J Optim Theory Appl (2016) 171:121–145
DOI 10.1007/s10957-016-0999-6

# Stochastic Intermediate Gradient Method for Convex Problems with Stochastic Inexact Oracle

**Pavel Dvurechensky**[1,2] · **Alexander Gasnikov**[2,3]

Received: 8 January 2016 / Accepted: 16 August 2016 / Published online: 23 August 2016

**Abstract** In this paper, we introduce new methods for convex optimization problems with stochastic inexact oracle. Our first method is an extension of the Intermediate Gradient Method proposed by Devolder, Glineur and Nesterov for problems with deterministic inexact oracle. Our method can be applied to problems with composite objective function, both deterministic and stochastic inexactness of the oracle, and allows using a non-Euclidean setup. We estimate the rate of convergence in terms of the expectation of the non-optimality gap and provide a way to control the probability of large deviations from this rate. Also we introduce two modifications of this method for strongly convex problems. For the first modification, we estimate the rate of convergence for the non-optimality gap expectation and, for the second, we provide a bound for the probability of large deviations from the rate of convergence in terms of the expectation of the non-optimality gap. All the rates lead to the complexity estimates for the proposed methods, which up to a multiplicative constant coincide with the lower complexity bound for the considered class of convex composite optimization problems with stochastic inexact oracle.

Communicated by Yurii Nesterov.

✉ Pavel Dvurechensky
pavel.dvurechensky@wias-berlin.de

Alexander Gasnikov
gasnikov@yandex.ru

[1] Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstr. 39, 10117 Berlin, Germany

[2] Institute for Information Transmission Problems RAS, Bolshoy Karetny per. 19, build.1, Moscow, Russia 127051

[3] Moscow Institute of Physics and Technology, 9 Institutskiy per., Dolgoprudny, Moscow Region, Russia 141700

**Keywords** Convex optimization · Inexact oracle · Rate of convergence · Stochastic optimization

**Mathematics Subject Classification** 90C30 · 90C15 · 90C25

## 1 Introduction

In this paper, we introduce new first-order methods for problems, which belong to a rather wide class of convex composite optimization problems with stochastic inexact oracle. First-order methods are widely developed since the earliest years of optimization theory; see, e.g., [1,2]. The book [3] started an activity of providing upper complexity bounds for optimization methods and lower complexity bounds for different classes of problems (see also [4]). Later, ellipsoid methods (e.g., [5]) and interior-point methods [6] were proposed for convex problems with special structure. These methods possess very fast convergence rate but have rather costly iterations, which require the number of arithmetic operations proportional to cube or fourth power of the space dimension [4]. This makes them hardly applicable for general problems with dimensions greater than ten thousands. In the last decade, problems of large and huge dimension [7] have become one of the main focuses of research in optimization methods. The reason is large amount of application areas, such as telecommunications, the Internet, traffic flows, machine learning and mechanical design. Usually in these areas, the requirements for the precision of the approximation of the optimal value are not very strong. This allows to use first-order methods, which converge slower, but have nearly dimension-independent rate of convergence, and each of their iteration requires the number of arithmetic operations proportional to the square of the space dimension or less.

In some problems, e.g., bi-level optimization, a concept of inexact oracle naturally arises. This means that the value of the objective function and its subgradient are available only with an error of some kind. It could be a deterministic error and a random error. The recent work [8] considers the case of deterministic error. The authors propose the Dual Gradient Method and the Fast Gradient Method with inexact oracle and show that the first one does not accumulate the error of the oracle, and the second has faster rate of convergence, but accumulates the oracle error. In [9], the same authors propose the Intermediate Gradient Method. This method allows to choose the trade-off between the rate of convergence and the rate of the oracle error accumulation by choosing an appropriate value of some parameter. In the thesis [10], all the mentioned above methods are extended for non-Euclidean setup. On the other hand in [11,12], the authors construct a method for composite stochastic optimization problems, which is optimal for problems both with smooth and non-smooth objective functions, but they do not consider any deterministic error of the oracle.

In this paper, we consider the general framework of stochastic inexact oracle introduced in [10]. This means, that both stochastic and deterministic errors are present in the oracle information, and the methods we develop can solve more general problems than the methods proposed in [9,11,12]. Unlike [10], our method has more flexibility in using the trade-off between the rate of convergence and the rate of the oracle error accumulation. In contrast to [9,10], we develop two modifications of our method, which converge faster under additional assumption of strong convexity of the objective function. Note that the considered class of problems with stochastic inexact oracle is very wide and includes, for example, problems of stochastic optimization, smooth and non-smooth problems (see [8]), problems with an error in the gradient of the objective function, such as LASSO [13].

The paper is organized as follows. In Sect. 2, we introduce the problem and provide the definition of stochastic inexact oracle. In Sect. 3, we generalize the Intermediate Gradient Method [9] for the case of composite optimization problems [14] with stochastic oracle error. The result is the Stochastic Intermediate Gradient Method (Algorithm 1), which also can be used in non-Euclidean setup. We estimate its rate of convergence in terms of the non-optimality gap expectation (Theorem 3.3). With some so-called light-tail assumption about the nature of the stochastic error, we obtain (Theorem 3.4) a bound for the probability of large deviations from the rate given in the previous theorem. In Sect. 4, we propose an accelerated method for strongly convex problems (Algorithm 2) and estimate its rate of convergence (Theorem 4.1). Finally, we introduce Algorithm 3, which allows to control the probability of large deviations from the rate of convergence in terms of the non-optimality gap expectation for strongly convex case (Theorem 4.2).

## 2 Notation and Terminology

Let $E$ be a finite-dimensional real vector space and $E^*$ be its dual. We denote the value of a linear function $g \in E^*$ at $x \in E$ by $\langle g, x\rangle$. Let $\|\cdot\|$ be some norm on $E$. By $\partial f(x)$, we denote the subdifferential of the function $f(x)$ at a point $x$. In this paper, we consider the *composite optimization* problem of the form

$$\min_{x\in Q}\{\varphi(x) := f(x) + h(x)\}, \tag{1}$$

where $Q \subset E$ is a closed and convex set, $h(x)$ is a simple convex function, and $f(x)$ is a convex function with *stochastic inexact oracle*. This means that, for every $x \in Q$, there exist $f_{\delta,L}(x) \in \mathbb{R}$ and $g_{\delta,L}(x) \in E^*$, such that

$$0 \le f(y) - f_{\delta,L}(x) - \langle g_{\delta,L}(x), y - x\rangle \le \frac{L}{2}\|x - y\|^2 + \delta, \quad \forall y \in Q, \tag{2}$$

and also that, instead of $(f_{\delta,L}(x), g_{\delta,L}(x))$ (we will call this pair a $(\delta, L)$-oracle), we use their stochastic approximations $(F_{\delta,L}(x,\xi), G_{\delta,L}(x,\xi))$. The latter means that, for any point $x \in Q$, we associate with $x$ a random variable $\xi$ whose probability distribution is supported on a set $\Xi \subset \mathbb{R}$ and such that $\mathbb{E}_\xi F_{\delta,L}(x,\xi) = f_{\delta,L}(x)$, $\mathbb{E}_\xi G_{\delta,L}(x,\xi) = g_{\delta,L}(x)$ and

$$\mathbb{E}_\xi(\|G_{\delta,L}(x,\xi) - g_{\delta,L}(x)\|_*)^2 \le \sigma^2. \tag{3}$$

Here $\|\cdot\|_*$ is the dual norm corresponding to $\|\cdot\|_E$, i.e., $\|g\|_* = \sup_{y\in E}\{\langle g, y\rangle : \|y\|_E \leq 1\}$.

To deal with such problems, we will need a *prox-function* $d(x)$, which is differential and strongly convex with parameter 1 on $Q$ with respect to $\|\cdot\|$. Let $x_0$ be the minimizer of $d(x)$ on $Q$. By translating and scaling $d(x)$, if necessary, we can always ensure that

$$d(x_0) = 0, \quad d(x) \geq \frac{1}{2}\|x - x_0\|^2, \quad \forall x \in Q. \tag{4}$$

We define also the corresponding *Bregman distance*:

$$V(x, z) = d(x) - d(z) - \langle \nabla d(z), x - z\rangle. \tag{5}$$

Due to the strong convexity of $d(x)$ with parameter 1, we have:

$$V(x, z) \geq \frac{1}{2}\|x - z\|^2, \quad \forall x, z \in Q. \tag{6}$$

## 3 Stochastic Intermediate Gradient Method

Let $\{\alpha_i\}_{i\geq 0}, \{\beta_i\}_{i\geq 0}, \{B_i\}_{i\geq 0} \subset \mathbb{R}$ be three sequences of coefficients satisfying

$$\alpha_0 \in ]0, 1], \quad \beta_{i+1} \geq \beta_i > L, \quad \forall i \geq 0, \tag{7}$$

$$0 \leq \alpha_i \leq B_i, \quad \forall i \geq 0, \tag{8}$$

$$\alpha_k^2 \beta_k \leq B_k \beta_{k-1} \leq \left(\sum_{i=0}^{k} \alpha_i\right)\beta_{k-1}, \quad \forall k \geq 1. \tag{9}$$

We define also

$$A_k := \sum_{i=0}^{k} \alpha_i, \tag{10}$$

$$\tau_i := \frac{\alpha_{i+1}}{B_{i+1}}. \tag{11}$$

Note that, by definition, $\alpha_0 = A_0 = B_0$. The Stochastic Intermediate Gradient Method is described below as Algorithm 1.

