsymbol{\xi}^{k+1}) \mid \tilde{y}^{k+1}\right]\right)\right\|_2 \\
&\quad + \frac{1}{A_N}\left\|\sum_{k=0}^{N-1}\alpha_{k+1}\left(\mathbb{E}\left[\tilde{x}(A^\top\tilde{y}^{k+1},\boldsymbol{\xi}^{k+1}) \mid \tilde{y}^{k+1}\right] - x(A^\top\tilde{y}^{k+1})\right)\right\|_2 \\
&\leq \frac{\sqrt{2}}{A_N}\left(1+\sqrt{3\ln\frac{1}{\beta}}\right)\sqrt{\sum_{k=0}^{N-1}\frac{\alpha_{k+1}^2\sigma_x^2}{r_{k+1}^2}} \\
&\quad + \frac{1}{A_N}\sum_{k=0}^{N-1}\alpha_{k+1}\left\|\mathbb{E}\left[\tilde{x}(A^\top\tilde{y}^{k+1},\boldsymbol{\xi}^{k+1}) \mid \tilde{y}^{k+1}\right] - x(A^\top\tilde{y}^{k+1})\right\|_2 \\
&\overset{(33)}{\leq} \frac{2}{A_N}\sqrt{6\ln\frac{1}{\beta}}\frac{1}{\sqrt{\ln\frac{N}{\beta}}}\sqrt{\sum_{k=0}^{N-1}\frac{C_2\alpha_{k+1}\varepsilon}{\lambda_{\max}(A^\top A)}} + \frac{1}{A_N}\sum_{k=0}^{N-1}\alpha_{k+1}\delta_y \\
&\leq \frac{2}{A_N}\sqrt{\frac{6C_2}{\lambda_{\max}(A^\top A)}}\sqrt{\sum_{k=0}^{N-1}\frac{(k+2)H\tilde{L}R_y^2}{2\tilde{L}N^2}} \\
&+ \frac{1}{A_N}\sum_{k=0}^{N-1}\frac{k+2}{2\tilde{L}}\cdot\frac{G\tilde{L}R_y}{(N+1)^2\sqrt{\lambda_{\max}(A^\top A)}} \\
&\leq \frac{2R_y}{A_N}\left(\sqrt{\frac{6C_2H}{\lambda_{\max}(A^\top A)}} + \frac{G}{4\sqrt{\lambda_{\max}(A^\top A)}}\right). \qquad (143)
\end{aligned}$$

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

$$\begin{aligned}
\|A\tilde{x}^N - A\hat{x}^N\|_2 &\leq \|A\|_2\cdot\|\tilde{x}^N - \hat{x}^N\|_2 \\
&\overset{(143)}{\leq} \sqrt{\lambda_{\max}(A^\top A)}\frac{2R_y}{A_N}\left(\sqrt{\frac{6C_2H}{\lambda_{\max}(A^\top A)}} + \frac{G}{4\sqrt{\lambda_{\max}(A^\top A)}}\right) \\
&= \frac{R_y}{2A_N}\left(\sqrt{96C_2H} + G\right) \qquad (144)
\end{aligned}$$

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

$$\begin{aligned}
2R_y\|A\hat{x}^N\|_2 &\geq 2R_y\|A\tilde{x}^N\|_2 - 2R_yA_N\|A\hat{x}^N - A\tilde{x}^N\|_2 \\
&\overset{(144)}{\geq} 2R_y\|A\tilde{x}^N\|_2 - \frac{R_y^2\left(\sqrt{96C_2H}+G\right)}{A_N}. \qquad (145)
\end{aligned}$$

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

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

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

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

From this and (133) we get that with probability at least $1-2\beta$ the inequality

$$\|x(A^\top\tilde{y}^{k+1})\|_2 \overset{(133)}{\le} \left(\frac{\sqrt{\lambda_{\max}(A^\top A)}J}{\mu} + \frac{R_x}{R_y}\right)R_y \tag{146}$$

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

$$\|\hat{x}^N\|_2 \le \frac{1}{A_N}\sum_{k=0}^{N-1}\alpha_{k+1}\|x(A^\top\tilde{y}^{k+1})\|_2 \overset{(146)}{\le} \left(\frac{\sqrt{\lambda_{\max}(A^\top A)}J}{\mu} + \frac{R_x}{R_y}\right)R_y. \tag{147}$$

We notice that the last inequality lies in the same probability event when (133) holds.

Consider the probability event $E = \{\text{inequalities (142) } - \text{ (147) hold simultaneously}\}$. Using union bound we get that $\mathbb{P}\{E\} \ge 1 - 4\beta$. Combining (143) and (147) we get that inequality

$$\|\tilde{x}^N\|_2 \le \|\tilde{x}^N - \hat{x}^N\|_2 + \|\hat{x}^N\|_2 \le \left(\frac{\left(\sqrt{96C_2H} + G\right)}{2A_N\sqrt{\lambda_{\max}(A^\top A)}} + \frac{\sqrt{\lambda_{\max}(A^\top A)}J}{\mu} + \frac{R_x}{R_y}\right)R_y \tag{148}$$

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 $\tilde{x}^N$ 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 $\tilde{x}^N$ and $\hat{x}^N$ and get that inequalities

$$|f(\hat{x}^N) - f(\tilde{x}^N)| \le L_f\|\hat{x}^N - \tilde{x}^N\|_2 \overset{(143)}{\le} \frac{L_f R_y\left(\sqrt{96C_2H} + G\right)}{2A_N\sqrt{\lambda_{\max}(A^\top A)}} \tag{149}$$

and

$$f(\hat{x}^N) = f(\tilde{x}^N) + \left(f(\hat{x}^N) - f(\tilde{x}^N)\right) \overset{(149)}{\ge} f(\tilde{x}^N) - \frac{L_f R_y\left(\sqrt{96C_2H} + G\right)}{2A_N\sqrt{\lambda_{\max}(A^\top A)}} \tag{150}$$

also lie in the event E. It remains to use inequalities (145) and (150) to bound first and second terms in the right hand side of inequality (142) and obtain that with probability at least $1-4\beta$

$$\begin{aligned}\psi(y^N) + f(\tilde{x}^N) + 2R_y\|A\tilde{x}^N\|_2 \le \frac{R_y^2}{A_N}\Bigg(&8\sqrt{HC_2} + 2 + 12CH + \frac{G(6J+4)}{2}\\ &+\frac{L_f\left(\sqrt{96C_2H} + G\right)}{2R_y\sqrt{\lambda_{\max}(A^\top A)}} + \frac{G^2}{2(N+1)}\\ &+C_1\sqrt{\frac{CHJg(N)}{2}} + \sqrt{96C_2H} + G\Bigg).\end{aligned} \tag{151}$$

Using that A_N grows as $\Omega\left(\frac{N^2}{\tilde{L}}\right)$ [55], $\tilde{L} \leq \frac{2\lambda_{\max}(A^\top A)}{\mu}$ and $R_y \leq \frac{\|\nabla f(x^*)\|_2^2}{\lambda^+_{\min}(A^\top A)}$ (see Section V-D from [16] 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 $1 - 4\beta$

$$f(\tilde{x}^N) - f(x^*) \leq f(\tilde{x}^N) + \psi(y^*) \leq f(\tilde{x}^N) + \psi(y^N) \leq \varepsilon. \tag{152}$$

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 $x = \tilde{x}^N$, we get

$$f(\tilde{x}^N) \geq f(x^*) - R_y\|A\tilde{x}^N\|_2 \tag{153}$$

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

$$\psi(y^N) + f(\tilde{x}^N) \geq \psi(y^*) + f(\tilde{x}^N) \geq -f(x^*) + f(\tilde{x}^N) \geq -R_y\|A\tilde{x}^N\|_2,$$

which implies that inequality

$$\|Ax^N\|_2 \overset{(151)+(152)}{\leq} \frac{\varepsilon}{R_y} \tag{154}$$

holds together with (152) 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 $\sum_{k=1}^N \alpha_{k+1} = A_N$.

F Missing Proofs from Section 5.2

F.1 Proof of Theorem 6

For simplicity we analyse only the first restart since the analysis of the later restarts is the same. We apply Theorem 4 with $N = \bar{N}$ such that

$$\frac{CL_\psi^2 \ln^4 \bar{N}}{\mu_\psi^2 \bar{N}^4} \leq \frac{1}{32}$$

and batch-size

$$r_1 = \max\left\{1, \frac{64C\sigma_\psi^2 \ln^6 \bar{N}}{\bar{N}\|\nabla\Psi(y^0, \boldsymbol{\xi}^0, \hat{r}_1)\|_2^2}\right\}$$

together with simple inequality $\|\nabla\psi(y^0)\|_2 \geq \mu_\psi\|y^0 - y^*\|_2$ and get for all $p = 1, \ldots, p_1$

$$\begin{aligned}\mathbb{E}\left[\|\nabla\psi(\bar{y}^{1,p})\|_2^2 \mid y^0, r_1, \hat{r}_1\right] &\leq \frac{\|\nabla\psi(y^0)\|_2^2}{32} + \frac{\|\nabla\Psi(y^0, \boldsymbol{\xi}^0, \hat{r}_1)\|_2^2}{64} \\ &\overset{(101)}{\leq} \frac{\|\nabla\psi(y^0)\|_2^2}{16} + \frac{\|\nabla\Psi(y^0, \boldsymbol{\xi}^0, \hat{r}_1) - \nabla\psi(y^0)\|_2^2}{32}. \end{aligned} \tag{155}$$

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

$$\|\nabla\psi(\bar{y}^{1,p})\|_2^2 \geq \frac{\|\nabla\psi(y^0)\|_2^2}{8} + \frac{\|\nabla\Psi(y^0, \boldsymbol{\xi}^0, \hat{r}_1) - \nabla\psi(y^0)\|_2^2}{16}.$$