### 3.1 General Convergence Rate

Let us obtain the convergence rate of the proposed method in terms of the sequences $A_i$, $B_i$ and $\beta_i$, $i \geq 0$. Denote by

**ALGORITHM 1:** Stochastic Intermediate Gradient Method (SIGM)

```
Input: The sequences $\{\alpha_i\}_{i\geq 0}$, $\{\beta_i\}_{i\geq 0}$, $\{B_i\}_{i\geq 0}$, functions $d(x)$, $V(x,z)$.
Output: The point $y_k$.
Compute $x_0 := \arg\min_{x\in Q}\{d(x)\}$.
Let $\xi_0$ be a realization of the random variable $\xi$. Calculate $G_{\delta,L}(x_0,\xi_0)$.
Find
    $y_0 := \arg\min_{x\in Q}\{\beta_0 d(x) + \alpha_0\langle G_{\delta,L}(x_0,\xi_0), x - x_0\rangle + \alpha_0 h(x)\}.$   (12)
Set $k = 0$.
repeat
    Find
        $z_k := \arg\min_{x\in Q}\{\beta_k d(x) + \sum_{i=0}^{k}\alpha_i\langle G_{\delta,L}(x_i,\xi_i), x - x_i\rangle + A_k h(x)\}.$   (13)
    Let
        $x_{k+1} := \tau_k z_k + (1-\tau_k)y_k.$   (14)
    Let $\xi_{k+1}$ be a realization of the random variable $\xi$. Calculate $G_{\delta,L}(x_{k+1},\xi_{k+1})$.
    Find
        $\hat{x}_{k+1} := \arg\min_{x\in Q}\{\beta_k V(x,z_k) + \alpha_{k+1}\langle G_{\delta,L}(x_{k+1},\xi_{k+1}), x - z_k\rangle + \alpha_{k+1}h(x).\}.$   (15)
    Let
        $w_{k+1} := \tau_k\hat{x}_{k+1} + (1-\tau_k)y_k.$   (16)
    Let
        $y_{k+1} := \frac{A_{k+1}-B_{k+1}}{A_{k+1}}y_k + \frac{B_{k+1}}{A_{k+1}}w_{k+1}.$   (17)
until;
```

$$\Psi_k(x) := \beta_k d(x) + \sum_{i=0}^{k}\alpha_i\left[F_{\delta,L}(x_i,\xi_i) + \langle G_{\delta,L}(x_i,\xi_i), x - x_i\rangle\right] + A_k h(x), \quad (18)$$

the *model* of the objective function, $\Psi_k^* := \min_{x\in Q}\Psi_k(x)$ its minimal value on the feasible set and $\xi_{[k]} := (\xi_0,\dots,\xi_k)$ the history of the random process after $k$ iterations. Let us show that $\{y_k\}_{k\geq 0}$ and $\{\Psi_k(x)\}_{k\geq 0}$ define a sequence of estimate functions. We denote $f_i := f_{\delta,L}(x_i)$, $g_i := g_{\delta,L}(x_i)$, $F_i := F_{\delta,L}(x_i,\xi_i)$, $G_i := G_{\delta,L}(x_i,\xi_i)$.

**Lemma 3.1** *For all $k \geq 0$, the following inequality holds*

$$A_k\varphi(y_k) \leq \Psi_k^* + E_k, \quad (19)$$

*where*

$$E_k := \sum_{i=0}^{k} B_i \delta + \sum_{i=0}^{k} \frac{B_i}{\beta_i - L} (\|G_i - g_i\|_*)^2 + \sum_{i=0}^{k} \alpha_i (f_i - F_i) + \sum_{i=1}^{k} (B_i - \alpha_i) \frac{\alpha_i}{B_i} \langle g_i - G_i, z_{i-1} - y_{i-1} \rangle.$$

*Proof* The proof is rather technical and can be found in "Appendix A." □

**Lemma 3.2** *For all $k \geq 0$, the following inequality holds*

$$\Psi_k(x) \leq A_k \varphi(x) + \beta_k d(x) + \bar{E}_k(x), \quad \forall x \in Q, \tag{20}$$

*where* $\bar{E}_k(x) := \sum_{i=0}^{k} \alpha_i \big[F_i - f_i + \langle G_i - g_i, x - x_i \rangle\big]$.

*Proof* We have:

$$\begin{aligned} \Psi_k(x) &= \beta_k d(x) + \textstyle\sum_{i=0}^{k} \alpha_i \big[f_i + \langle g_i, x - x_i \rangle\big] \\ &\quad + \textstyle\sum_{i=0}^{k} \alpha_i \big[F_i - f_i + \langle G_i - g_i, x - x_i \rangle\big] + A_k h(x) \\ &\overset{(2)}{\leq} \beta_k d(x) + A_k \varphi(x) + \bar{E}_k(x). \end{aligned}$$

□

Combining Lemmas 3.1 and 3.2, we obtain the following result.

**Theorem 3.1** *Assume that the function $f$ is endowed with a stochastic inexact oracle with noise level $\sigma$, bias $\delta$ and constant $L$. Then the sequence $y_k$ generated by the Algorithm 1, when applied to the Problem (1), satisfies*

$$\varphi(y_k) - \varphi^* \leq \frac{1}{A_k} \Bigg( \beta_k d(x^*) + \sum_{i=0}^{k} B_i \delta + \sum_{i=0}^{k} \frac{B_i}{\beta_i - L} \|G_i - g_i\|_*^2 + \sum_{i=0}^{k} \alpha_i \langle G_i - g_i, x^* - x_i \rangle + \sum_{i=1}^{k} (B_i - \alpha_i) \frac{\alpha_i}{B_i} \langle G_i - g_i, y_{i-1} - z_{i-1} \rangle \Bigg). \tag{21}$$

*Moreover,*

$$\mathbb{E}_{\xi_0,\ldots,\xi_k} \varphi(y_k) - \varphi^* \leq \frac{\beta_k d(x^*)}{A_k} + \frac{\sum_{i=0}^{k} B_i \delta}{A_k} + \frac{1}{A_k} \sum_{i=0}^{k} \frac{B_i}{\beta_i - L} \sigma^2.$$

*Proof* From the inequalities (19) and (20), by the definition of $\Psi_k(x)$ and $\Psi_k^*$, we have:

$$A_k \varphi(y_k) \leq \Psi_k^* + E_k \leq \Psi_k(x^*) + E_k \leq A_k \varphi^* + \beta_k d(x^*) + \bar{E}_k(x^*) + E_k,$$

which immediately gives the first statement of the theorem.

Since $\mathbb{E}_{\xi_i}\left[G_i|\xi_{[i-1]}\right] = g_i$ and $x_i$, $y_{i-1}$, and $z_{i-1}$ are deterministic functions of $(\xi_0, \dots, \xi_{i-1})$, we have

$$\mathbb{E}_{\xi_i}\left[\langle G_i - g_i, x^* - x_i\rangle|\xi_{[i-1]}\right] = \mathbb{E}_{\xi_i}\left[\langle G_i - g_i, y_{i-1} - z_{i-1}\rangle|\xi_{[i-1]}\right] = 0.$$

Therefore, the expectation of fourth and fifth terms in (21) with respect to $\xi_0, \dots, \xi_k$ is zero. Also, by our assumption, $\mathbb{E}_{\xi_i}\left[\|G_i - g_i\|_*^2|\xi_{[i-1]}\right] \le \sigma^2$, and, hence, $\mathbb{E}_{\xi_0,\dots,\xi_k}\left[\sum_{i=0}^k \frac{B_i}{\beta_i - L}\|G_i - g_i\|_*^2\right] \le \sum_{i=0}^k \frac{B_i}{\beta_i - L}\sigma^2$. This proves the second part of the theorem. □

### 3.2 General Probability of Large Deviations

In this section, we obtain an upper bound on the probability of large deviations for the $\varphi(y_k) - \varphi^*$. To obtain our results, we make the following additional assumptions.

1. $\xi_0, \dots, \xi_k$ are i.i.d random variables.
2. $G_{\delta,L}(x, \xi)$ satisfies the light-tail condition

$$\mathbb{E}_\xi\left[\exp\left(\frac{\|G_{\delta,L}(x,\xi) - g_{\delta,L}(x)\|_*^2}{\sigma^2}\right)\right] \le \exp(1).$$

3. Set $Q$ is bounded, and we know a number $D > 0$, such that $\max_{x,y\in Q}\|x - y\| \le D$.

**Lemma 3.3** *([15], [10]) Let $\xi_0, \dots, \xi_k$ be a sequence of realizations of the i.i.d. random variables $X_0, \dots, X_k$ and let $\Delta_i := \Delta_i(\xi_{[i]})$ be a deterministic function of $\xi_{[i]}$ such that, for all $i \ge 0$, $\mathbb{E}\left[\exp\left(\frac{\Delta_i^2}{\sigma^2}\right)|\xi_{[i-1]}\right] \le \exp(1)$, and $c_0, \dots, c_k$ be a sequence of positive coefficients. Then we have for any $k \ge 0$ and any $\Omega \ge 0$:*

$$\mathbb{P}\left(\sum_{i=0}^k c_i\Delta_i^2 \ge (1+\Omega)\sum_{i=0}^k c_i\sigma^2\right) \le \exp(-\Omega).$$

**Lemma 3.4** *([16], [10]) Let $\xi_0, \dots, \xi_k$ be a sequence of realizations of i.i.d. random variables $X_0, \dots, X_k$ and let $\Gamma_i$ and $\eta_i$ be deterministic functions of $\xi_{[i]}$ such that 1) $\mathbb{E}\left[\Gamma_i|\xi_{[i-1]}\right] = 0$; 2) $|\Gamma_i| \le c_i\eta_i$, where $c_i$ is positive deterministic constant; and 3) $\mathbb{E}\left[\exp\left(\frac{\eta_i^2}{\sigma^2}\right)|\xi_{[i-1]}\right] \le \exp(1)$. Then, for any $k \ge 0$ and any $\Omega \ge 0$,*

$$\mathbb{P}\left(\sum_{i=0}^k \Gamma_i \ge \sqrt{3\Omega}\sigma\sqrt{\sum_{i=0}^k c_i^2}\right) \le \exp(-\Omega).$$

**Theorem 3.2** *If the assumptions 1, 2, 3 are satisfied, then, for all $k \geq 0$ and all $\Omega \geq 0$, the sequence generated by the SIGM satisfies*

$$\mathbb{P}\Bigg(\varphi(y_k) - \varphi^* \geq \frac{\beta_k d(x^*)}{A_k} + \frac{\sum_{i=0}^k B_i \delta}{A_k} + \frac{1+\Omega}{A_k}\sum_{i=0}^k \frac{B_i}{\beta_i - L}\sigma^2 + \frac{2D\sigma\sqrt{3\Omega}}{A_k}\sqrt{\sum_{i=0}^k \alpha_i^2}\Bigg) \leq 3\exp(-\Omega).$$

*Proof* From Theorem 3.1, we know that for the SIGM, the gap $\varphi(y_k) - \varphi^*$ can be bounded from above by the sum of four quantities:

1. deterministic $I_1(k) := \frac{\beta_k d(x^*)}{A_k} + \frac{\sum_{i=0}^k B_i \delta}{A_k}$,
2. random $I_2(k, \xi_{[k]}) := \frac{1}{A_k}\sum_{i=0}^k \frac{B_i}{\beta_i - L}\|G_i - g_i\|_*^2$,
3. random $I_3(k, \xi_{[k]}) := \frac{1}{A_k}\sum_{i=1}^k (B_i - \alpha_i)\frac{\alpha_i}{B_i}\langle G_i - g_i, y_{i-1} - z_{i-1}\rangle$,
4. random $I_4(k, \xi_{[k]}) := \frac{1}{A_k}\sum_{i=0}^k \alpha_i \langle G_i - g_i, x^* - x_i\rangle$.

For $I_2(k, \xi_{[k]})$, using Lemma 3.3 with $\Delta_i = \|G_i - g_i\|_*$ and $c_i = \frac{B_i}{A_k(\beta_i - L)}$, we obtain, for all $k \geq 0$ and $\Omega \geq 0$,

$$\mathbb{P}\left(I_2(k, \xi_{[k]}) \geq \frac{1+\Omega}{A_k}\sum_{i=0}^k \frac{B_i}{\beta_i - L}\sigma^2\right) \leq \exp(-\Omega).$$

For $I_3(k, \xi_{[k]})$, using Lemma 3.4 with $\Gamma_i = (B_i - \alpha_i)\frac{\alpha_i}{A_k B_i}\langle G_i - g_i, y_{i-1} - z_{i-1}\rangle$, $\eta_i = \|G_i - g_i\|_*$ and $c_i = \frac{\alpha_i D}{A_k}$, we obtain, for all $k \geq 0$ and $\Omega \geq 0$,

$$\mathbb{P}\left(I_3(k, \xi_{[k]}) \geq \frac{D\sigma\sqrt{3\Omega}}{A_k}\sqrt{\sum_{i=1}^k \alpha_i^2}\right) \leq \exp(-\Omega).$$

For $I_4(k, \xi_{[k]})$, using Lemma 3.4 with $\Gamma_i = \frac{\alpha_i}{A_k}\langle G_i - g_i, x^* - x_i\rangle$, $\eta_i = \| G_i - g_i \|_*$ and $c_i = \frac{\alpha_i D}{A_k}$, we obtain, for all $k \geq 0$ and $\Omega \geq 0$,

$$\mathbb{P}\left(I_4(k, \xi_{[k]}) \geq \frac{D\sigma\sqrt{3\Omega}}{A_k}\sqrt{\sum_{i=0}^k \alpha_i^2}\right) \leq \exp(-\Omega).$$

Combining these results, we obtain the statement of the theorem. □

### 3.3 Choice of the Coefficients

In Theorem 3.1, we have obtained the rate of convergence for the SIGM in terms of the non-optimality gap expectation, and in Theorem 3.2, we have obtained the bound on probability of large deviations for the non-optimality gap. These results are formulated in terms of the sequences $\{\alpha_i\}_{i\geq 0}$, $\{\beta_i\}_{i\geq 0}$, $\{B_i\}_{i\geq 0}$ satisfying (7), (8) and (9). In this section, we specify these sequences in order to obtain the rate of convergence $\mathbb{E}\varphi(y_k)-\varphi^* \leq \Theta\left(\frac{LR^2}{k^p}+\frac{\sigma R}{\sqrt{k}}+k^{p-1}\delta\right)$, where $p\in[0,1]$ can be chosen before the start of the algorithm. Let $a\geq 1$ and $b\geq 0$ be some parameters. Let us assume that we know a number $R$ such that $\sqrt{2d(x^*)}\leq R$. We set

$$\alpha_i = \frac{1}{a}\left(\frac{i+p}{p}\right)^{p-1}, \quad \forall i\geq 0, \tag{22}$$

$$\beta_i = L+\frac{b\sigma}{R}(i+p+1)^{\frac{2p-1}{2}}, \quad \forall i\geq 0, \tag{23}$$

$$B_i = a\alpha_i^2 = \frac{1}{a}\left(\frac{i+p}{p}\right)^{2p-2}, \quad \forall i\geq 0. \tag{24}$$

Then inequalities (7) and (8) hold, and we only need to check that (9) also holds. We have

$$A_k=\sum_{i=0}^{k}\alpha_i \geq \frac{1}{a}\int_0^k\left(\frac{x+p}{p}\right)^{p-1}dx+\alpha_0 \geq \frac{1}{a}\left(\frac{k+p}{p}\right)^p. \tag{25}$$

Clearly, for any $i\geq 0$,

$$\alpha_k^2=\frac{1}{a^2}\left(\frac{k+p}{p}\right)^{2p-2} \leq \frac{1}{a}\left(\frac{k+p}{p}\right)^{2p-2} \leq \frac{1}{a}\left(\frac{k+p}{p}\right)^{p} \leq A_k.$$

If we choose $a=2^{\frac{2p-1}{2}}$, then

$$\frac{1}{a^2}\left(\frac{k+p}{p}\right)^{2p-2}(k+p+1)^{\frac{2p-1}{2}} \\ \leq \frac{1}{a}\left(\frac{k+p}{p}\right)^{2p-2}(k+p)^{\frac{2p-1}{2}} \leq \frac{1}{a}\left(\frac{k+p}{p}\right)^{p}(k+p)^{\frac{2p-1}{2}}.$$

Last two sequences of inequalities prove that (9) holds. Using (25), we have

$$\frac{\beta_k d(x^*)}{A_k} \leq \frac{\beta_k R^2}{2A_k} \leq \left(L+\frac{b\sigma}{R}(k+p+1)^{\frac{2p-1}{2}}\right)R^2 2^{\frac{2p-3}{2}}\left(\frac{p}{k+p}\right)^p. \tag{26}$$

Also, using (25) and the fact that $p\in[1,2]$, we have the following chain of inequalities

$$\frac{\delta}{A_k}\sum_{i=0}^{k}B_i=\frac{a\delta}{A_k}\sum_{i=0}^{k}\alpha_i^2 \leq \frac{a\delta}{A_k}\left(\int_0^k\left(\frac{x+p}{p}\right)^{2p-2}dx+\left(\frac{k+p}{p}\right)^{2p-2}\right) \\ \leq \frac{a\delta}{A_k}\left(\left(\frac{k+p}{p}\right)^{2p-1}+\left(\frac{k+p}{p}\right)^{2p-2}\right)$$

$$\leq 2^{2p-1}\delta\left(\frac{p}{k+p}\right)^p\left(\left(\frac{k+p}{p}\right)^{2p-1}+\left(\frac{k+p}{p}\right)^{2p-2}\right)$$
$$\leq 2^{2p-1}\left(\left(\frac{k+p}{p}\right)^{p-1}+1\right)\delta. \tag{27}$$

Again, by (25), we have the following inequalities

$$\frac{\sigma^2}{A_k}\sum_{i=0}^{k}\frac{B_i}{\beta_i-L} \leq \frac{\sigma R}{bp^{2p-2}}\left(\frac{p}{k+p}\right)^p\sum_{i=0}^{k}\frac{(i+p)^{2p-2}}{(i+p+1)^{\frac{2p-1}{2}}}$$
$$\leq \frac{\sigma R p^{2-p}}{b(k+p)^p}\sum_{i=0}^{k}(i+p+1)^{p-\frac{3}{2}} \leq \frac{\sigma R p^{2-p}}{b(k+p)^p}\int_1^{k+1}(x+p+1)^{p-\frac{3}{2}}dx$$
$$\leq \frac{\sigma R p^{2-p}}{b(p-\frac{1}{2})}\frac{(k+p+2)^{p-\frac{1}{2}}}{(k+p)^p}. \tag{28}$$

Combining the estimates (26), (27) and (28), we get for the rate of convergence obtained in Theorem 3.1

$$\mathbb{E}_{\xi_0,\ldots,\xi_k}\varphi(y_k)-\varphi^* \leq \left(L+\frac{b\sigma}{R}(k+p+1)^{\frac{2p-1}{2}}\right)R^2 2^{\frac{2p-3}{2}}\left(\frac{p}{k+p}\right)^p$$
$$+\frac{\sigma R p^{2-p}}{b(p-\frac{1}{2})}\frac{(k+p+2)^{p-\frac{1}{2}}}{(k+p)^p}+2^{2p-1}\left(\left(\frac{k+p}{p}\right)^{p-1}+1\right)\delta$$
$$\leq \frac{LR^2p^p2^{\frac{2p-3}{2}}}{(k+p)^p}+\frac{\sigma R(k+p+2)^{p-\frac{1}{2}}}{(k+p)^p}\left(b2^{p-\frac{3}{2}}p^p+\frac{2p^{1-p}}{b}\right)$$
$$+2^{2p-1}\left(\left(\frac{k+p}{p}\right)^{p-1}+1\right)\delta.$$

Choosing optimal $b=2^{\frac{5-2p}{4}}p^{\frac{1-2p}{2}}$, we get the following theorem.

**Theorem 3.3** *If the sequences* $\{\alpha_i\}_{i\geq 0}$, $\{\beta_i\}_{i\geq 0}$, $\{B_i\}_{i\geq 0}$ *are chosen according to* (22)*,* (23)*,* (24) *with* $a=2^{\frac{2p-1}{2}}$ *and* $b=2^{\frac{5-2p}{4}}p^{\frac{1-2p}{2}}$*, then the sequence* $y_k$ *generated by the SIGM satisfies*

$$\mathbb{E}_{\xi_0,\ldots,\xi_k}\varphi(y_k)-\varphi^* \leq \frac{LR^2p^p2^{\frac{2p-3}{2}}}{(k+p)^p}+\frac{\sigma R2^{\frac{3+2p}{4}}\sqrt{p}(k+p+2)^{p-\frac{1}{2}}}{(k+p)^p}+$$
$$+2^{2p-1}\left(\left(\frac{k+p}{p}\right)^{p-1}+1\right)\delta \leq \frac{C_1LR^2}{k^p}+\frac{C_2\sigma R}{\sqrt{k}}+C_3k^{p-1}\delta$$
$$=\Theta\left(\frac{LR^2}{k^p}+\frac{\sigma R}{\sqrt{k}}+k^{p-1}\delta\right),$$

*where* $C_1 = 4\sqrt{2}, C_2 = 16\sqrt{2}, C_3 = 48.$

Similarly to what we have done to prove (27), we obtain the following inequality $\frac{1}{A_k^2}\sum_{i=0}^{k}\alpha_i^2 \le \frac{2p}{k+p}$. Combining this inequality with (26), (27) and (28), we prove the following corollary of Theorem 3.2.

**Theorem 3.4** *If the sequences* $\{\alpha_i\}_{i\ge 0}, \{\beta_i\}_{i\ge 0}, \{B_i\}_{i\ge 0}$ *are chosen according to* (22), (23) *and* (24) *with* $a = 2^{\frac{2p-1}{2}}$ *and* $b = 2^{\frac{5-2p}{4}} p^{\frac{1-2p}{2}}$, *then the sequence* $y_k$ *generated by the SIGM satisfies*

$$\mathbb{P}\left(\varphi(y_k) - \varphi^* > \frac{C_1 L R^2}{k^p} + \frac{C_2(1+\Omega)\sigma R}{\sqrt{k}} + C_3 k^{p-1}\delta + \frac{C_4 D\sigma\sqrt{\Omega}}{\sqrt{k}}\right)$$
$$\le \mathbb{P}\left(\varphi(y_k) - \varphi^* > \frac{LR^2 p^p 2^{\frac{2p-3}{2}}}{(k+p)^p} + \frac{(1+\Omega)\sigma R 2^{\frac{3+2p}{4}}\sqrt{p}(k+p+2)^{p-\frac{1}{2}}}{(k+p)^p}\right.$$
$$\left. + 2^{2p-1}\left(\left(\frac{k+p}{p}\right)^{p-1} + 1\right)\delta + \frac{2D\sigma\sqrt{6\Omega p}}{\sqrt{k+p}}\right) \le 3\exp(-\Omega),$$

*where* $C_1 = 4\sqrt{2}, C_2 = 16\sqrt{2}, C_3 = 48, C_4 = 4\sqrt{3}.$

## 4 Stochastic Intermediate Gradient Method for Strongly Convex Problems

In this section, we consider two modifications of the SIGM for strongly convex problems. For the first modification, we obtain the rate of convergence in terms of the non-optimality gap expectation, and for the second, we bound the probability of large deviations from this rate. Both modifications are based on the restart technique, which was previously used in [12] and [17].