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

$$\|\nabla\psi(\bar{y}^{1,\hat{p}_1})\|_2^2 \leq \frac{\|\nabla\psi(y^0)\|_2^2}{8} + \frac{\|\nabla\Psi(y^0, \boldsymbol{\xi}^0, \hat{r}_1) - \nabla\psi(y^0)\|_2^2}{16}, \tag{156}$$

where $\hat{p}_1$ is such that $\|\nabla\psi(\bar{y}^{1,\hat{p}_1})\|_2^2 = \min_{p=1,\ldots,p_1} \|\nabla\psi(\bar{y}^{1,p})\|_2^2$. From Lemma 9 we have for all $p = 1, \ldots, p_1$

$$\mathbb{P}\left\{ \left\|\nabla\Psi(\bar{y}^{1,p}, \boldsymbol{\xi}^{1,p}, \bar{r}_1) - \nabla\psi(\bar{y}^{1,p})\right\|_2 \geq \left(\sqrt{2} + \sqrt{2\gamma}\right)\sqrt{\frac{\sigma_\psi^2}{\bar{r}_1}} \mid \bar{y}^{1,p} \right\} \leq \exp\left(-\frac{\gamma^2}{3}\right).$$

Since $\bar{r}_1 = \max\left\{1, \frac{128\sigma_\psi^2\left(1+\sqrt{3\ln\frac{lp_1}{\beta}}\right)^2 R_y^2}{\varepsilon^2}\right\}$ we can take $\gamma = \sqrt{3\ln\frac{lp_1}{\beta}}$ in the previous inequality and get that for all $p = 1, \ldots, p_1$ and fixed points $\bar{y}^{1,p}$ with probability at least $1 - \beta/(lp_1)$

$$\left\|\nabla\Psi(\bar{y}^{1,p}, \boldsymbol{\xi}^{1,p}, \bar{r}_1) - \nabla\psi(\bar{y}^{1,p})\right\|_2^2 \leq \frac{\varepsilon^2}{64R_y^2}.$$

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

$$\left\|\nabla\Psi(\bar{y}^{1,p}, \boldsymbol{\xi}^{1,p}, \bar{r}_1) - \nabla\psi(\bar{y}^{1,p})\right\|_2^2 \leq \frac{\varepsilon^2}{64R_y^2}. \tag{157}$$

holds for all $p = 1, \ldots, p_1$ simultaneously with fixed points $\bar{y}^{1,p}$. Using union bound again we get that with probability at least $1 - 2\beta/l$ for fixed $\nabla\Psi(y^0, \boldsymbol{\xi}^0, \hat{r}_1)$

$$\begin{aligned}
\|\nabla\psi(\bar{y}^{1,p(1)})\|_2^2 &\overset{(101)}{\leq} 2\left\|\nabla\Psi(\bar{y}^{1,p(1)}, \boldsymbol{\xi}^{1,p(1)}, \bar{r}_1)\right\|_2^2 + 2\left\|\nabla\Psi(\bar{y}^{1,p(1)}, \boldsymbol{\xi}^{1,p(1)}, \bar{r}_1) - \nabla\psi(\bar{y}^{1,p(1)})\right\|_2^2 \\
&\overset{(157)}{\leq} 2\left\|\nabla\Psi(\bar{y}^{1,\hat{p}_1}, \boldsymbol{\xi}^{1,\hat{p}_1}, \bar{r}_1)\right\|_2^2 + \frac{\varepsilon^2}{32R_y^2} \\
&\overset{(101)}{\leq} 4\|\nabla\psi(\bar{y}^{1,\hat{p}_1})\|_2^2 + 4\left\|\nabla\Psi(\bar{y}^{1,\hat{p}_1}, \boldsymbol{\xi}^{1,\hat{p}_1}, \bar{r}_1) - \nabla\psi(\bar{y}^{1,\hat{p}_1})\right\|_2^2 + \frac{\varepsilon^2}{32R_y^2} \\
&\overset{(156)+(157)}{\leq} \frac{\|\nabla\psi(y^0)\|_2^2}{2} + \frac{\|\nabla\Psi(y^0, \boldsymbol{\xi}^0, \hat{r}_1) - \nabla\psi(y^0)\|_2^2}{4} + \frac{\varepsilon^2}{8R_y^2}.
\end{aligned} \tag{158}$$

Using Lemma 9 with $\gamma = \sqrt{3\ln\frac{l}{\beta}}$ and $\hat{r}_1 = \max\left\{1, \frac{4\sigma_\psi^2\left(1+\sqrt{3\ln\frac{l}{\beta}}\right)^2 R_y^2}{\varepsilon^2}\right\}$ we get that with probability at least $1 - \beta/l$

$$\|\nabla\Psi(y^0, \boldsymbol{\xi}^0, \hat{r}_1) - \nabla\psi(y^0)\|_2^2 \leq \frac{\varepsilon^2}{2R_y^2}. \tag{159}$$

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

$$\|\nabla\psi(\bar{y}^{1,p(1)})\|_2^2 \overset{(158)+(159)}{\leq} \frac{\|\nabla\psi(y^0)\|_2^2}{2} + \frac{\varepsilon^2}{4R_y^2}.$$

Similarly, for all $k = 1, \ldots, l$ with probability at least $1 - 3\beta/l$

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

Using union bound we get that with probability at least $1-3\beta$ the inequality

$$\|\nabla\psi(\bar{y}^{k,p(k)})\|_2^2 \leq \frac{\|\nabla\psi(\bar{y}^{k-1,p(k-1)})\|_2^2}{2} + \frac{\varepsilon^2}{4R_y^2} \tag{160}$$

holds for all $k = 1,\ldots,l$ simultaneously. Finally, unrolling the recurrence an using our choice of $l = \max\left\{1, \log_2\left(2R_y^2\|\nabla\psi(y^0)\|_2^2/\varepsilon^2\right)\right\}$ we obtain that with probability at least $1-3\beta$

$$\begin{aligned}
\|\nabla\psi(\bar{y}^{l,p(l)})\|_2^2 &\overset{(160)}{\leq} \frac{\|\nabla\psi(y^0)\|_2^2}{2^l} + \frac{\varepsilon^2}{4R_y^2}\sum_{k=0}^{l-1} 2^{-k} \\
&\leq \frac{\varepsilon^2}{2R_y^2} + \frac{\varepsilon^2}{4R_y^2}\sum_{k=0}^{\infty} 2^{-k} \\
&= \frac{\varepsilon^2}{2R_y^2} + \frac{\varepsilon^2}{4R_y^2}\cdot 2 = \frac{\varepsilon^2}{R_y^2},
\end{aligned}$$

which concludes the proof. To get (59) we need to estimate $\sum\limits_{k=1}^{l}(\hat{r}_k + \bar{N}p_k r_k + p_k\bar{r}_k)$ using our choice of parameters stated in (57).

F.2 Proof of Corollary 3

Theorem 6, Corollary 2 and inequality $\varepsilon \leq \mu_\psi R_y^2$ imply that with probability at least $1-3\beta$

$$\|\nabla\psi(\bar{y}^{l,p(l)})\|_2 \leq \frac{\varepsilon}{R_y}, \quad \|\bar{y}^{l,p(l)}\|_2 \leq \|\bar{y}^{l,p(l)} - y^*\|_2 + \|y^*\|_2 \overset{(60)}{\leq} 2R_y. \tag{161}$$

Applying Theorem 3 we get that with probability $1-3\beta$ we also have

$$f(\hat{x}^l) - f(x^*) \leq 2\varepsilon, \quad \|A\hat{x}^l\|_2 \leq \frac{\varepsilon}{R_y}, \tag{162}$$

where $\hat{x}^l \overset{\text{def}}{=} x(A^\top\bar{y}^{l,p(l)})$. Next, we show that points $\hat{x}^{l,p} = x(A^\top\bar{y}^{l,p})$ and $x^{l,p} \overset{\text{def}}{=} x(A^\top\bar{y}^{l,p}, \boldsymbol{\xi}^l, \bar{r}_l)$ 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 9 states that

$$\mathbb{P}\left\{\left\|\hat{x}^{l,p} - x^{l,p}\right\|_2 \geq (\sqrt{2}+\sqrt{2\gamma})\sqrt{\frac{\sigma_x^2}{\bar{r}_l}} \mid \bar{y}^{l,p(l)}\right\} \leq \exp\left(-\frac{\gamma^2}{3}\right).$$

Taking $\gamma = \sqrt{3\ln\frac{p_l}{\beta}}$ and using $\bar{r}_l = \max\left\{1, \frac{128\sigma_\psi^2\left(1+\sqrt{3\ln\frac{lp_l}{\beta}}\right)R_y^2}{\varepsilon^2}\right\}$ we get that for all $p = 1,\ldots,p_l$ with probability at least $1 - \beta/p_l$

$$\|\hat{x}^{l,p} - x^{l,p}\|_2 \leq \frac{\varepsilon}{8R_y}\cdot\sqrt{\frac{\sigma_x^2}{\sigma_\psi^2}} = \frac{\varepsilon}{8R_y\sqrt{\lambda_{\max}(A^\top A)}},$$

where we use $\sigma_\psi = \sqrt{\lambda_{\max}(A^\top A)}\sigma_x$. Using union bound we get that with probability at least $1-\beta$ the inequality

$$\|\hat{x}^{l,p} - x^{l,p}\|_2 \leq \frac{\varepsilon}{8R_y\sqrt{\lambda_{\max}(A^\top A)}},$$

holds for all $p = 1, \ldots, p(l)$ simultaneously and, in particular, we get that with probability at least $1 - \beta$

$$\|\hat{x}^l - x^l\|_2 \leq \frac{\varepsilon}{8R_y\sqrt{\lambda_{\max}(A^\top A)}}. \tag{163}$$

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

$$\begin{aligned} \|A\hat{x}^l - Ax^l\|_2 &\leq \|A\|_2 \cdot \|\hat{x}^l - x^l\|_2 \\ &\overset{(163)}{\leq} \sqrt{\lambda_{\max}(A^\top A)}\frac{\varepsilon}{8R_y\sqrt{\lambda_{\max}(A^\top A)}} = \frac{\varepsilon}{8R_y}, \end{aligned} \tag{164}$$

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

$$\|A\hat{x}^l\|_2 \geq \|Ax^l\|_2 - \|A\hat{x}^l - Ax^l\|_2 \overset{(164)}{\geq} \|Ax^l\|_2 - \frac{\varepsilon}{8R_y}. \tag{165}$$