Throughout this section, we assume that $E$ is a Euclidean space with scalar product $\langle\cdot,\cdot\rangle$ and norm $\|x\| := \sqrt{\langle x, Hx\rangle}$, where $H$ is a symmetric positive definite matrix. Without loss of generality, we assume that the function $d(x)$ satisfies conditions $0 = \arg\min_{x\in Q} d(x)$ and $d(0) = 0$. Also we assume that the function $\varphi(x)$ in (1) is strongly convex, i.e., $\frac{\mu}{2}\|x - y\|^2 \le \varphi(y) - \varphi(x) - \langle g(x), y - x\rangle$ for all $x, y \in Q, g(x) \in \partial\varphi(x)$. As a corollary, we have

$$\varphi(x) - \varphi(x^*) \ge \frac{\mu}{2}\|x - x^*\|^2, \quad \forall x \in Q, \tag{29}$$

where $x^*$ is the solution of the Problem (1).

### 4.1 Modified Algorithm with Rate of Convergence for Expectation of Non-Optimality Gap

In this subsection, we assume that $d(x)$ satisfies the following property. If $x_0$ is a random vector such that $\mathbb{E}_{x_0}\|x - x_0\|^2 \le R_0^2$ for some fixed point $x$ and number $R_0$, then, for some $V > 0$,

$$\mathbb{E}_{x_0} d\left(\frac{x - x_0}{R_0}\right) \le \frac{V^2}{2}. \tag{30}$$

This assumption is satisfied, for example, for prox-functions, which have quadratic growth with constant $V^2$. The latter means that $d(x) \le \frac{V^2}{2}\|x\|^2$ for all $x \in E$. Several examples of such prox-functions can be found in [17].

**Lemma 4.1** *Assume that we start Algorithm 1 from a random point $x_0$ such that $\mathbb{E}_{x_0}\|x^* - x_0\|^2 \le R_0^2$ and, hence, (30) holds with $x = x^*$. We use the function $d\left(\frac{x-x_0}{R_0}\right)$ as the prox-function in the algorithm. Also assume that on kth iteration of Algorithm 1, we ask the oracle m times, getting answers $G_{\delta,L}(x_{k+1}, \xi_{k+1}^i)$, $i = 1, \dots, m$, and use $\tilde{G}_{\delta,L}(x_{k+1}) := \frac{1}{m}\sum_{i=1}^m G_{\delta,L}(x_{k+1}, \xi_{k+1}^i)$ in (15) instead of $G_{\delta,L}(x_{k+1}, \xi_{k+1})$. We assume that $\xi_{k+1}^i$, $i = 1, \dots, m$ are i.i.d for fixed $k+1$. Also let the assumptions of Theorem 3.3 hold. Then*

$$\mathbb{E}\varphi(y_k) - \varphi^* \le \frac{C_1 L R_0^2 V^2}{k^p} + \frac{C_2 \sigma R_0 V}{\sqrt{mk}} + C_3 k^{p-1}\delta,$$

*where $C_1 = 4\sqrt{2}$, $C_2 = 16\sqrt{2}$, $C_3 = 48$ and the expectation is taken with respect to all the randomness.*

*Proof* Note that $d\left(\frac{x-x_0}{R_0}\right)$ is strongly convex with respect to the norm $\frac{1}{R_0}\|\cdot\|$ with parameter 1 and that the dual for this norm is the norm $R_0\|\cdot\|_*$. Also note that with respect to the norm $\frac{1}{R_0}\|\cdot\|$ $(f_{\delta,L}(x), g_{\delta,L}(x))$ is a $(\delta, LR_0^2)$-oracle for $f(x)$. Also we have $\mathbb{E}_{\xi_{k+1}^1,\dots,\xi_{k+1}^m}\tilde{G}_{\delta,L}(x_{k+1}) = g_{\delta,L}(x_{k+1})$, and

$$\begin{aligned}&\mathbb{E}_{\xi_{k+1}^1,\dots,\xi_{k+1}^m} R_0^2\|\tilde{G}_{\delta,L}(x_{k+1}) - g_{\delta,L}(x_{k+1})\|_*^2\\ &\quad = \mathbb{E}_{\xi_{k+1}^1,\dots,\xi_{k+1}^m} R_0^2\left\|\frac{1}{m}\sum_{i=1}^m G_{\delta,L}(x_{k+1}, \xi_{k+1}^i) - g_{\delta,L}(x_{k+1})\right\|_*^2 \overset{(3)}{\le} \frac{\sigma^2 R_0^2}{m}.\end{aligned}$$

Applying Theorems 3.1 and 3.3 with changing $L$ to $LR_0^2$, $\sigma$ to $\frac{\sigma R_0}{\sqrt{m}}$ and $R$ to $V$, we obtain

$$\begin{aligned}\mathbb{E}\varphi(y_k) - \varphi^* &\le \frac{\beta_k \mathbb{E}_{x_0} d\left(\frac{x^*-x_0}{R_0}\right)}{A_k} + \frac{\sum_{i=0}^k B_i\delta}{A_k} + \frac{1}{A_k}\sum_{i=0}^k \frac{B_i}{\beta_i - L}\sigma^2\\ &\le \frac{C_1 L R_0^2 V^2}{k^p} + \frac{C_2\sigma R_0 V}{\sqrt{mk}} + C_3 k^{p-1}\delta.\end{aligned}$$

□

Now we are ready to formulate the new algorithm for strongly convex problems and convergence result for this algorithm.

**ALGORITHM 2:** Stochastic Intermediate Gradient Method for Strongly Convex Problems

```
Input: The function $d(x)$, point $u_0$, number $R_0$ such that $\|u_0 - x^*\| \le R_0$, number $p \in [1, 2]$.
Output: The point $u_{k+1}$.
Set $k = 0$.
Calculate
    $$N_k := \left\lceil \left(\frac{4eC_1LV^2}{\mu}\right)^{\frac{1}{p}} \right\rceil. \tag{31}$$
repeat
    Calculate
        $$m_k := \max\left\{1, \left\lceil \frac{16e^{k+2}C_2^2\sigma^2V^2}{\mu^2R_0^2N_k} \right\rceil\right\}, \tag{32}$$
    $$R_k^2 := R_0^2e^{-k} + \frac{2^p eC_3\delta}{\mu(e-1)}\left(\frac{4eC_1LV^2}{\mu}\right)^{\frac{p-1}{p}}\left(1 - e^{-k}\right). \tag{33}$$
    Run Algorithm 1 with $x_0 = u_k$ and prox-function $d\left(\frac{x-u_k}{R_k}\right)$ for $N_k$ steps, using oracle
    $\tilde{G}^k_{\delta,L}(x) := \frac{1}{m_k}\sum_{i=1}^{m_k} G_{\delta,L}(x, \xi^i)$, where $\xi^i$, $i = 1, ..., m_k$ are i.i.d, on each step and sequences
    $\{\alpha_i\}_{i\ge0}$, $\{\beta_i\}_{i\ge0}$, $\{B_i\}_{i\ge0}$ defined in Theorem 3.3.
    Set $u_{k+1} = y_{N_k}$, $k = k + 1$.
until;
```

**Theorem 4.1** *After $k \ge 1$ outer iterations of Algorithm 2, we have*

$$\mathbb{E}\varphi(u_k) - \varphi^* \le \frac{\mu R_0^2}{2}e^{-k} + \frac{C_3e2^{p-1}}{e-1}\left(\frac{4eC_1LV^2}{\mu}\right)^{\frac{p-1}{p}}\delta, \tag{34}$$

$$\mathbb{E}\|u_k - x^*\|^2 \le R_0^2e^{-k} + \frac{C_3e2^p}{\mu(e-1)}\left(\frac{4eC_1LV^2}{\mu}\right)^{\frac{p-1}{p}}\delta. \tag{35}$$

*As a consequence, if we choose the error $\delta$ of the oracle satisfying*

$$\delta \le \frac{\varepsilon(e-1)}{2^pC_3e}\left(\frac{4eC_1LV^2}{\mu}\right)^{\frac{1-p}{p}}, \tag{36}$$

*then we need $N = \left\lceil \ln\left(\frac{\mu R_0^2}{\varepsilon}\right)\right\rceil$ outer iterations and not more than*

$$\left(1 + \left(\frac{4eC_1LV^2}{\mu}\right)^{\frac{1}{p}}\right)\left(1 + \ln\left(\frac{\mu R_0^2}{\varepsilon}\right)\right) + \frac{16e^3C_2^2\sigma^2V^2}{\mu\varepsilon(e-1)}$$

*oracle calls to guarantee that* $\mathbb{E}\varphi(u_N) - \varphi^* \leq \varepsilon$.

*Proof* The proof is rather technical and can be found in "Appendix B." □

### 4.2 Modified Algorithm with Controlled Probability of Large Deviations

In this subsection, we assume that the prox-function has quadratic growth with parameter $V^2$ with respect to the chosen norm, i.e.,

$$d(x) \leq \frac{V^2}{2}\|x\|^2, \quad \forall x \in \mathbb{R}^n. \tag{37}$$

Several examples of such prox-functions can be found in [17].

Now we present a modification of Algorithm 2 and a theorem with a bound for the probability of large deviations for the non-optimality gap of this algorithm.