Applying Demyanov-Danskin's theorem, L_φ-smoothness of φ with $L_\varphi = 1/\mu$ and $\varepsilon \leq \mu_\psi R_y^2$ we obtain that with probability at least $1 - \beta$

$$\begin{aligned} \|\hat{x}^l\|_2 &= \|\nabla\varphi(A^\top\bar{y}^{l,p(l)})\|_2 \leq \|\nabla\varphi(A^\top\bar{y}^{l,p(l)}) - \nabla\varphi(A^\top y^*)\|_2 + \|\nabla\varphi(A^\top y^*)\|_2 \\ &\leq L_\varphi\|A^\top\bar{y}^{l,p(l)} - A^\top y^*\|_2 + \|x(A^\top y^*)\|_2 \leq \frac{\sqrt{\lambda_{\max}(A^\top A)}}{\mu}\|\bar{y}^{l,p(l)} - y^*\|_2 + R_x \\ &\overset{(60)}{\leq} \frac{\sqrt{\lambda_{\max}(A^\top A)}\varepsilon}{\mu\mu_\psi R_y} + R_x \leq \left(\frac{\sqrt{\lambda_{\max}(A^\top A)}}{\mu} + \frac{R_x}{R_y}\right)R_y \end{aligned} \tag{166}$$

and also

$$\|x^l\|_2 \leq \|x^l - \hat{x}^l\|_2 + \|\hat{x}^l\|_2 \overset{(163)+(166)}{\leq} \left(\frac{\mu_\psi}{8\sqrt{\lambda_{\max}(A^\top A)}} + \frac{\sqrt{\lambda_{\max}(A^\top A)}}{\mu} + \frac{R_x}{R_y}\right)R_y \tag{167}$$

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 $1 - \beta$

$$\begin{aligned} f(\hat{x}^l) &= f(x^l) + f(\hat{x}^l) - f(x^l) \geq f(x^l) - |f(\hat{x}^l) - f(x^l)| \\ &\geq f(x^l) - L_f\|\hat{x}^l - x^l\|_2 \overset{(163)}{\geq} f(x^l) - \frac{\varepsilon L_f}{8R_y\sqrt{\lambda_{\max}(A^\top A)}}. \end{aligned} \tag{168}$$

Combining inequalities (162), (165) and (168) and using union bound we get that with probability at least $1 - 4\beta$

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

G Missing Proofs from Section 5.3

G.1 Proof of Lemma 5

We prove (65) by induction. For $k = 0$ this inequality is trivial since $A_k = \frac{1}{L}$, $\tilde{y}^1 = y^0$ and $z^0 = \tilde{y}^0$. Next, assume that (65) holds for some $k \geq 0$ and prove it for $k + 1$. By definition of $g_{k+1}(z)$ we have

$$\begin{aligned} \tilde{g}_{k+1}(z^{k+1}) &= \tilde{g}_k(z^{k+1}) \\ &\quad + \alpha_{k+1}\left(\psi(\tilde{y}^{k+1}) + \langle\tilde{\nabla}\Psi(\tilde{y}^{k+1}, \boldsymbol{\xi}^{k+1}), z^{k+1} - \tilde{y}^{k+1}\rangle + \frac{\mu_\psi}{2}\|z^{k+1} - \tilde{y}^{k+1}\|_2^2\right). \end{aligned} \tag{169}$$

Since $\tilde{g}_k(z)$ is $(1 + A_k\mu_\psi)$-strongly convex we can estimate the first term in the r.h.s. of the previous inequality as follows:

$$\begin{aligned}\tilde{g}_k(z^{k+1}) &\geq \tilde{g}_k(z) + \frac{1 + A_k\mu_\psi}{2}\|z^{k+1} - z^k\|_2^2 \\ &\overset{(65)}{\geq} A_k\psi(y^k) + \frac{1 + A_k\mu_\psi}{2}\|z^{k+1} - z^k\|_2^2 \\ &\quad + \sum_{l=0}^{k-1}\frac{A_l\mu_\psi}{2}\|y^l - \tilde{y}^{l+1}\|_2^2 - \sum_{l=0}^{k}\frac{\alpha_l}{2\mu_\psi}\left\|\tilde{\nabla}\Psi(\tilde{y}^l, \boldsymbol{\xi}^l) - \nabla\psi(\tilde{y}^l)\right\|_2^2\end{aligned}$$

Applying μ_ψ-strong convexity of ψ and the relation

$$y^{k+1} = \frac{A_k y^k + \alpha_{k+1}z^{k+1}}{A_{k+1}} = \frac{A_k y^k + \alpha_{k+1}z^k}{A_{k+1}} + \frac{\alpha_{k+1}}{A_{k+1}}\left(z^{k+1} - z^k\right) = \tilde{y}^{k+1} + \frac{\alpha_{k+1}}{A_{k+1}}\left(z^{k+1} - z^k\right)$$

to the previous inequality we get

$$\begin{aligned}\tilde{g}_k(z^{k+1}) &\geq A_k\psi(\tilde{y}^{k+1}) + \langle\nabla\psi(\tilde{y}^{k+1}), A_k(y^k - \tilde{y}^{k+1})\rangle + \frac{A_k\mu_\psi}{2}\|y^k - \tilde{y}^{k+1}\|_2^2 \\ &\quad + \frac{A_{k+1}^2(1 + A_k\mu_\psi)}{2\alpha_{k+1}^2}\|y^{k+1} - \tilde{y}^{k+1}\|_2^2 + \sum_{l=0}^{k-1}\frac{A_l\mu_\psi}{2}\|y^l - \tilde{y}^{l+1}\|_2^2 \\ &\quad - \sum_{l=0}^{k}\frac{\alpha_l}{2\mu_\psi}\left\|\tilde{\nabla}\Psi(\tilde{y}^l, \boldsymbol{\xi}^l) - \nabla\psi(\tilde{y}^l)\right\|_2^2. \qquad (170)\end{aligned}$$

Next, we use (170) in (169) together with relations $A_{k+1} = A_k + \alpha_{k+1}$, $A_{k+1}(1 + A_k\mu_\psi) = \alpha_{k+1}^2 L_\psi$ and $A_k(y^k - \tilde{y}^{k+1}) + \alpha_{k+1}(z^{k+1} - \tilde{y}^{k+1}) = A_{k+1}(y^{k+1} - \tilde{y}^{k+1})$:

$$\begin{aligned}\tilde{g}_{k+1}(z^{k+1}) &\geq A_{k+1}\psi(\tilde{y}^{k+1}) + \langle\nabla\psi(\tilde{y}^{k+1}), A_k(y^k - \tilde{y}^{k+1}) + \alpha_{k+1}(z^{k+1} - \tilde{y}^{k+1})\rangle \\ &\quad + \frac{A_{k+1}^2(1 + A_k\mu_\psi)}{2\alpha_{k+1}^2}\|y^{k+1} - \tilde{y}^{k+1}\|_2^2 + \sum_{l=0}^{k}\frac{A_l\mu_\psi}{2}\|y^l - \tilde{y}^{l+1}\|_2^2 \\ &\quad - \sum_{l=0}^{k}\frac{\alpha_l}{2\mu_\psi}\left\|\tilde{\nabla}\Psi(\tilde{y}^l, \boldsymbol{\xi}^l) - \nabla\psi(\tilde{y}^l)\right\|_2^2 \\ &\quad + \alpha_{k+1}\left\langle\tilde{\nabla}\Psi(\tilde{y}^{l+1}, \boldsymbol{\xi}^{l+1}) - \nabla\psi(\tilde{y}^{l+1}), z^{k+1} - \tilde{y}^{k+1}\right\rangle + \frac{\alpha_{k+1}\mu_\psi}{2}\|z^{k+1} - \tilde{y}^{k+1}\|_2^2 \\ &= A_{k+1}\left(\psi(\tilde{y}^{k+1}) + \langle\nabla\psi(\tilde{y}^{k+1}), y^{k+1} - \tilde{y}^{k+1}\rangle + \frac{L_\psi}{2}\|y^{k+1} - \tilde{y}^{k+1}\|_2^2\right) \\ &\quad + \sum_{l=0}^{k}\frac{A_l\mu_\psi}{2}\|y^l - \tilde{y}^{l+1}\|_2^2 - \sum_{l=0}^{k}\frac{\alpha_l}{2\mu_\psi}\left\|\tilde{\nabla}\Psi(\tilde{y}^l, \boldsymbol{\xi}^l) - \nabla\psi(\tilde{y}^l)\right\|_2^2 \\ &\quad + \alpha_{k+1}\left\langle\tilde{\nabla}\Psi(\tilde{y}^{l+1}, \boldsymbol{\xi}^{l+1}) - \nabla\psi(\tilde{y}^{l+1}), z^{k+1} - \tilde{y}^{k+1}\right\rangle + \frac{\alpha_{k+1}\mu_\psi}{2}\|z^{k+1} - \tilde{y}^{k+1}\|_2^2.\end{aligned}$$

From L_ψ-smoothness of ψ we have

$$\psi(\tilde{y}^{k+1}) + \langle\nabla\psi(\tilde{y}^{k+1}), y^{k+1} - \tilde{y}^{k+1}\rangle + \frac{L_\psi}{2}\|y^{k+1} - \tilde{y}^{k+1}\|_2^2 \geq \psi(y^{k+1}).$$

Next, Fenchel-Young inequality (see inequality (100)) implies that

$$\begin{aligned}\left\langle\tilde{\nabla}\Psi(\tilde{y}^{l+1}, \boldsymbol{\xi}^{l+1}) - \nabla\psi(\tilde{y}^{l+1}), z^{k+1} - \tilde{y}^{k+1}\right\rangle &\geq -\tfrac{1}{2\mu_\psi}\left\|\tilde{\nabla}\Psi(\tilde{y}^{l+1}, \boldsymbol{\xi}^{l+1}) - \nabla\psi(\tilde{y}^{l+1})\right\|_2^2 \\ &\quad - \tfrac{\mu_\psi}{2}\|z^{k+1} - \tilde{y}^{k+1}\|_2^2.\end{aligned}$$

Putting all together and rearranging the terms we get

$$\tilde{g}_{k+1}(z^{k+1}) \geq A_{k+1}\psi(y^{k+1}) + \sum_{l=0}^{k} \frac{A_l\mu_\psi}{2}\|y^l - \tilde{y}^{l+1}\|_2^2 - \sum_{l=0}^{k+1} \frac{\alpha_l}{2\mu_\psi}\left\|\tilde{\nabla}\Psi(\tilde{y}^l, \boldsymbol{\xi}^l) - \nabla\psi(\tilde{y}^l)\right\|_2^2.$$

G.2 Proof of Lemma 6

The idea behind the proof of this lemma is exactly the same as for Lemma 4. We start with applying Cauchy-Schwarz inequality to the second and the third terms, i.e.

$$\begin{aligned}
h\delta(R_k + \widetilde{R}_k) &\leq Dh^2\delta^2 + \frac{R_k^2}{4D} + Dh^2\delta^2 + \frac{\widetilde{R}_k^2}{4D} = 2Dh^2\delta^2 + \frac{R_k^2 + \widetilde{R}_k^2}{4D},\\
u\langle\eta^k, a^k + \tilde{a}^k\rangle &\leq u\|\eta^k\|_2 \cdot \|a^k\|_2 + u\|\eta^k\|_2 \cdot \|\tilde{a}^k\|_2 \leq u\|\eta^k\|_2 R_k + u\|\eta^k\|_2\widetilde{R}_k\\
&\leq u^2D\|\eta^k\|_2^2 + \frac{R_k^2}{4D} + u^2D\|\eta^k\|_2^2 + \frac{\widetilde{R}_k^2}{4D} \leq 2u^2D\|\eta^k\|_2^2 + \frac{R_k^2 + \widetilde{R}_k^2}{4D},
\end{aligned}$$

in the right-hand side of (66):

$$\begin{aligned}
A_lR_l^2 + \sum_{k=0}^{l-1} A_k\widetilde{R}_k^2 &\leq A + 2Dh^2\delta^2\underbrace{\sum_{k=0}^{l-1}\alpha_{k+1}}_{A_l} + \frac{1}{2D}\sum_{k=0}^{l-1}\alpha_{k+1}(R_k^2 + \widetilde{R}_k^2)\\
&\quad + \left(c + 2Du^2\right)\sum_{k=0}^{l-1}\alpha_{k+1}\|\eta^k\|_2^2. \qquad (171)
\end{aligned}$$

Using Lemma 9 we get that with probability at least $1 - \frac{\beta}{N}$

$$\begin{aligned}
\|\eta^k\|_2 &\leq \sqrt{2}\left(1 + \sqrt{3\ln\frac{N}{\beta}}\right)\sigma_k \leq \sqrt{2}\left(1 + \sqrt{3\ln\frac{N}{\beta}}\right)\frac{\sqrt{C\varepsilon}}{N\left(1 + \sqrt{3\ln\frac{N}{\beta}}\right)}\\
&= \sqrt{2C\varepsilon}. \qquad (172)
\end{aligned}$$

Using union bound and $\alpha_{k+1} \leq DA_k$ we get that with probability $\geq 1 - \beta$ inequalities

$$\begin{aligned}
A_lR_l^2 + \sum_{k=0}^{l-1} A_k\widetilde{R}_k^2 &\leq A + 2Dh^2\delta^2A_l + \frac{1}{2}\sum_{k=0}^{l-1} A_k(R_k^2 + \widetilde{R}_k^2) + 2C\left(c + 2Du^2\right)A_l\varepsilon,\\
A_lR_l^2 + \frac{1}{2}\sum_{k=0}^{l-1} A_k\widetilde{R}_k^2 &\leq A + 2Dh^2\delta^2A_l + \frac{1}{2}\sum_{k=0}^{l-1} A_kR_k^2 + 2C\left(c + 2Du^2\right)A_l\varepsilon \qquad (173)
\end{aligned}$$

hold for all $l = 1, \ldots, N$ simultaneously. Therefore, with probability $\geq 1 - \beta$ the inequality

$$\begin{aligned}
A_lR_l^2 &\leq A + 2Dh^2\delta^2A_l + 2C\left(c + 2Du^2\right)A_l\varepsilon + \frac{1}{2}\sum_{k=0}^{l-1} A_kR_k^2\\
&\leq \frac{3}{2}A + 2Dh^2\delta^2\underbrace{\left(A_l + \frac{1}{2}A_{l-1}\right)}_{\leq\frac{3}{2}A_l} + 2C\left(c + 2Du^2\right)\varepsilon\underbrace{\left(A_l + \frac{1}{2}A_{l-1}\right)}_{\leq\frac{3}{2}A_l} + \frac{3}{2}\cdot\frac{1}{2}\sum_{k=0}^{l-2} A_kR_k^2\\
&\leq \frac{3}{2}\left(A + 2Dh^2\delta^2A_l + 2C\left(c + 2Du^2\right)A_l\varepsilon + \frac{1}{2}\sum_{k=0}^{l-2} A_kR_k^2\right),
\end{aligned}$$

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

$$A_l R_l^2 \quad \leq \quad \left(\frac{3}{2}\right)^l \left(A + 2Dh^2\delta^2 A_l + 2C\left(c + 2Du^2\right) A_l\varepsilon\right),$$

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 $\geq 1 - \beta$

$$\sum_{k=0}^{l-1} A_k R_k^2 \leq l\left(\frac{3}{2}\right)^l \left(A + 2Dh^2\delta^2 A_l + 2C\left(c + 2Du^2\right) A_l\varepsilon\right), \tag{174}$$

for all $l = 1, \ldots, N$ simultaneously. Moreover, since (173) holds we have in the same probability event that inequalities

$$\sum_{k=0}^{l-1} A_k \widetilde{R}_k^2 \leq \left(l\left(\frac{3}{2}\right)^l + 2\right) \left(A + 2Dh^2\delta^2 A_l + 2C\left(c + 2Du^2\right) A_l\varepsilon\right) \tag{175}$$

hold with probability $\geq 1 - \beta$ for all $l = 1, \ldots, N$ simultaneously with (174). Next we apply delicate result from [34] which is presented in Section B as Lemma 8. We consider random variables $\xi^k = \alpha_{k+1}\langle \eta^k, a^k + \tilde{a}^k\rangle$. Note that $\mathbb{E}\left[\xi^k \mid \xi^0, \ldots, \xi^{k-1}\right] = \alpha_{k+1}\left\langle \mathbb{E}\left[\eta^k \mid \eta^0, \ldots, \eta^{k-1}\right], a^k\right\rangle = 0$ and

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\frac{(\xi^k)^2}{2\sigma_k^2\alpha_{k+1}^2(R_k^2 + \widetilde{R}_k^2)}\right) \mid \xi^0, \ldots, \xi^{k-1}\right] \\
&\leq \mathbb{E}\left[\exp\left(\frac{\alpha_{k+1}^2\|\eta^k\|_2^2\|a^k + \tilde{a}^k\|_2^2}{2\sigma_k^2\alpha_{k+1}^2(R_k^2 + \widetilde{R}_k^2)}\right) \mid \eta^0, \ldots, \eta^{k-1}\right] \\
&= \mathbb{E}\left[\exp\left(\frac{\|\eta^k\|_2^2}{\sigma_k^2}\right) \mid \eta^0, \ldots, \eta^{k-1}\right] \leq \exp(1)
\end{aligned}$$

due to Cauchy-Schwarz inequality and assumptions of the lemma. If we denote $\hat{\sigma}_k^2 = 2\sigma_k^2\alpha_{k+1}^2(R_k^2 + \widetilde{R}_k^2)$ and apply Lemma 8 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 $b = \hat{\sigma}_0^2$, we get that for all $l = 1, \ldots, N$ with probability $\geq 1 - \frac{\beta}{N}$

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

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

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

and it holds for all $l = 1, \ldots, N$ simultaneously. Note that $\alpha_{k+1} \leq A_{k+1}$, $\varepsilon \leq \frac{HR_0^2}{A_N}$, $\delta \leq \frac{GR_0}{N\sqrt{A_N}}$ and with probability at least $1 - \beta$

$$\begin{aligned}
\sum_{k=0}^{l-1} \hat{\sigma}_k^2 &= 2\sum_{k=0}^{l-1} \sigma_k^2 \alpha_{k+1}^2 (R_k^2 + \widetilde{R}_k^2) \leq \frac{2C\varepsilon}{N^2\left(1 + \sqrt{3\ln\frac{N}{\beta}}\right)^2} \sum_{k=0}^{l-1} A_{k+1} \cdot DA_k (R_k^2 + \widetilde{R}_k^2) \\
&\leq 2\varepsilon CDA_N \sum_{k=0}^{l-1} A_k (R_k^2 + \widetilde{R}_k^2) \\
&\overset{(174)+(175)}{\leq} 4\varepsilon CDA_N \left(l\left(\frac{3}{2}\right)^l + 1\right)\left(A + 2Dh^2\delta^2 A_l + 2C\left(c + 2Du^2\right)A_l\varepsilon\right) \\
&\leq 4HCDR_0^2 \left(N\left(\frac{3}{2}\right)^N + 1\right)\left(A + 2Dh^2G^2R_0^2 + 2C\left(c + 2Du^2\right)HR_0^2\right) \\
&= \frac{B}{2}
\end{aligned}$$

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

$$\left|\sum_{k=0}^{l-1} \xi^k\right| \leq C_1 \sqrt{\sum_{k=0}^{l-1} \hat{\sigma}_k^2 \left(\ln\left(\frac{N}{\beta}\right) + \ln\ln\left(\frac{B}{b}\right)\right)} \tag{176}$$