**ALGORITHM 3:** Stochastic Intermediate Gradient Method for Strongly Convex Problems 2

```
Input: The function $d(x)$, point $u_0$, number $R_0$ such that $\|u_0 - x^*\| \leq R_0$, number $p \in [1, 2]$,
    number $N \geq 1$ of outer iterations, confidence level $\Lambda$.
Output: The point $u_N$.
Set $k = 0$.
Calculate
    $$N_k := \left\lceil \left( \frac{6\mathrm{e}C_1LV^2}{\mu} \right)^{\frac{1}{p}} \right\rceil. \tag{38}$$
repeat
    Calculate
        $$m_k := \max\left\{1, \left\lceil \frac{36\mathrm{e}^{k+2}C_2^2\sigma^2V^2\left(1+\ln\left(\frac{3N}{\Lambda}\right)\right)^2}{\mu^2R_0^2N_k} \right\rceil, \left\lceil \frac{144\mathrm{e}^{k+2}C_4^2\sigma^2\ln\left(\frac{3N}{\Lambda}\right)}{\mu^2R_0^2N_k} \right\rceil \right\}, \tag{39}$$
        $$R_k^2 := R_0^2\mathrm{e}^{-k} + \frac{2^p\mathrm{e}C_3\delta}{\mu(\mathrm{e}-1)}\left(\frac{6\mathrm{e}C_1LV^2}{\mu}\right)^{\frac{p-1}{p}}\left(1-\mathrm{e}^{-k}\right), \tag{40}$$
        $$Q_k := \left\{x \in Q : \|x - u_k\|^2 \leq R_k^2\right\}. \tag{41}$$
    Run Algorithm 1 applied to the problem $\min_{x \in Q_k}\varphi(x)$ with $x_0 = u_k$ and prox-function
    $d\left(\frac{x-u_k}{R_k}\right)$ for $N_k$ steps using oracle $\tilde{G}^k_{\delta,L}(x) := \frac{1}{m_k}\sum_{i=1}^{m_k} G_{\delta,L}(x, \xi^i)$, where $\xi^i$, $i = 1, ..., m_k$
    are i.i.d, on each step and sequences $\{\alpha_i\}_{i\geq 0}$, $\{\beta_i\}_{i\geq 0}$, $\{B_i\}_{i\geq 0}$ defined in Theorem 3.3.
    Set $u_{k+1} = y_{N_k}$, $k = k + 1$.
until $k = N - 1$;
```

**Theorem 4.2** *After N outer iterations of Algorithm 3, we have*

$$\mathbb{P}\left\{\varphi(u_N)-\varphi^*>\frac{\mu R_0^2}{2}\mathrm{e}^{-N}+\frac{2^{p-1}\mathrm{e}C_3\delta}{(\mathrm{e}-1)}\left(\frac{6\mathrm{e}C_1LV^2}{\mu}\right)^{\frac{p-1}{p}}\delta\right\}\le\Lambda. \tag{42}$$

*As a consequence, if we choose error of the oracle $\delta$ satisfying*

$$\delta\le\frac{\varepsilon(\mathrm{e}-1)}{2^pC_3\mathrm{e}}\left(\frac{6\mathrm{e}C_1LV^2}{\mu}\right)^{\frac{1-p}{p}}, \tag{43}$$

*then we need not more than $N=\left\lceil\ln\left(\frac{\mu R_0^2}{\varepsilon}\right)\right\rceil$ outer iterations and not more than*

$$\begin{aligned}&\left(1+\left(\frac{6\mathrm{e}C_1LV^2}{\mu}\right)^{\frac{1}{p}}\right)\left(1+\ln\left(\frac{\mu R_0^2}{\varepsilon}\right)\right)+\\&+\frac{36\mathrm{e}^3C_2^2\sigma^2V^2}{\mu(\mathrm{e}-1)\varepsilon}\left(1+\ln\left(\frac{3}{\Lambda}\left(1+\ln\left(\frac{\mu R_0^2}{\varepsilon}\right)\right)\right)\right)^2\\&+\frac{144\mathrm{e}^3C_4^2\sigma^2}{\mu\varepsilon(\mathrm{e}-1)}\ln\left(\frac{3}{\Lambda}\left(1+\ln\left(\frac{\mu R_0^2}{\varepsilon}\right)\right)\right)\end{aligned} \tag{44}$$

*oracle calls to guarantee that $\mathbb{P}\{\varphi(u_N)-\varphi^*>\varepsilon\}\le\Lambda$.*

*Proof* The proof of this theorem can be found in "Appendix C." □

## 5 Conclusions

In this paper, we propose the Stochastic Intermediate Gradient Method, which can be used for convex composite optimization problems with stochastic inexact oracle. We estimate its rate of convergence in terms of the non-optimality gap expectation and provide the bound on the probability of large deviations for the non-optimality gap, which has the same asymptotic dependence on the iteration number. The main advantage of this method is that it provides several degrees of freedom for adapting it to the problem at hand.

1. Depending on the relations between the error of the oracle and the Lipschitz constant, one can choose an optimal trade-off between error accumulation and rate of convergence.
2. One can introduce a randomization to the problem if some stochastic approximation of the gradient is cheaper to obtain than the real gradient. Since the rate of convergence depends only on the number of iterations, but not on the number of calls of the oracle, one can use Monte Carlo approach and generate several realizations of the stochastic approximation of the gradient on each iteration. This can reduce the variance of the stochastic approximation.

3. The concept of stochastic inexact oracle allows to use the proposed method to solve non-smooth problems; see [8].
4. Since the method uses a general prox-function and norm, one can choose them optimally, depending on the geometry of the feasible set. The classical example is the standard simplex and the entropy prox-function.
5. The method allows to solve composite optimization problems such as the LASSO.

We provide an extension of this method for the case of problems with a strongly convex objective function and estimate its rate of convergence in terms of the expected non-optimality gap, and another extension for the case when one needs a solution with controlled level of confidence in corresponding non-optimality gap. It follows from the results of [3], [10] that the obtained rates of convergence of our algorithms lead to the complexity estimates, which up to a multiplicative constant coincide with the lower complexity bound for the considered class of convex composite optimization problems with stochastic inexact oracle.

**Acknowledgments** The research presented in Sect. 4 of this paper was conducted in IITP RAS and was supported by the Russian Science Foundation Grant (project 14-50-00150), and the research presented in other sections was supported by RFBR, research project No. 15-31-20571 mol_a_ved. Authors would like to thank professor Yurii Nesterov and professor Arkadi Nemirovski for useful discussions. Also we are grateful to two anonymous reviewers for their suggestions which helped to improve the text.

## Appendix A: Proof of Lemma 3.1

Note that for all $g \in E^*, x \in E, \zeta > 0$,

$$\langle g, x\rangle + \frac{\zeta}{2}\|x\|^2 \ge -\frac{1}{\zeta}\|g\|_*^2. \tag{45}$$

Let us first prove that the statement is true for $k = 0$. Since $\alpha_0 = A_0$, we have

$$\begin{aligned}
\Psi_0^* &\overset{(18),(12)}{=} \beta_0 d(y_0) + \alpha_0\left[F_0 + \langle G_0, y_0 - x_0\rangle + h(y_0)\right]\\
&\overset{(4)}{\ge} \tfrac{\beta_0}{2}\|y_0 - x_0\|^2 + \alpha_0\left[F_0 + \langle G_0, y_0 - x_0\rangle + h(y_0)\right]\\
&\overset{(7)}{\ge} \alpha_0\left[F_0 + \langle G_0, y_0 - x_0\rangle + h(y_0) + \tfrac{\beta_0}{2}\|y_0 - x_0\|^2\right]\\
&= \alpha_0\left[f_0 + \langle g_0, y_0 - x_0\rangle + h(y_0) + \tfrac{L}{2}\|y_0 - x_0\|^2\right]\\
&\quad + \alpha_0\left[F_0 - f_0 + \langle G_0 - g_0, y_0 - x_0\rangle + \tfrac{\beta_0 - L}{2}\|y_0 - x_0\|^2\right]\\
&\overset{(2),(45)}{\ge} \alpha_0\left[f(y_0) + h(y_0) - \delta\right] + \alpha_0\left[F_0 - f_0\right] - \tfrac{\alpha_0}{\beta_0 - L}\|G_0 - g_0\|_*^2.
\end{aligned}$$

In view of $\alpha_0 = A_0 = B_0$, this proves (19) for $k = 0$.

Let us assume that (19) holds for some $k \ge 0$ and prove that it also holds for $k + 1$. Let $gh(z_k) \in \partial h(z_k)$. From the optimality condition in (13), we have

$$\left\langle \beta_k \nabla d(z_k) + \sum_{i=0}^{k} \alpha_i G_i + A_k gh(z_k), x - z_k \right\rangle \ge 0, \quad \forall x \in Q$$

and, hence,

$$\beta_k \langle \nabla d(z_k), x - z_k\rangle \geq \sum_{i=0}^{k} \alpha_i \langle G_i, z_k - x\rangle + A_k\langle gh(z_k), z_k - x\rangle, \quad \forall x \in Q. \tag{46}$$

At the same time, due to the convexity of $h(x)$

$$\begin{aligned} A_{k+1}h(x) + A_k\langle gh(z_k), z_k - x\rangle &\overset{(10)}{=} A_k h(x) + A_k\langle gh(z_k), z_k - x\rangle + \alpha_{k+1}h(x) \\ &\geq A_k h(z_k) + A_k\langle gh(z_k), x - z_k\rangle + A_k\langle gh(z_k), z_k - x\rangle + \alpha_{k+1}h(x) \\ &\overset{(10)}{=} A_k h(z_k) + \alpha_{k+1}h(x) \end{aligned} \tag{47}$$

Now we get for all $x \in Q$

$$\begin{aligned} \Psi_{k+1}(x) &\overset{(18)}{=} \beta_{k+1}d(x) + \textstyle\sum_{i=0}^{k+1} \alpha_i \left[F_i + \langle G_i, x - x_i\rangle\right] + A_{k+1}h(x) \\ &\overset{(7),(5)}{\geq} \beta_k V(x, z_k) + \beta_k d(z_k) + \beta_k\langle \nabla d(z_k), x - z_k\rangle + \\ &\quad + \textstyle\sum_{i=0}^{k+1} \alpha_i \left[F_i + \langle G_i, x - x_i\rangle\right] + A_{k+1}h(x) \\ &\overset{(46)}{\geq} \beta_k V(x, z_k) + \beta_k d(z_k) + \textstyle\sum_{i=0}^{k} \alpha_i\langle G_i, z_k - x\rangle + A_k\langle gh(z_k), z_k - x\rangle \\ &\quad + \textstyle\sum_{i=0}^{k+1} \alpha_i \left[F_i + \langle G_i, x - x_i\rangle\right] + A_{k+1}h(x) \\ &= \beta_k V(x, z_k) + \beta_k d(z_k) + A_{k+1}h(x) + A_k\langle gh(z_k), z_k - x\rangle \\ &\quad + \textstyle\sum_{i=0}^{k} \alpha_i \left[F_i + \langle G_i, z_k - x_i\rangle\right] + \alpha_{k+1}\left[F_{k+1} + \langle G_{k+1}, x - x_{k+1}\rangle\right] \\ &\overset{(47)}{\geq} \beta_k V(x, z_k) + \beta_k d(z_k) + \textstyle\sum_{i=0}^{k} \alpha_i \left[F_i + \langle G_i, z_k - x_i\rangle\right] + A_k h(z_k) + \\ &\quad + \alpha_{k+1}\left[F_{k+1} + \langle G_{k+1}, x - x_{k+1}\rangle + h(x)\right] \\ &\overset{(18),(13)}{=} \beta_k V(x, z_k) + \Psi_k^* + \alpha_{k+1}\left[F_{k+1} + \langle G_{k+1}, x - x_{k+1}\rangle + h(x)\right]. \end{aligned} \tag{48}$$