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

Note that we also proved that (172) is in the same event together with (176) and holds with probability $\geq 1 - 2\beta$. Putting all together in (66), we get that with probability at least $1 - 2\beta$ the inequality

$$\begin{aligned}
&A_l R_l^2 + \sum_{k=0}^{l-1} A_k \widetilde{R}_k^2 \overset{(66)}{\leq} A + h\delta \sum_{k=0}^{l-1} \alpha_{k+1}(R_k + \widetilde{R}_k) + u\sum_{k=0}^{l-1} \alpha_{k+1}\langle \eta^k, a^k + \tilde{a}^k\rangle \\
&+ c\sum_{k=0}^{l-1} \alpha_{k+1}\|\eta^k\|_2^2 \\
&\overset{(172)+(176)}{\leq} A + h\delta \sum_{k=0}^{l-1} \alpha_{k+1}(R_k + \widetilde{R}_k) \\
&+ uC_1 \sqrt{\sum_{k=0}^{l-1} \hat{\sigma}_k^2 \left(\ln\left(\frac{N}{\beta}\right) + \ln\ln\left(\frac{B}{b}\right)\right)} + 2cC\varepsilon A_l
\end{aligned}$$

holds for all $l = 1, \ldots, N$ simultaneously. For brevity, we introduce new notation(neglecting constant factor):

$$g(N) = \frac{\ln\left(\frac{N}{\beta}\right) + \ln\ln\left(\frac{B}{b}\right)}{\left(1 + \sqrt{3\ln\left(\frac{N}{\beta}\right)}\right)^2} \approx 1.$$

Using our assumptions $\sigma_k^2 \leq \frac{C\varepsilon}{N^2\left(1+\sqrt{3\ln\left(\frac{N}{\beta}\right)}\right)^2}$, $\varepsilon \leq \frac{HR_0^2}{A_N}$, $\delta \leq \frac{GR_0}{N\sqrt{A_N}}$ and definition $\hat{\sigma}_k^2 =$

$2\sigma_k^2\alpha_{k+1}^2(R_k^2+\widetilde{R}_k^2)$ we obtain that with probability at least $1-2\beta$ the inequality

$$A_lR_l^2+\sum_{k=0}^{l-1}A_k\widetilde{R}_k^2\le A+h\delta\sum_{k=0}^{l-1}\alpha_{k+1}(R_k+\widetilde{R}_k)+u\sum_{k=0}^{l-1}\alpha_{k+1}\langle\eta^k,a^k+\tilde{a}^k\rangle+c\sum_{k=0}^{l-1}\alpha_{k+1}\|\eta^k\|_2^2$$
$$\le A+\frac{hGR_0}{N\sqrt{A_N}}\sum_{k=0}^{l-1}\alpha_{k+1}(R_k+\widetilde{R}_k)+\frac{uC_1R_0\sqrt{2HCg(N)}}{N\sqrt{A_N}}\sqrt{\sum_{k=0}^{l-1}\alpha_{k+1}^2(R_k^2+\widetilde{R}_k^2)}$$
$$+2cHCR_0^2$$
$$\le\left(\frac{A}{R_0^2}+2cHC\right)R_0^2+\frac{\left(hG+uC_1\sqrt{2HCg(N)}\right)R_0}{N\sqrt{A_N}}\sum_{k=0}^{l-1}\alpha_{k+1}(R_k+\widetilde{R}_k)\tag{177}$$

holds for all $l=1,\dots,N$ simultaneously, where in the last row we applied well-known inequality: $\sqrt{\sum_{i=1}^m a_i^2}\le\sum_{i=1}^m a_i$ for $a_i\ge0$, $i=1,\dots,m$. Next we use Lemma 16 with $A=\frac{A}{R_0^2}+2cHC$, $B=hG+uC_1\sqrt{2HCg(N)}$, $r_k=R_k$, $\tilde{r}_k=\widetilde{R}_k$ and get that with probability at least $1-2\beta$ inequalities

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

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

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

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

G.3 Proof of Theorem 7

From Lemma 5 we have

$$A_k\psi(y^k)\le\tilde{g}_k(z^k)-\sum_{l=0}^{k-1}\frac{A_l\mu_\psi}{2}\|y^l-\tilde{y}^{l+1}\|_2^2+\sum_{l=0}^{k}\frac{\alpha_l}{2\mu_\psi}\left\|\tilde{\nabla}\Psi(\tilde{y}^l,\boldsymbol{\xi}^l)-\nabla\psi(\tilde{y}^l)\right\|_2^2\tag{178}$$

for all $k \geq 0$. By definition of z^k we get that

$$\begin{aligned}
\tilde{g}_k(z^k) &= \min_{z\in\mathbb{R}^n}\left\{\frac{1}{2}\|z-z^0\|_2^2 + \sum_{l=0}^{k}\alpha_l\left(\psi(\tilde{y}^l) + \langle\tilde{\nabla}\Psi(\tilde{y}^l,\boldsymbol{\xi}^l), z-\tilde{y}^l\rangle + \frac{\mu_\psi}{2}\|z-\tilde{y}^l\|_2^2\right)\right\} \\
&\leq \frac{1}{2}\|y^*-z^0\|_2^2 + \sum_{l=0}^{k}\alpha_l\left(\psi(\tilde{y}^l) + \langle\tilde{\nabla}\Psi(\tilde{y}^l,\boldsymbol{\xi}^l), y^*-\tilde{y}^l\rangle + \frac{\mu_\psi}{2}\|y^*-\tilde{y}^l\|_2^2\right) \\
&= \frac{1}{2}\|y^*-z^0\|_2^2 + \sum_{l=0}^{k}\alpha_l\left(\psi(\tilde{y}^l) + \langle\nabla\psi(\tilde{y}^l), y^*-\tilde{y}^l\rangle + \frac{\mu_\psi}{2}\|y^*-\tilde{y}^l\|_2^2\right) \\
&\quad + \sum_{l=0}^{k}\alpha_l\langle\tilde{\nabla}\Psi(\tilde{y}^l,\boldsymbol{\xi}^l) - \nabla\psi(\tilde{y}^l), y^*-\tilde{y}^l\rangle \\
&\leq \frac{1}{2}\|y^*-y^0\|_2^2 + A_k\psi(y^*) + \sum_{l=0}^{k}\alpha_l\langle\tilde{\nabla}\Psi(\tilde{y}^l,\boldsymbol{\xi}^l) - \nabla\psi(\tilde{y}^l), y^*-\tilde{y}^l\rangle, \qquad (179)
\end{aligned}$$

where the last inequality follows from μ_ψ-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 $\widetilde{R}_k \overset{\text{def}}{=} \|y^k - \tilde{y}^{k+1}\|_2$ for all $k \geq 0$. Using this and another consequence of strong convexity, i.e. $\psi(y) - \psi(y^*) \geq \frac{\mu_\psi}{2}\|y-y^*\|_2^2$, we obtain

$$\begin{aligned}
\frac{A_k\mu_\psi}{2}R_k^2 + \sum_{l=0}^{k-1}\frac{A_l\mu_\psi}{2}\widetilde{R}_l^2 &\leq A_k\left(\psi(y^k)-\psi(y^*)\right) + \sum_{l=0}^{k-1}\frac{A_l\mu_\psi}{2}\widetilde{R}_l^2 \\
&\overset{(178)+(179)}{\leq} \frac{1}{2}R_0^2 + \sum_{l=0}^{k}\alpha_l\langle\tilde{\nabla}\Psi(\tilde{y}^l,\boldsymbol{\xi}^l) - \nabla\psi(\tilde{y}^l), y^*-\tilde{y}^l\rangle \\
&\quad + \sum_{l=0}^{k}\frac{\alpha_l}{2\mu_\psi}\left\|\tilde{\nabla}\Psi(\tilde{y}^l,\boldsymbol{\xi}^l) - \nabla\psi(\tilde{y}^l)\right\|_2^2. \qquad (180)
\end{aligned}$$

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

$$\begin{aligned}
\langle\tilde{\nabla}\Psi(\tilde{y}^l,\boldsymbol{\xi}^l) - \nabla\psi(\tilde{y}^l), y^*-\tilde{y}^l\rangle &= \left\langle\mathbb{E}\left[\tilde{\nabla}\Psi(\tilde{y}^l,\boldsymbol{\xi}^l)\right] - \nabla\psi(\tilde{y}^l), y^*-\tilde{y}^l\right\rangle \\
&\quad + \left\langle\tilde{\nabla}\Psi(\tilde{y}^l,\boldsymbol{\xi}^l) - \mathbb{E}\left[\tilde{\nabla}\Psi(\tilde{y}^l,\boldsymbol{\xi}^l)\right], y^*-\tilde{y}^l\right\rangle \\
&\overset{(38)}{\leq} \delta\|y^*-\tilde{y}^l\|_2 + \left\langle\tilde{\nabla}\Psi(\tilde{y}^l,\boldsymbol{\xi}^l) - \mathbb{E}\left[\tilde{\nabla}\Psi(\tilde{y}^l,\boldsymbol{\xi}^l)\right], y^*-\tilde{y}^l\right\rangle, \\
\left\|\tilde{\nabla}\Psi(\tilde{y}^l,\boldsymbol{\xi}^l) - \nabla\psi(\tilde{y}^l)\right\|_2^2 &\leq 2\left\|\mathbb{E}\left[\tilde{\nabla}\Psi(\tilde{y}^l,\boldsymbol{\xi}^l)\right] - \nabla\psi(\tilde{y}^l)\right\|_2^2 \\
&\quad + 2\left\|\tilde{\nabla}\Psi(\tilde{y}^l,\boldsymbol{\xi}^l) - \mathbb{E}\left[\tilde{\nabla}\Psi(\tilde{y}^l,\boldsymbol{\xi}^l)\right]\right\|_2^2 \\
&\overset{(38)}{\leq} 2\delta^2 + 2\left\|\tilde{\nabla}\Psi(\tilde{y}^l,\boldsymbol{\xi}^l) - \mathbb{E}\left[\tilde{\nabla}\Psi(\tilde{y}^l,\boldsymbol{\xi}^l)\right]\right\|_2^2
\end{aligned}$$

for all $l \geq 0$. Next, we introduce new notation

$$\begin{aligned}
\tilde{A} &\overset{\text{def}}{=} \frac{1}{2}R_0^2 + \delta\alpha_0 R_0 + \frac{A_N\delta^2}{\mu_\psi} + \alpha_0\left\langle\tilde{\nabla}\Psi(\tilde{y}^0,\boldsymbol{\xi}^0) - \mathbb{E}\left[\tilde{\nabla}\Psi(\tilde{y}^0,\boldsymbol{\xi}^0)\right], y^*-\tilde{y}^0\right\rangle \\
&\quad + \frac{\alpha_0}{\mu_\psi}\left\|\tilde{\nabla}\Psi(\tilde{y}^0,\boldsymbol{\xi}^0) - \mathbb{E}\left[\tilde{\nabla}\Psi(\tilde{y}^0,\boldsymbol{\xi}^0)\right]\right\|_2^2. \qquad (181)
\end{aligned}$$