Also, since $A_k \overset{(10)}{=} (A_{k+1} - B_{k+1}) + (B_{k+1} - \alpha_{k+1})$, we have

$$\begin{aligned} &\Psi_k^* + \alpha_{k+1}\left[F_{k+1} + \langle G_{k+1}, x - x_{k+1}\rangle + h(x)\right] \\ &\overset{(19)}{\geq} A_k\varphi(y_k) - E_k + \alpha_{k+1}\left[F_{k+1} + \langle G_{k+1}, x - x_{k+1}\rangle + h(x)\right] \\ &\overset{(1)}{=} (A_{k+1} - B_{k+1})f(y_k) + (B_{k+1} - \alpha_{k+1})f(y_k) + A_k h(y_k) - E_k \\ &\quad + \alpha_{k+1}\left[F_{k+1} + \langle G_{k+1}, x - x_{k+1}\rangle + h(x)\right] \\ &\overset{(2)}{\geq} (A_{k+1} - B_{k+1})f(y_k) + (B_{k+1} - \alpha_{k+1})(f_{k+1} + \langle g_{k+1}, y_k - x_{k+1}\rangle) + \\ &\quad + A_k h(y_k) - E_k + \alpha_{k+1}\left[F_{k+1} + \langle G_{k+1}, x - x_{k+1}\rangle\right] + \alpha_{k+1}h(x) \\ &= (A_{k+1} - B_{k+1})f(y_k) + (B_{k+1} - \alpha_{k+1})f_{k+1} + \alpha_{k+1}F_{k+1} \\ &\quad + (B_{k+1} - \alpha_{k+1})\langle g_{k+1}, y_k - x_{k+1}\rangle + \alpha_{k+1}\langle G_{k+1}, x - x_{k+1}\rangle \\ &\quad + A_k h(y_k) - E_k + \alpha_{k+1}h(x) \\ &= (A_{k+1} - B_{k+1})f(y_k) + (B_{k+1} - \alpha_{k+1})(f_{k+1} - F_{k+1}) + B_{k+1}F_{k+1} \end{aligned} \tag{49}$$

$$
\begin{aligned}
&+(B_{k+1}-\alpha_{k+1})\langle g_{k+1}-G_{k+1}, y_k-x_{k+1}\rangle \\
&+\langle G_{k+1}, (B_{k+1}-\alpha_{k+1})(y_k-x_{k+1})+\alpha_{k+1}(x-x_{k+1})\rangle \\
&+A_k h(y_k)-E_k+\alpha_{k+1}h(x) \\
\overset{(10)}{=}\ &(A_{k+1}-B_{k+1})f(y_k)+(B_{k+1}-\alpha_{k+1})(f_{k+1}-F_{k+1})+B_{k+1}F_{k+1} \\
&+(B_{k+1}-\alpha_{k+1})\langle g_{k+1}-G_{k+1}, y_k-x_{k+1}\rangle \\
&+\langle G_{k+1}, (B_{k+1}-\alpha_{k+1})(y_k-x_{k+1})+\alpha_{k+1}(x-x_{k+1})\rangle \\
&+(A_{k+1}-B_{k+1}+B_{k+1}-\alpha_{k+1})h(y_k)-E_k+\alpha_{k+1}h(x) \\
\overset{(14)}{=}\ &(A_{k+1}-B_{k+1})(f(y_k)+h(y_k)) \\
&+(B_{k+1}-\alpha_{k+1})\left[f_{k+1}-F_{k+1}+\langle g_{k+1}-G_{k+1}, y_k-x_{k+1}\rangle+h(y_k)\right] \\
&+B_{k+1}F_{k+1}+\alpha_{k+1}\langle G_{k+1}, x-z_k\rangle-E_k+\alpha_{k+1}h(x).
\end{aligned}
$$

In the last equality we used, from (14) and (11), it follows that

$$
(B_{k+1}-\alpha_{k+1})x_{k+1}+\alpha_{k+1}x_{k+1}=\alpha_{k+1}z_k+(B_{k+1}-\alpha_{k+1})y_k.
$$

Hence,

$$
(B_{k+1}-\alpha_{k+1})(y_k-x_{k+1})-\alpha_{k+1}x_{k+1}=-\alpha_{k+1}z_k
$$

and

$$
(B_{k+1}-\alpha_{k+1})(y_k-x_{k+1})+\alpha_{k+1}(x-x_{k+1})=\alpha_{k+1}(x-z_k).
$$

Thus, for all $x \in Q$, we have

$$
\begin{aligned}
&\Psi_{k+1}(x) \overset{(48),(50),(1)}{\geq} (A_{k+1}-B_{k+1})\varphi(y_k)+B_{k+1}F_{k+1} \\
&+\beta_k V(x,z_k)+\alpha_{k+1}\langle G_{k+1}, x-z_k\rangle+\alpha_{k+1}h(x) \\
&+(B_{k+1}-\alpha_{k+1})\left[f_{k+1}-F_{k+1}+\langle g_{k+1}-G_{k+1}, y_k-x_{k+1}\rangle+h(y_k)\right]-E_k \quad (50)
\end{aligned}
$$

At the same time,

$$
\frac{\beta_k}{B_{k+1}} \overset{(9)}{\geq} \frac{\alpha_{k+1}^2\beta_{k+1}}{B_{k+1}^2} \overset{(11)}{=} \tau_k^2\beta_{k+1}. \tag{51}
$$

Using (50), we obtain

$$\begin{aligned}
\Psi_{k+1}^{*} \quad & \geq \quad B_{k+1}F_{k+1} + \min\nolimits_{x\in Q}\{\beta_k V(x,z_k) + \alpha_{k+1}\langle G_{k+1}, x - z_k\rangle + \\
& \qquad + \alpha_{k+1}h(x)\} + (A_{k+1} - B_{k+1})\varphi(y_k) + \\
& \qquad (B_{k+1} - \alpha_{k+1})\left[f_{k+1} - F_{k+1} + \langle g_{k+1} - G_{k+1}, y_k - x_{k+1}\rangle + h(y_k)\right] \\
& \overset{(15)}{=} \quad B_{k+1}F_{k+1} + \beta_k V(\hat{x}_{k+1}, z_k) + \alpha_{k+1}\langle G_{k+1}, \hat{x}_{k+1} - z_k\rangle + \\
& \qquad + \alpha_{k+1}h(\hat{x}_{k+1}) - E_k + (A_{k+1} - B_{k+1})\varphi(y_k) + \\
& \qquad (B_{k+1} - \alpha_{k+1})\left[f_{k+1} - F_{k+1} + \langle g_{k+1} - G_{k+1}, y_k - x_{k+1}\rangle + h(y_k)\right] \\
& \overset{(6)}{\geq} \quad B_{k+1}\big[F_{k+1} + \tau_k\langle G_{k+1}, \hat{x}_{k+1} - z_k\rangle + \tfrac{\beta_k}{B_{k+1}}\|\hat{x}_{k+1} - z_k\|^2\big] - \\
& \qquad - E_k + (B_{k+1} - \alpha_{k+1})\left[f_{k+1} - F_{k+1} + \langle g_{k+1} - G_{k+1}, y_k - x_{k+1}\rangle\right] \\
& \qquad + B_{k+1}(\tau_k h(\hat{x}_{k+1}) + (1 - \tau_k)h(y_k)) + (A_{k+1} - B_{k+1})\varphi(y_k) \\
& \overset{(51),(16)}{\geq} \quad B_{k+1}\big[F_{k+1} + \tau_k\langle G_{k+1}, \hat{x}_{k+1} - z_k\rangle + \tfrac{\tau_k^2\beta_{k+1}}{2}\|\hat{x}_{k+1} - z_k\|^2\big] - E_k \\
& \qquad + (B_{k+1} - \alpha_{k+1})\left[f_{k+1} - F_{k+1} + \langle g_{k+1} - G_{k+1}, y_k - x_{k+1}\rangle\right] \\
& \qquad + B_{k+1}h(w_{k+1}) + (A_{k+1} - B_{k+1})\varphi(y_k) \\
& \overset{(14),(16)}{\geq} \quad B_{k+1}\big[F_{k+1} + \langle G_{k+1}, w_{k+1} - x_{k+1}\rangle + \tfrac{\beta_{k+1}}{2}\|w_{k+1} - x_{k+1}\|^2 \\
& \qquad + h(w_{k+1})\big] - E_k + (B_{k+1} - \alpha_{k+1})\big[f_{k+1} - F_{k+1} \\
& \qquad + \langle g_{k+1} - G_{k+1}, y_k - x_{k+1}\rangle\big] + (A_{k+1} - B_{k+1})\varphi(y_k) \\
& = \quad B_{k+1}\big[f_{k+1} + \langle g_{k+1}, w_{k+1} - x_{k+1}\rangle + \tfrac{L}{2}\|w_{k+1} - x_{k+1}\|^2 \\
& \qquad + h(w_{k+1})\big] - E_k + B_{k+1}\big[F_{k+1} - f_{k+1} \\
& \qquad + \langle G_{k+1} - g_{k+1}, w_{k+1} - x_{k+1}\rangle + \tfrac{\beta_{k+1} - L}{2}\|w_{k+1} - x_{k+1}\|^2\big] \\
& \qquad + (B_{k+1} - \alpha_{k+1})\left[f_{k+1} - F_{k+1} + \langle g_{k+1} - G_{k+1}, y_k - x_{k+1}\rangle\right] \\
& \qquad + (A_{k+1} - B_{k+1})\varphi(y_k) \\
& \overset{(2)}{\geq} \quad B_{k+1}(f(w_{k+1}) + h(w_{k+1}) - \delta) - E_k + \alpha_{k+1}(F_{k+1} - f_{k+1}) \\
& \qquad + (B_{k+1} - \alpha_{k+1})\langle g_{k+1} - G_{k+1}, y_k - x_{k+1}\rangle + \\
& \qquad + B_{k+1}\big[\langle G_{k+1} - g_{k+1}, w_{k+1} - x_{k+1}\rangle + \tfrac{\beta_{k+1} - L}{2}\|w_{k+1} - x_{k+1}\|^2\big] \\
& \qquad + (A_{k+1} - B_{k+1})\varphi(y_k) \\
& = \quad A_{k+1}\left(\tfrac{B_{k+1}}{A_{k+1}}\varphi(w_{k+1}) + \tfrac{A_{k+1} - B_{k+1}}{A_{k+1}}\varphi(y_k)\right) - B_{k+1}\delta - E_k \\
& \qquad + \alpha_{k+1}(F_{k+1} - f_{k+1}) + (B_{k+1} - \alpha_{k+1})\langle g_{k+1} - G_{k+1}, y_k - x_{k+1}\rangle + \\
& \qquad + B_{k+1}\big[\langle G_{k+1} - g_{k+1}, w_{k+1} - x_{k+1}\rangle + \tfrac{\beta_{k+1} - L}{2}\|w_{k+1} - x_{k+1}\|^2\big] \\
& \overset{(17),(45)}{\geq} \quad A_{k+1}\varphi(y_{k+1}) - E_k - B_{k+1}\delta + \alpha_{k+1}(F_{k+1} - f_{k+1}) \\
& \qquad + (B_{k+1} - \alpha_{k+1})\langle g_{k+1} - G_{k+1}, y_k - x_{k+1}\rangle \\
& \qquad - \tfrac{B_{k+1}}{\beta_{k+1} - L}\|g_{k+1} - G_{k+1}\|_*^2.
\end{aligned} \tag{52}$$