Putting all together in (180) we get

$$\begin{aligned}
&\frac{A_k\mu_\psi}{2}R_k^2+\sum_{l=0}^{k-1}\frac{A_l\mu_\psi}{2}\widetilde{R}_l^2\\
&\le\frac{1}{2}R_0^2+\delta\sum_{l=0}^{k}\alpha_l\|y^*-\tilde{y}^l\|_2+\sum_{l=0}^{k}\alpha_l\left\langle\tilde{\nabla}\Psi(\tilde{y}^l,\boldsymbol{\xi}^l)-\mathbb{E}\left[\tilde{\nabla}\Psi(\tilde{y}^l,\boldsymbol{\xi}^l)\right],y^*-\tilde{y}^l\right\rangle\\
&\quad+\frac{\delta^2}{\mu_\psi}\sum_{l=0}^{k}\alpha_l+\frac{1}{\mu_\psi}\sum_{l=0}^{k}\alpha_l\left\|\tilde{\nabla}\Psi(\tilde{y}^l,\boldsymbol{\xi}^l)-\mathbb{E}\left[\tilde{\nabla}\Psi(\tilde{y}^l,\boldsymbol{\xi}^l)\right]\right\|_2^2\\
&\le\tilde{A}+\delta\sum_{l=0}^{k-1}\alpha_{l+1}\|y^*-\tilde{y}^{l+1}\|_2\\
&\quad+\sum_{l=0}^{k-1}\alpha_{l+1}\left\langle\tilde{\nabla}\Psi(\tilde{y}^{l+1},\boldsymbol{\xi}^{l+1})-\mathbb{E}\left[\tilde{\nabla}\Psi(\tilde{y}^{l+1},\boldsymbol{\xi}^{l+1})\right],y^*-\tilde{y}^{l+1}\right\rangle\\
&\quad+\frac{1}{\mu_\psi}\sum_{l=0}^{k-1}\alpha_{l+1}\left\|\tilde{\nabla}\Psi(\tilde{y}^{l+1},\boldsymbol{\xi}^{l+1})-\mathbb{E}\left[\tilde{\nabla}\Psi(\tilde{y}^{l+1},\boldsymbol{\xi}^{l+1})\right]\right\|_2^2. \qquad (182)
\end{aligned}$$

To simplify previous inequality we define new vectors $a^l\overset{\text{def}}{=}y^*-y^l$, $\tilde{a}^l\overset{\text{def}}{=}y^l-\tilde{y}^{l+1}$, $\eta^l\overset{\text{def}}{=}\tilde{\nabla}\Psi(\tilde{y}^{l+1},\boldsymbol{\xi}^{l+1})-\mathbb{E}\left[\tilde{\nabla}\Psi(\tilde{y}^{l+1},\boldsymbol{\xi}^{l+1})\right]$ for all $l\ge0$. Note that $\|a^l\|_2=R_l$, $\|\tilde{a}^l\|_2=\widetilde{R}_l$ and $\tilde{a}^0=y^0-\tilde{y}^1=0$. Using this we can rewrite (182) in the following form:

$$\begin{aligned}
A_kR_k^2+\sum_{l=0}^{k-1}A_l\widetilde{R}_l^2 \;\le\; A+\frac{2\delta}{\mu_\psi}\sum_{l=0}^{k-1}\alpha_{l+1}(R_l+\widetilde{R}_l)+\frac{2}{\mu_\psi}\sum_{l=0}^{k-1}\alpha_{l+1}\langle\eta^l,a^l+\tilde{a}^l\rangle\\
+\frac{2}{\mu_\psi^2}\sum_{l=0}^{k-1}\alpha_{l+1}\|\eta^l\|_2^2, \qquad (183)
\end{aligned}$$

where we used $A\overset{\text{def}}{=}\frac{2\tilde{A}}{\mu_\psi}$ and triangle inequality, i.e. $\|y^*-\tilde{y}^{l+1}\|_2\le\|y^*-y^l\|_2+\|y^l-\tilde{y}^{l+1}\|_2=R_l+\widetilde{R}_l$. Next, we apply Lemma 6 with $h=u=\frac{2}{\mu_\psi}$, $c=\frac{2}{\mu_\psi^2}$ and get that with probability at least $1-2\beta$

$$R_N^2\le\frac{J^2R_0^2}{A_N} \qquad (184)$$

where

$$g(N)=\frac{\ln\left(\frac{N}{\beta}\right)+\ln\ln\left(\frac{B}{b}\right)}{\left(1+\sqrt{3\ln\left(\frac{N}{\beta}\right)}\right)^2},\; b=\frac{2\sigma_1^2\alpha_1^2R_0^2}{r_1},\; D\overset{(200)}{=}1+\frac{\mu_\psi}{L_\psi}+\sqrt{1+\frac{\mu_\psi}{L_\psi}},$$

$$B=8HC\left(\frac{L_\psi}{\mu_\psi}\right)^{3/2}DR_0^2\left(N\left(\frac{3}{2}\right)^N+1\right)\left(A+2Dh^2G^2R_0^2+2C\left(\frac{L_\psi}{\mu_\psi}\right)^{3/2}\left(c+2Du^2\right)HR_0^2\right),$$

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

$$B_1 = hG + uC_1\sqrt{2HC\left(\frac{L_\psi}{\mu_\psi}\right)^{3/2} g(N)}$$

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 $A = \frac{2\tilde{A}}{\mu_\psi}$ and

$$A \overset{(181)}{=} \frac{R_0^2}{\mu_\psi} + \frac{2\delta\alpha_0 R_0}{\mu_\psi} + \frac{2A_N\delta^2}{\mu_\psi^2} + \frac{2\alpha_0}{\mu_\psi}\left\langle \tilde{\nabla}\Psi(\tilde{y}^0,\boldsymbol{\xi}^0) - \mathbb{E}\left[\tilde{\nabla}\Psi(\tilde{y}^0,\boldsymbol{\xi}^0)\right], y^* - \tilde{y}^0\right\rangle + \frac{2\alpha_0}{\mu_\psi^2}\left\|\tilde{\nabla}\Psi(\tilde{y}^0,\boldsymbol{\xi}^0) - \mathbb{E}\left[\tilde{\nabla}\Psi(\tilde{y}^0,\boldsymbol{\xi}^0)\right]\right\|_2^2. \tag{185}$$

Lemma 9 implies that

$$\mathbb{P}\left\{\left\|\tilde{\nabla}\Psi(\tilde{y}^0,\boldsymbol{\xi}^0) - \mathbb{E}\left[\tilde{\nabla}\Psi(\tilde{y}^0,\boldsymbol{\xi}^0)\right]\right\|_2 \geq \sqrt{2}(1+\sqrt{\gamma})\sqrt{\frac{\sigma_\psi^2}{r_0}}\right\} \leq \exp\left(-\frac{\gamma^2}{3}\right).$$

Taking $\gamma = \sqrt{3\ln\frac{1}{\beta}}$ and using $r_0 \geq \left(\frac{\mu_\psi}{L_\psi}\right)^{3/2}\frac{N^2\sigma_\psi^2\left(1+\sqrt{3\ln\frac{N}{\beta}}\right)^2}{C\varepsilon}$, $\varepsilon \leq \frac{HR_0^2}{A_N}$ we get that with probability at least $1-\beta$

$$\left\langle \tilde{\nabla}\Psi(\tilde{y}^0,\boldsymbol{\xi}^0) - \mathbb{E}\left[\tilde{\nabla}\Psi(\tilde{y}^0,\boldsymbol{\xi}^0)\right], y^* - \tilde{y}^0\right\rangle \leq \left\|\tilde{\nabla}\Psi(\tilde{y}^0,\boldsymbol{\xi}^0) - \mathbb{E}\left[\tilde{\nabla}\Psi(\tilde{y}^0,\boldsymbol{\xi}^0)\right]\right\|_2 \cdot \|y^* - y^0\|_2$$
$$\leq \left(\frac{L_\psi}{\mu_\psi}\right)^{3/4}\frac{\sqrt{2C\varepsilon}R_0}{N} \leq \left(\frac{L_\psi}{\mu_\psi}\right)^{3/4}\frac{\sqrt{2CH}R_0^2}{N\sqrt{A_N}}, \tag{186}$$

$$\left\|\tilde{\nabla}\Psi(\tilde{y}^0,\boldsymbol{\xi}^0) - \mathbb{E}\left[\tilde{\nabla}\Psi(\tilde{y}^0,\boldsymbol{\xi}^0)\right]\right\|_2^2 \leq \left(\frac{L_\psi}{\mu_\psi}\right)^{3/2}\frac{2C\varepsilon}{N^2} \leq \left(\frac{L_\psi}{\mu_\psi}\right)^{3/2}\frac{2CHR_0^2}{N^2A_N}. \tag{187}$$

From this and $\delta \leq \frac{GR_0}{N\sqrt{A_N}}$ we obtain that with probability $\geq 1-\beta$

$$A \overset{(185)+(186)+(187)}{\leq} \hat{A}R_0^2,$$
$$\hat{A} \overset{\text{def}}{=} \frac{1}{\mu_\psi} + \frac{2G}{L_\psi\mu_\psi N\sqrt{A_N}} + \frac{2G^2}{\mu_\psi^2N^2} + \left(\frac{L_\psi}{\mu_\psi}\right)^{3/4}\frac{2\sqrt{2CH}}{L_\psi\mu_\psi N\sqrt{A_N}} + \left(\frac{L_\psi}{\mu_\psi}\right)^{3/2}\frac{4CH}{L_\psi\mu_\psi^2N^2A_N}.$$

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

$$\hat{g}(N) = \frac{\ln\left(\frac{N}{\beta}\right) + \ln\ln\left(\frac{\hat{B}}{b}\right)}{\left(1+\sqrt{3\ln\left(\frac{N}{\beta}\right)}\right)^2},$$

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

$$\hat{J} = \max\left\{\sqrt{A_0}, \frac{3\hat{B}_1 D + \sqrt{9\hat{B}_1^2 D^2 + 4\hat{A} + 8cHC\left(\frac{L_\psi}{\mu_\psi}\right)^{3/2}}}{2}\right\},$$

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

Note that

$$A_k \overset{(199)}{\geq} \frac{1}{L_\psi}\left(1 + \frac{1}{2}\sqrt{\frac{\mu_\psi}{L_\psi}}\right)^{2k}.$$

It means that in order to achieve $R_N^2 \leq \varepsilon$ with probability at least $1 - 3\beta$ the method requires $N = \tilde{O}\left(\sqrt{\frac{L_\psi}{\mu_\psi}}\ln\frac{1}{\varepsilon}\right)$ iterations and

$$\sum_{k=0}^{N} r_k = \widetilde{O}\left(\max\left\{\sqrt{\frac{L_\psi}{\mu_\psi}}, \frac{\sigma_\psi^2}{\varepsilon}\ln\frac{1}{\beta}\right\}\ln\frac{1}{\varepsilon}\right)$$

oracle calls where $\widetilde{O}(\cdot)$ hides polylogarithmic factors depending on $L_\psi, \mu_\psi, R_0, \varepsilon$ and β.

G.4 Proof of Corollary 5

Corollary 4 implies that with probability at least $1 - 3\beta$

$$\|y^N\|_2 \leq 2R_y, \quad \|\nabla\psi(y^N)\|_2 \leq \frac{\varepsilon}{R_y}$$

and the total number of oracle calls to get this is of order (79). Together with Theorem 3 it gives us that with probability at least $1 - 3\beta$

$$f(\hat{x}^N) - f(x^*) \leq 2\hat{\varepsilon}, \quad \|A\hat{x}^N\|_2 \leq \frac{\hat{\varepsilon}}{R_y}, \tag{188}$$

where $\hat{x}^N \overset{\text{def}}{=} x(A^\top y^N)$. It remains to show that $\tilde{x}^N$ and $\hat{x}^N$ are *close* to each other with high probability. Lemma 9 states that

$$\mathbb{P}\left\{\left\|\tilde{x}^N - \mathbb{E}\left[\tilde{x}^N \mid y^N\right]\right\|_2 \geq (\sqrt{2} + \sqrt{2\gamma})\sqrt{\frac{\sigma_x^2}{r_N}} \mid y^N\right\} \leq \exp\left(-\frac{\gamma^2}{3}\right).$$

Taking $\gamma = \sqrt{3\ln\frac{1}{\beta}}$ and using $r_N \geq \frac{1}{C}\frac{\sigma_\psi^2 R_y^2\left(1+\sqrt{3\ln\frac{1}{\beta}}\right)^2}{\varepsilon^2}$ we get that with probability at least $1 - \beta$

$$\begin{aligned}
\left\|\tilde{x}^N - \mathbb{E}\left[\tilde{x}^N \mid y^N\right]\right\|_2 &\leq \sqrt{2C\frac{\sigma_x^2\varepsilon^2}{\sigma_\psi^2 R_y^2}} = \frac{\sqrt{2C}\varepsilon}{R_y\sqrt{\lambda_{\max}(A^\top A)}}, \\
\left\|\tilde{x}^N - \hat{x}^N\right\|_2 &\leq \left\|\tilde{x}^N - \mathbb{E}\left[\tilde{x}^N \mid y^N\right]\right\|_2 + \left\|\mathbb{E}\left[\tilde{x}^N \mid y^N\right] - \hat{x}^N\right\|_2 \\
&\overset{(33)}{\leq} \frac{\sqrt{2C}\varepsilon}{R_y\sqrt{\lambda_{\max}(A^\top A)}} + \frac{G_1\varepsilon}{NR_y} \\
&\leq \left(\sqrt{\frac{2C}{\lambda_{\max}(A^\top A)}} + G_1\right)\frac{\varepsilon}{R_y}.
\end{aligned} \tag{189}$$

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

$$\begin{aligned}\|A\tilde{x}^N - A\hat{x}^N\|_2 &\leq \|A\|_2 \cdot \|\tilde{x}^N - \hat{x}^N\|_2 \\ &\overset{(189)}{\leq} \sqrt{\lambda_{\max}(A^\top A)}\left(\sqrt{\frac{2C}{\lambda_{\max}(A^\top A)}} + G_1\right)\frac{\varepsilon}{R_y} \\ &= \left(\sqrt{2C} + G_1\sqrt{\lambda_{\max}(A^\top A)}\right)\frac{\varepsilon}{R_y}, \end{aligned} \tag{190}$$

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

$$\begin{aligned}\|A\hat{x}^N\|_2 &\geq \|A\tilde{x}^N\|_2 - \|A\hat{x}^N - A\tilde{x}^N\|_2 \\ &\overset{(190)}{\geq} \|A\tilde{x}^N\|_2 - \left(\sqrt{2C} + G_1\sqrt{\lambda_{\max}(A^\top A)}\right)\frac{\varepsilon}{R_y}. \end{aligned} \tag{191}$$

Applying Demyanov-Danskin theorem and L_φ-smoothness of φ with $L_\varphi = 1/\mu$ we obtain that with probability at least $1-\beta$

$$\begin{aligned}\|\hat{x}^N\|_2 &= \|\nabla\varphi(A^\top y^N)\|_2 \leq \|\nabla\varphi(A^\top y^N) - \nabla\varphi(A^\top y^*)\|_2 + \|\nabla\varphi(A^\top y^*)\|_2 \\ &\leq L_\varphi\|A^\top y^N - A^\top y^*\|_2 + \|x(A^\top y^*)\|_2 \leq \frac{\sqrt{\lambda_{\max}(A^\top A)}}{\mu}\|y^N - y^*\|_2 + R_x \\ &\overset{(74)}{\leq} \frac{\sqrt{\lambda_{\max}(A^\top A)}\varepsilon}{\mu R_y} + R_x \end{aligned} \tag{192}$$

and also

$$\begin{aligned}\|\tilde{x}^N\|_2 &\leq \|\tilde{x}^N - \hat{x}^N\|_2 + \|\hat{x}^N\|_2 \\ &\overset{(189)+(192)}{\leq} \left(\sqrt{\frac{2C}{\lambda_{\max}(A^\top A)}} + G_1 + \frac{\sqrt{\lambda_{\max}(A^\top A)}}{\mu}\right)\frac{\varepsilon}{R_y} + R_x. \end{aligned} \tag{193}$$

That is, we proved that with probability at least $1-\beta$ points $\hat{x}^l$ and $\tilde{x}^l$ lie in the ball $B_{R_f}(0)$. In this ball function f is L_f-Lipschitz continuous, therefore, with probability at least $1-\beta$

$$\begin{aligned}f(\hat{x}^N) &= f(\tilde{x}^N) + f(\hat{x}^N) - f(\tilde{x}^N) \geq f(\tilde{x}^N) - |f(\hat{x}^N) - f(\tilde{x}^N)| \\ &\geq f(\tilde{x}^N) - L_f\|\hat{x}^N - \tilde{x}^N\|_2 \\ &\overset{(189)}{\geq} f(\tilde{x}^N) - \left(\sqrt{\frac{2C}{\lambda_{\max}(A^\top A)}} + G_1\right)\frac{L_f\varepsilon}{R_y}. \end{aligned} \tag{194}$$

Combining inequalities (188), (191) and (194) and using union bound we get that with probability at least $1-4\beta$

$$\begin{aligned}f(\tilde{x}^N) - f(x^*) &\leq \left(2 + \left(\sqrt{\frac{2C}{\lambda_{\max}(A^\top A)}} + G_1\right)\frac{L_f}{R_y}\right)\varepsilon, \\ \|A\tilde{x}^N\|_2 &\leq \left(1 + \sqrt{2C} + G_1\sqrt{\lambda_{\max}(A^\top A)}\right)\frac{\varepsilon}{R_y}. \end{aligned}$$

H Technical Results

Lemma 12. For the sequence $\alpha_{k+1} \geq 0$ such that

$$A_{k+1} = A_k + \alpha_{k+1}, \quad A_{k+1} = 2L\alpha_{k+1}^2 \tag{195}$$

we have for all $k \geq 0$

$$\alpha_{k+1} \leq \widetilde{\alpha}_{k+1} \overset{\text{def}}{=} \frac{k+2}{2L}. \tag{196}$$

Moreover, $A_k = \Omega\left(\frac{N^2}{L}\right)$.

Proof. We prove (196) by induction. For $k = 0$ equation (195) gives us $\alpha_1 = 2L\alpha_1^2 \iff \alpha_1 = \frac{1}{2L}$. Next we assume that (196) holds for all $k \leq l-1$ and prove it for $k = l$:

$$2L\alpha_{l+1}^2 \overset{(195)}{=} \sum_{i=1}^{l+1} \alpha_i \overset{(196)}{\leq} \alpha_{l+1} + \frac{1}{2L}\sum_{i=1}^{l}(i+1) = \alpha_{l+1} + \frac{l(l+3)}{4L}.$$

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 $A_k = \Omega\left(\frac{N^2}{L}\right)$ is proved in Lemma 1 from [25] (see also [55]). □

Lemma 13 (See Lemma 3 from [24] and Lemma 4 from [14])**.** For the sequence $\alpha_{k+1} \geq 0$ such that

$$A_{k+1} = A_k + \alpha_{k+1}, \quad A_{k+1}(1 + A_k\mu) = L\alpha_{k+1}^2, \quad \alpha_0 = A_0 = \frac{1}{L} \tag{197}$$

we have for all $k \geq 0$

$$\alpha_{k+1} = \frac{1 + A_k\mu}{2L} + \sqrt{\frac{(1+A_k\mu)^2}{4L^2} + \frac{A_k(1+A_k\mu)}{L}}, \tag{198}$$

$$A_k \geq \frac{1}{L}\left(1 + \frac{1}{2}\sqrt{\frac{\mu}{L}}\right)^{2k}, \tag{199}$$

$$\alpha_{k+1} \leq \left(1 + \frac{\mu}{L} + \sqrt{1 + \frac{\mu}{L}}\right) A_k. \tag{200}$$