Finally, from (14), we have

$$y_k - x_{k+1} \overset{(14)}{=} \tau_k(y_k - z_k) \overset{(11)}{=} \frac{\alpha_{k+1}}{B_{k+1}}(y_k - z_k).$$

Thus, by (52), we obtain (19) for $k + 1$. □

## Appendix B: Proof of Theorem 4.1

Obviously, (35) follows from (34) and (29). Let us prove the inequality

$$\mathbb{E}\varphi(u_k) - \varphi^* \leq \frac{\mu R_0^2}{2}\mathrm{e}^{-k} + \frac{C_3 \mathrm{e} 2^{p-1}}{e-1}\left(\frac{4\mathrm{e}C_1 L V^2}{\mu}\right)^{\frac{p-1}{p}}\left(1-\mathrm{e}^{-k}\right)\delta \tag{53}$$

for all $k \geq 1$. Then we will have (34) as a consequence. Let us prove (53) for $k = 1$. It follows from Lemma 4.1 that

$$\mathbb{E}\varphi(y_{N_0}) - \varphi^* \leq \frac{C_1 L R_0^2 V^2}{N_0^p} + \frac{C_2 \sigma R_0 V}{\sqrt{m_0 N_0}} + C_3 N_0^{p-1}\delta. \tag{54}$$

From (31), we have

$$\begin{aligned} \frac{C_1 L R_0^2 V^2}{N_0^p} &\leq \frac{C_1 L R_0^2 V^2}{\frac{4\mathrm{e}C_1 L V^2}{\mu}} \leq \frac{\mu R_0^2}{4\mathrm{e}}, \\ C_3 N_0^{p-1}\delta &\leq \frac{C_3 \mathrm{e} 2^{p-1}}{\mathrm{e}-1}\left(\frac{4\mathrm{e}C_1 L V^2}{\mu}\right)^{\frac{p-1}{p}}\left(1-\mathrm{e}^{-1}\right)\delta. \end{aligned} \tag{55}$$

Using (32), we obtain

$$\frac{C_2 \sigma R_0 V}{\sqrt{m_0 N_0}} \leq \frac{C_2 \sigma R_0 V}{\sqrt{\frac{16\mathrm{e}^2 C_2^2 \sigma^2 V^2}{\mu^2 R_0^2 N_0} N_0}} \leq \frac{\mu R_0^2}{4\mathrm{e}}.$$

This with (54) and (55) proves (53) for $k = 1$.

Let us now assume that (53) holds for $k = j$ and prove that it holds for $k = j + 1$. It follows from (29) and(53) for $k = j$ that

$$\begin{aligned} \mathbb{E}\|u_j - x^*\|^2 &\leq \frac{2}{\mu}\left(\mathbb{E}\varphi(u_j) - \varphi^*\right) \\ &\leq \frac{2}{\mu}\left(\frac{\mu R_0^2}{2}\mathrm{e}^{-j} + \frac{C_3 \mathrm{e} 2^{p-1}}{\mathrm{e}-1}\left(\frac{4\mathrm{e}C_1 L V^2}{\mu}\right)^{\frac{p-1}{p}}\left(1-\mathrm{e}^{-j}\right)\delta\right) \overset{(33)}{=} R_j^2. \end{aligned}$$

After $N_j$ iterations of Algorithm 1 with starting point $u_j$ (or $y_{N_{j-1}}$, which is the same), applying Lemma 4.1, we have

$$\mathbb{E}\varphi(y_{N_j}) - \varphi^* \leq \frac{C_1 L R_j^2 V^2}{N_j^p} + \frac{C_2 \sigma R_j V}{\sqrt{m_j N_j}} + C_3 N_j^{p-1}\delta. \tag{56}$$

From (31), we have

$$\frac{C_1LR_j^2V^2}{N_j^p} \leq \frac{C_1LR_j^2V^2}{\frac{4\mathrm{e}C_1LV^2}{\mu}} \leq \frac{\mu R_j^2}{4\mathrm{e}}, \quad C_3N_j^{p-1}\delta \leq C_3 2^{p-1}\left(\frac{4\mathrm{e}C_1LV^2}{\mu}\right)^{\frac{p-1}{p}}\delta. \tag{57}$$

By (32),

$$m_j \geq \frac{16\mathrm{e}^{j+2}C_2^2\sigma^2V^2}{\mu^2R_0^2N_j} \geq \frac{16\mathrm{e}^2C_2^2\sigma^2V^2}{\mu^2N_j\left(R_0^2\mathrm{e}^{-j}+\frac{2^p\mathrm{e}C_3\delta}{\mu(\mathrm{e}-1)}\left(\frac{4\mathrm{e}C_1LV^2}{\mu}\right)^{\frac{p-1}{p}}\left(1-\mathrm{e}^{-j}\right)\delta\right)}$$

$$\overset{(33)}{=} \frac{16\mathrm{e}^2C_2^2\sigma^2V^2}{\mu^2R_j^2N_j},$$

and, hence,

$$\frac{C_2\sigma R_jV}{\sqrt{m_jN_j}} \leq \frac{C_2\sigma R_jV}{\sqrt{\frac{16\mathrm{e}^2C_2^2\sigma^2V^2}{\mu^2R_j^2N_j}N_j}} \leq \frac{\mu R_j^2}{4\mathrm{e}}. \tag{58}$$

Finally, we arrive at

$$\mathbb{E}\varphi(u_{j+1})-\varphi^* = \mathbb{E}\varphi(y_{N_j})-\varphi^* \overset{(56),(57),(58)}{\leq} \frac{\mu R_j^2}{2\mathrm{e}} + C_3 2^{p-1}\left(\frac{4\mathrm{e}C_1LV^2}{\mu}\right)^{\frac{p-1}{p}}\delta \overset{(33)}{=}$$

$$= \frac{1}{\mathrm{e}}\left(\frac{\mu R_0^2}{2}\mathrm{e}^{-j} + \frac{C_3\mathrm{e}2^{p-1}}{\mathrm{e}-1}\left(\frac{4\mathrm{e}C_1LV^2}{\mu}\right)^{\frac{p-1}{p}}\left(1-\mathrm{e}^{-j}\right)\delta\right)$$

$$+ C_3 2^{p-1}\left(\frac{4\mathrm{e}C_1LV^2}{\mu}\right)^{\frac{p-1}{p}}\delta$$

$$= \frac{\mu R_0^2}{2}\mathrm{e}^{-(j+1)} + \frac{C_3\mathrm{e}2^{p-1}}{\mathrm{e}-1}\left(\frac{4\mathrm{e}C_1LV^2}{\mu}\right)^{\frac{p-1}{p}}\left(1-\mathrm{e}^{-(j+1)}\right)\delta.$$

Thus, we have obtained that (53) holds for $k = j+1$ and, by induction, it holds for all $k \geq 1$. If we choose $\delta$ satisfying (36) and perform $N = \left\lceil \ln\left(\frac{\mu R_0^2}{\varepsilon}\right)\right\rceil$ outer iterations of Algorithm 2, we will obtain from (34)

$$\mathbb{E}\varphi(u_N)-\varphi^* \leq \frac{\mu R_0^2}{2}\mathrm{e}^{-N} + \frac{C_3\mathrm{e}2^{p-1}}{e-1}\left(\frac{4\mathrm{e}C_1LV^2}{\mu}\right)^{\frac{p-1}{p}}\delta \leq \frac{\varepsilon}{2}+\frac{\varepsilon}{2} = \varepsilon.$$

It remains to calculate the number of oracle calls to obtain an $\varepsilon$-solution in the sense that $\mathbb{E}\varphi(u_N)-\varphi^* \leq \varepsilon$. We perform $N$ outer iterations ($k$ runs from 0 to $N-1$), on each outer iteration $k$, we perform $N_k$ inner iterations, and, on each inner iteration, we call the oracle $m_k$ times. Hence, the total number of oracle calls is

$$\mathcal{C}(\varepsilon) = \sum_{k=0}^{N-1} N_k m_k \leq \sum_{k=0}^{N-1} N_k \left(1 + \frac{16C_2^2 \mathrm{e}^2 \sigma^2 V^2 e^k}{\mu^2 R_0^2 N_k}\right) \leq N N_0 + \frac{16C_2^2 \mathrm{e}^2 \sigma^2 V^2 \mathrm{e}^N}{\mu^2 R_0^2 (\mathrm{e}-1)}$$

$$\leq \left(1 + \ln\left(\frac{\mu R_0^2}{\varepsilon}\right)\right)\left(1 + \left(\frac{4\mathrm{e}C_1 L V^2}{\mu}\right)^{\frac{1}{p}}\right) + \frac{16\mathrm{e}^3 C_2^2 \sigma^2 V^2}{\mu\varepsilon(\mathrm{e}-1)}$$

$$\leq \left(1 + \ln\left(\frac{\mu R_0^2}{\varepsilon}\right)\right)\left(1 + \left(\frac{62 L V^2}{\mu}\right)^{\frac{1}{p}}\right) + \frac{96000\sigma^2 V^2}{\mu\varepsilon}.$$

□

## Appendix C: Proof of Theorem 4.2

Let $A_k, k \geq 0$ be the event $A_k := \left\{\varphi(u_k) - \varphi^* \leq \frac{\mu R_k^2}{2}\right\}$ and $\bar{A}_k$ be its complement. Let us first prove that, for $k \geq 1$,

$$\mathbb{P}\left\{\varphi(u_k) - \varphi^* > \frac{\mu R_k^2}{2} \,\middle|\, A_{k-1}\right\} \leq \frac{\Lambda}{N}. \tag{59}$$

Since the event $A_{k-1}$ holds, we have from (29)

$$\left\|u_{k-1} - x^*\right\|^2 \leq \frac{2}{\mu}\left(\varphi(u_{k-1}) - \varphi^*\right) \leq R_{k-1}^2.$$

Hence, the solution of the problem $\min_{x\in Q_{k-1}} \varphi(x)$ is the same as the solution of the initial Problem (1). Let us denote $D_{k-1} := \max_{x,y\in Q_{k-1}} \|x - y\|$. Clearly, $D_{k-1} \leq 2R_{k-1}$. Note that $D_{k-1} = R_{k-1} \max_{x,y\in Q_{k-1}} \frac{\|x-y\|}{R_{k-1}}$ and the diameter of the set $Q_{k-1}$ with respect to the norm $\frac{\|\cdot\|}{R_{k-1}}$ is not greater than 2. We apply Theorems 3.2 and 3.4 with $LR_{k-1}^2$ in the role of $L$, $\frac{\sigma R_{k-1}}{\sqrt{m_{k-1}}}$ in the role of $\sigma$, $V$ in the role of $R$, 2 in the role of $D$, use (37) and make the same argument as in the proof of Lemma 4.1. This leads to the following inequality

$$\mathbb{P}\left\{\varphi(u_k) - \varphi^* > \frac{C_1 L R_{k-1}^2 V^2}{N_{k-1}^p} + \frac{C_2(1+\Omega)\sigma R_{k-1} V}{\sqrt{m_{k-1}N_{k-1}}} + C_3 N_{k-1}^{p-1}\delta + \right.$$
$$\left. \frac{2C_4 R_{k-1}\sigma\sqrt{\Omega}}{\sqrt{m_{k-1}N_{k-1}}} \,\middle|\, A_{k-1}\right\} \leq \frac{\Lambda}{N}, \tag{60}$$

where $C_1 = 4\sqrt{2}$, $C_2 = 16\sqrt{2}$, $C_3 = 48$, $C_4 = 4\sqrt{3}$, $\Omega = \ln\left(\frac{3N}{\Lambda}\right)$. Using (38), we have

$$\frac{C_1 L R_{k-1}^2 V^2}{N_{k-1}^p} \le \frac{C_1 L R_{k-1}^2 V^2}{\frac{6\mathrm{e}C_1 L V^2}{\mu}} \le \frac{\mu R_{k-1}^2}{6\mathrm{e}}, \;\; C_3 N_{k-1}^{p-1}\delta \le C_3 2^{p-1}\left(\frac{6\mathrm{e}C_1 L V^2}{\mu}\right)^{\frac{p-1}{p}}\delta. \tag{61}$$

From (39), we have

$$m_{k-1} \ge \frac{36\mathrm{e}^{k+1} C_2^2 \sigma^2 V^2 (1+\Omega)^2}{\mu^2 R_0^2 N_{k-1}} \ge \frac{36\mathrm{e}^2 C_2^2 \sigma^2 V^2 (1+\Omega)^2}{\mu^2 N_{k-1}\left(R_0^2 \mathrm{e}^{-(k-1)} + \frac{2^p \mathrm{e} C_3 \delta}{\mu(\mathrm{e}-1)}\left(\frac{6\mathrm{e}C_1 L V^2}{\mu}\right)^{\frac{p-1}{p}}\left(1-\mathrm{e}^{-(k-1)}\right)\delta\right)} \overset{(40)}{=} \frac{36\mathrm{e}^2 C_2^2 \sigma^2 V^2 (1+\Omega)^2}{\mu^2 R_{k-1}^2 N_{k-1}},$$

and, hence,

$$\frac{C_2 \sigma R_{k-1} V (1+\Omega)}{\sqrt{m_{k-1} N_{k-1}}} \le \frac{C_2 \sigma R_{k-1} V (1+\Omega)}{\sqrt{\frac{36\mathrm{e}^2 C_2^2 \sigma^2 V^2 (1+\Omega)^2}{\mu^2 R_{k-1}^2 N_{k-1}} N_{k-1}}} \le \frac{\mu R_{k-1}^2}{6\mathrm{e}}. \tag{62}$$

Also, by (39), we have

$$m_{k-1} \ge \frac{144\mathrm{e}^{k+1} C_4^2 \sigma^2 \Omega}{\mu^2 R_0^2 N_{k-1}} \ge \frac{144\mathrm{e}^2 C_4^2 \sigma^2 \Omega}{\mu^2 N_{k-1}\left(R_0^2 \mathrm{e}^{-(k-1)} + \frac{2^p \mathrm{e} C_3 \delta}{\mu(\mathrm{e}-1)}\left(\frac{6\mathrm{e}C_1 L V^2}{\mu}\right)^{\frac{p-1}{p}}\left(1-\mathrm{e}^{-(k-1)}\right)\delta\right)} \overset{(40)}{=} \frac{144\mathrm{e}^2 C_4^2 \sigma^2 \Omega}{\mu^2 R_{k-1}^2 N_{k-1}},$$

and, hence,

$$\frac{2C_4 R_{k-1} \sigma \sqrt{\Omega}}{\sqrt{m_{k-1} N_{k-1}}} \le \frac{2C_4 R_{k-1} \sigma \sqrt{\Omega}}{\sqrt{\frac{144\mathrm{e}^2 C_4^2 \sigma^2 \Omega}{\mu^2 R_{k-1}^2 N_{k-1}} N_{k-1}}} \le \frac{\mu R_{k-1}^2}{6\mathrm{e}}. \tag{63}$$

Finally, we arrive at

$$\frac{C_1 L R_{k-1}^2 V^2}{N_{k-1}^p} + \frac{C_2(1+\Omega)\sigma R_{k-1} V}{\sqrt{m_{k-1}N_{k-1}}} + C_3 N_{k-1}^{p-1}\delta + \frac{2C_4 R_{k-1}\sigma\sqrt{\Omega}}{\sqrt{m_{k-1}N_{k-1}}} \overset{(61),(62),(63)}{\leq}$$

$$\leq \frac{\mu R_{k-1}^2}{2\mathrm{e}} + C_3 2^{p-1}\left(\frac{6\mathrm{e}C_1 L V^2}{\mu}\right)^{\frac{p-1}{p}}\delta$$

$$\overset{(40)}{=} \frac{1}{\mathrm{e}}\left(\frac{\mu R_0^2}{2}\mathrm{e}^{-(k-1)} + \frac{C_3 \mathrm{e} 2^{p-1}}{\mathrm{e}-1}\left(\frac{6\mathrm{e}C_1 L V^2}{\mu}\right)^{\frac{p-1}{p}}\left(1-\mathrm{e}^{-(k-1)}\right)\delta\right)$$

$$+ C_3 2^{p-1}\left(\frac{6\mathrm{e}C_1 L V^2}{\mu}\right)^{\frac{p-1}{p}}\delta$$

$$= \frac{\mu R_0^2}{2}\mathrm{e}^{-k} + \frac{C_3 \mathrm{e} 2^{p-1}}{\mathrm{e}-1}\left(\frac{6\mathrm{e}C_1 L V^2}{\mu}\right)^{\frac{p-1}{p}}\left(1-\mathrm{e}^{-k}\right)\delta = \frac{\mu R_k^2}{2}.$$

Thus, (59) follows from (60).

Also, for all $k = 1, \dots, N$, we have

$$\mathbb{P}\left\{\varphi(u_k) - \varphi^* > \frac{\mu R_k^2}{2}\right\} = \mathbb{P}\left\{\varphi(u_k) - \varphi^* > \frac{\mu R_k^2}{2}\,\middle|\, A_{k-1}\cup \bar{A}_{k-1}\right\}$$

$$= \mathbb{P}\left\{\varphi(u_k) - \varphi^* > \frac{\mu R_k^2}{2}\,\middle|\, A_{k-1}\right\}\mathbb{P}\{A_{k-1}\}$$

$$+ \mathbb{P}\left\{\varphi(u_k) - \varphi^* > \frac{\mu R_k^2}{2}\,\middle|\, \bar{A}_{k-1}\right\}\mathbb{P}\{\bar{A}_{k-1}\} \overset{(59)}{\leq}$$

$$\leq \frac{\Lambda}{N} + \mathbb{P}\{\bar{A}_{k-1}\} = \frac{\Lambda}{N} + \mathbb{P}\left\{\varphi(u_{k-1}) - \varphi^* > \frac{\mu R_{k-1}^2}{2}\right\}.$$

Using that $\mathbb{P}\{A_0\} = 1$ and summing up these inequalities, we obtain

$$\mathbb{P}\left\{\varphi(u_N) - \varphi^* > \frac{\mu R_0^2}{2}\mathrm{e}^{-N} + \frac{2^{p-1}\mathrm{e}C_3\delta}{(\mathrm{e}-1)}\left(\frac{6\mathrm{e}C_1 L V^2}{\mu}\right)^{\frac{p-1}{p}}\delta\right\} \leq$$

$$\mathbb{P}\left\{\varphi(u_N) - \varphi^* > \frac{\mu R_N^2}{2}\right\} \leq \Lambda.$$

Making the same arguments as in the proof of Theorem 4.1, we obtain the complexity bound (44). □

## References

1. Evtushenko, Y.: Methods of Solving Extremal Problems and Their Application in Optimization Systems. Nauka, Moscow (1982)
2. Polyak, B.T.: Introduction to Optimization. Optimization Software Inc, New York (1987)
3. Nemirovski, A., Yudin, D.: Problem Complexity and Method Efficiency in Optimization. Wiley, NewYork (1983)
4. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course. Kluwer Academic Publishers, Massachusetts (2004)
5. Khachiyan, L., Tarasov, S., Erlich, E.: The inscribed ellipsoid method. Soviet Math. Dokl. **298**, (1988) (**In Russian**)
6. Nemirovski, A., Nesterov, Y.: Interior Point Polynomial Methods in Convex Programming: Theory and Applications. SIAM, Philadelphia (1994)
7. Nesterov, Y.: Subgradient Methods for Huge-Scale Optimization Problems. CORE Discussion Paper 2012/2, Louvain-la-Neuve (2012)
8. Devolder, O., Glineur, F., Nesterov, Y.: First-order methods of smooth convex optimization with inexact oracle. Math. Program. **146**(1–2), 37–75 (2014)
9. Devolder, O., Glineur, F., Nesterov, Y.: Intermediate Gradient Methods for Smooth Convex Problems with Inexact Oracle. CORE Discussion Paper 2013/17, Louvain-la-Neuve (2013)
10. Devolder, O.: Exactness, Inexactness and Stochasticity in First-Order Methods for Large-Scale Convex Optimization, Ph.D. thesis, Louvain-la-Neuve (2013)
11. Ghadimi, S., Lan, G.: Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization I: a generic algorithmic framework. SIAM J. Optim. **22**(4), 1469–1492 (2012)
12. Ghadimi, S., Lan, G.: Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization II: shrinking procedures and optimal algorithms. SIAM J. Optim. **23**(4), 2061–2089 (2013)
13. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B **58**(1), 267–288 (1996)
14. Nesterov, Y.: Gradient methods for minimizing composite functions. Math. Prog. B **140**(1), 125–161 (2013)
15. Juditsky, A., Lan, G., Nemirovski, A., Shapiro, A.: Stochastic approximation approach to stochastic programming. SIAM J. Optim. **19**(4), 1574–1609 (2009)
16. Lan, G., Nemirovski, A., Shapiro, A.: Validation analysis of mirror descent stochastic approximation method. Math. Program. Ser. A **134**(2), 425–458 (2012)
17. Juditsky, A., Nesterov, Y.: Primal-dual subgradient methods for minimizing uniformly convex functions. Preprint ArXiv:1401.1792 (2014)