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 (198). Inequality (199) was established in Lemma 3 from [24] and Lemma 4 from [14]. It remains to prove (200). 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

$$\begin{aligned}
\alpha_{k+1} &\overset{(198)}{=} \frac{1 + A_k\mu}{2L} + \sqrt{\frac{(1+A_k\mu)^2}{4L^2} + \frac{A_k(1+A_k\mu)}{L}} \\
&\leq \frac{1}{2L} + \frac{\mu}{2L}A_k + \frac{1 + A_k\mu}{2L} + \sqrt{\frac{A_k}{L} + \frac{\mu}{L}A_k^2} \\
&\leq \frac{1}{L} + \frac{\mu}{L}A_k + A_k\sqrt{1 + \frac{\mu}{L}} = \left(1 + \frac{\mu}{L} + \sqrt{1 + \frac{\mu}{L}}\right) A_k.
\end{aligned}$$

□

Lemma 14. Let $A, B, D, r_0, r_1, \ldots, r_N$, where $N \geq 1$, be non-negative numbers such that

$$\frac{1}{2}r_l^2 \leq Ar_0^2 + \frac{Dr_0}{(N+1)^2}\sum_{k=0}^{l-1}(k+2)r_k + B\frac{r_0}{N}\sqrt{\sum_{k=0}^{l-1}(k+2)r_k^2}, \quad \forall l = 1, \ldots, N. \tag{201}$$

Then for all $l = 0, \ldots, N$ we have

$$r_l \leq Cr_0, \tag{202}$$

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 (202) by induction. For $l = 0$ the inequality $r_l \leq Cr_0$ trivially follows since $C \geq 1$. Next we assume that (202) holds for some $l < N$ and prove it for $l+1$:

$$\begin{aligned}
r_{l+1} &\overset{(201)}{\leq} \sqrt{2}\sqrt{Ar_0^2 + \frac{Dr_0}{(N+1)^2}\sum_{k=0}^{l}(k+2)r_k + B\frac{r_0}{N}\sqrt{\sum_{k=0}^{l}(k+2)r_k^2}} \\
&\overset{(202)}{\leq} r_0\sqrt{2}\sqrt{A + \frac{DC}{(N+1)^2}\sum_{k=0}^{l}(k+2) + \frac{BC}{N}\sqrt{\sum_{k=0}^{l}(k+2)}} \\
&\leq r_0\sqrt{2}\sqrt{A + \frac{DC}{(N+1)^2}\frac{(l+1)(l+2)}{2} + \frac{BC}{N}\sqrt{\frac{(l+1)(l+2)}{2}}} \\
&\leq r_0\sqrt{2}\sqrt{A + DC + \frac{BC}{N}\sqrt{\frac{N(N+1)}{2}}} \leq r_0\underbrace{\sqrt{2A + 2(B+D)C}}_{\leq C} \leq Cr_0.
\end{aligned}$$

□

Lemma 15. Let $C, r_0, r_1, \ldots, r_N$, where $N \geq 1$, be non-negative numbers such that

$$r_l^2 \leq r_0^2 + \frac{2C}{(N+1)^{3/2}}\sum_{k=0}^{l-1}(k+2)^{1/2}r_{k+1}^2, \quad \forall l = 1, \ldots, N, \tag{203}$$

and $C \in (0, 1/4)$. Then for all $l = 0, \ldots, N$ we have

$$r_l \leq 2r_0, \tag{204}$$

Proof. We prove (204) by induction. For $l = 0$ the inequality $r_l \leq 2r_0$ trivially follows. Next we assume that (204) holds for some $l \leq N-1$ and prove it for $l+1$. From (203), $C < 1/4$, $N \geq 1$ and $l \leq N-1$ we have

$$\begin{aligned}
\frac{3}{4}r_{l+1}^2 &\leq \left(1 - \frac{2C(l+2)^{1/2}}{(N+1)^{3/2}}\right)r_{l+1}^2 \\
&\overset{(203)}{\leq} r_0^2 + \frac{2C}{(N+1)^{3/2}}\sum_{k=0}^{l-1}(k+2)^{1/2}r_{k+1}^2 \\
&\overset{(204)}{\leq} r_0^2 + \frac{1}{2(N+1)^{3/2}}l\cdot(l+1)^{1/2}\cdot 4r_0^2 \leq 3r_0^2,
\end{aligned}$$

which implies $r_{l+1} \leq 2r_0$. □

Lemma 16. 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

$$A_l r_l^2 + \sum_{k=0}^{l-1} A_k \tilde{r}_k^2 \leq A r_0^2 + \frac{B r_0}{N\sqrt{A_N}} \sum_{k=0}^{l-1} \alpha_{k+1}(r_k + \tilde{r}_k), \quad \forall l = 1, \ldots, N, \tag{205}$$

where $\tilde{r}_0 = 0$, $A_0 = \alpha_0 > 0$, $A_l = A_{l-1} + \alpha_l$ and $\alpha_l \leq D A_{l-1}$ for $l = 1, \ldots, N$ and $D \geq 1$. Then for all $l = 1, \ldots, N$ we have

$$r_l \leq \frac{C r_0}{\sqrt{A_l}}, \quad \tilde{r}_{l-1} \leq \frac{C r_0}{\sqrt{A_{l-1}}} \tag{206}$$

and $r_0 \leq \frac{C r_0}{\sqrt{A_0}}$ where C is such positive number that

$$C \geq \max\left\{\sqrt{A_0}, \frac{BD}{2} + \sqrt{\frac{B^2D^2}{4} + A + 2BCD}\right\},$$

i.e. one can choose $C = \max\left\{\sqrt{A_0}, \frac{3BD + \sqrt{9B^2D^2 + 4A}}{2}\right\}$.

Proof. We prove (206) by induction. For $l = 1$ the inequality $\tilde{r}_0 \leq \frac{C r_0}{\sqrt{A_0}}$ trivially follows since $\tilde{r}_0 = 0$. What is more, (205) implies that

$$A_1 r_1^2 \leq A r_0^2 + \frac{B\alpha_1 r_0^2}{N\sqrt{A_N}} \implies r_1 \leq r_0 \sqrt{\frac{A}{A_1} + \frac{BDA_0}{A_1 N\sqrt{A_N}}} \leq r_0 \sqrt{\frac{A + BD\sqrt{A_0}}{A_1}} \leq \frac{C r_0}{\sqrt{A_1}},$$

since $C \geq \sqrt{A_0}$ and $C \geq \sqrt{A + BCD} \geq \sqrt{A + BD\sqrt{A_0}}$. Note that we also have $r_0 \leq \frac{C r_0}{\sqrt{A_0}}$. Next we assume that (206) holds for some $l \leq N - 1$ and prove it for $l + 1$:

$$\begin{aligned}
A_l \tilde{r}_l^2 &\overset{(205)}{\leq} A r_0^2 + \frac{B r_0}{N\sqrt{A_N}} \sum_{k=0}^{l} \alpha_{k+1}(r_k + \tilde{r}_k) \\
&\overset{(206)}{\leq} A r_0^2 + \frac{B C r_0^2}{N\sqrt{A_N}} \sum_{k=0}^{l} \frac{\alpha_{k+1}}{\sqrt{A_k}} + \frac{B C r_0^2}{N\sqrt{A_N}} \sum_{k=0}^{l-1} \frac{\alpha_{k+1}}{\sqrt{A_k}} + \frac{B r_0 \alpha_{l+1} \tilde{r}_l}{N\sqrt{A_N}} \\
&\leq A r_0^2 + \frac{BCD r_0^2}{N\sqrt{A_N}} \sum_{k=0}^{l} \sqrt{A_k} + \frac{BCD r_0^2}{N\sqrt{A_N}} \sum_{k=0}^{l-1} \sqrt{A_k} + \frac{BD r_0 A_l \tilde{r}_l}{\sqrt{A_N}} \\
&\leq A r_0^2 + \frac{BCD r_0^2}{N\sqrt{A_N}} (l+1)\sqrt{A_l} + \frac{BCD r_0^2}{N\sqrt{A_N}} l\sqrt{A_{l-1}} + \frac{BD r_0 A_l \tilde{r}_l}{\sqrt{A_N}} \\
&\leq (A + 2BCD) r_0^2 + \frac{BD r_0 A_l \tilde{r}_l}{\sqrt{A_N}} \\
0 &\geq \tilde{r}_l^2 - \frac{BD r_0 \tilde{r}_l}{\sqrt{A_N}} - \frac{(A + 2BCD) r_0^2}{A_l}.
\end{aligned}$$

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.

$$\begin{aligned}\tilde{r}_l &\leq \frac{BDr_0}{2\sqrt{A_N}} + \sqrt{\frac{B^2D^2r_0^2}{4A_N} + \frac{(A+2BCD)r_0^2}{A_l}} \\ &\leq \underbrace{\left(\frac{BD}{2} + \sqrt{\frac{B^2D^2}{4} + A + 2BCD}\right)}_{\leq C} \frac{r_0}{\sqrt{A_l}} \\ &\leq \frac{Cr_0}{\sqrt{A_l}}.\end{aligned}$$

It implies that

$$\begin{aligned}A_{l+1}r_{l+1}^2 &\overset{(205)}{\leq} Ar_0^2 + \frac{Br_0}{N\sqrt{A_N}}\sum_{k=0}^{l}\alpha_{k+1}(r_k + \tilde{r}_k) \\ &\overset{(206)}{\leq} Ar_0^2 + \frac{2BCr_0^2}{N\sqrt{A_N}}\sum_{k=0}^{l}\frac{\alpha_{k+1}}{\sqrt{A_k}} \\ &\leq Ar_0^2 + \frac{2BCDr_0^2}{N\sqrt{A_N}}(l+1)\sqrt{A_l} \leq Ar_0^2 + 2BCDr_0^2, \\ r_{l+1} &\leq r_0\sqrt{\frac{A+2BCD}{A_{l+1}}} \leq \frac{Cr_0}{\sqrt{A_{l+1}}}.\end{aligned}$$

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.

□